# IMPLEMENTASI ALGORITMA *K-MEANS* PADA KLASIFIKASI PENGELUARAN BELANJA PELANGGAN *MALL*

**SKRIPSI**

**Oleh:**
**GHIFARI DWI CAHYONO**
**NIM. 17650094**

**PROGRAM STUDI TEKNIK INFORMATIKA**
**FAKULTAS SAINS DAN TEKNOLOGI**
**UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM**
**MALANG**
**2023**

**IMPLEMENTASI ALGORITMA *K-MEANS* PADA KLASIFIKASI PENGELUARAN BELANJA PELANGGAN *MALL***

**SKRIPSI**

**Oleh:**
**GHIFARI DWI CAHYONO**
**NIM. 17650094**

**Diajukan kepada:**
**Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang**
**Untuk Memenuhi Salah Satu Persyaratan dalam**
**Memperoleh Gelar Sarjana Komputer (S.Kom)**

**PROGRAM STUDI TEKNIK INFORMATIKA**
**FAKULTAS SAINS DAN TEKNOLOGI**
**UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG**
**2023**

**HALAMAN PERSETUJUAN**

**IMPLEMENTASI ALGORITMA *K-MEANS* PADA KLASIFIKASI PENGELUARAN BELANJA PELANGGAN *MALL***

**SKRIPSI**

**Oleh:**
**GHIFARI DWI CAHYONO**
**NIM. 17650094**

**Telah Diperiksa dan Disetujui untuk Diuji**
**Tanggal: 12 Desember 2023**

| Pembimbing I | Pembimbing II |
|---|---|
| Dr. M. Faisal, M.T. | Fatchurrohman, M.Kom |
| NIP. 19740510 200501 1 007 | NIP. 19700731 200501 1 002 |

Mengetahui,
Ketua Program Studi Teknik InformatikaFakultas
Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

**Dr. Fachrul Kurniawan, M.MT, IPM**
**NIP. 19771020 200912 1 001**





iii

**HALAMAN PENGESAHAN**

**IMPLEMENTASI ALGORITMA *K-MEANS* PADA KLASIFIKASI PENGELUARAN BELANJA PELANGGAN *MALL***

**SKRIPSI**

**Oleh:**
**GHIFARI DWI CAHYONO**
**NIM. 17650094**

Telah Dipertahankan di Depan Dewan Penguji Skripsi
Dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer ( S.Kom )
Tanggal: 12 Desember 2023

Susunan Dewan Penguji

| | | |
|---|---|---|
| Ketua Penguji | : Dr. Fachrul Kurniawan, M.MT, IPM<br>NIP. 19771020 200912 1 001 | ( ) |
| Anggota Penguji I | : Dr. Yunifa Miftachul Arif, M.T<br>NIP. 19830616 201101 1 004 | ( ) |
| Anggota Penguji II | : Dr. M. Faisal, M.T.<br>NIP. 19740510 200501 1 007 | ( ) |
| Anggota Penguji III | : Fatchurrohman, M.Kom<br>NIP. 19700731 200501 1 002 | ( ) |

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

**Dr. Fachrul Kurniawan, M. MT, IPM**
**NIP. 19771020 200912 1 001**

## PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini :

| | |
|---|---|
| Nama | : Ghifari Dwi Cahyono |
| NIM | : 17650094 |
| Program Studi | : Teknik Informatika |
| Fakultas | : Sains dan Teknologi |
| Judul Skripsi | : Implementasi Algoritma *K-Means* Pada Klasifikasi Pengeluaran Belanja Pelanggan *Mall* |

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan dan pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka

Apabila dikemudian hari terbukti atau dapat dibuktikan Skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 20 Desember 2023

Yang Pernyataan,


Ghifari Dwi Cahyono

NIM. 17650094

## MOTTO

وَاخْفِضْ لَهُمَا جَنَاحَ الذُّلِّ مِنَ الرَّحْمَةِ وَقُلْ رَّبِّ ارْحَمْهُمَا كَمَا رَبَّيٰنِيْ صَغِيْرًا

*"Rendahkanlah dirimu terhadap keduanya dengan penuh kasih sayang dan ucapkanlah, "Wahai Tuhanku, sayangilah keduanya sebagaimana mereka berdua (menyayangiku ketika) mendidik aku pada waktu kecil." (Q.S. Al-Isra' : 24)*

## HALAMAN PERSEMBAHAN

*Skripsi ini saya persembahkan kepada keluarga dan seluruh sanak saudara saya, semoga skripsi ini menjadi keberhasilan atas kerja keras saya beserta do'a penuh kasih sayang keluarga dan menjadi tanda permintaan ma'af sebesar – besarnya atas rasa kecewa yang sering kali saya berikan.*

# KATA PENGANTAR

*Assalamu'alaikum Wr.Wb.*

Alhamdulillah patut bersyukur penulis ucapkan kehadirat Allah SWT yang telah memberikan Rahmat dan Hidayah-Nya, sehingga penulis dapat menyelesaikan studi di Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Penulis mengucapkan terimakasih dan do'a kepada semua pihak yang telah membantu terselesaikannya Skripsi ini. Ucapan terimakasih ini penulis sampaikan kepada:

1. Prof. Dr. HM. Zainuddin MA, selaku rektor UIN Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Harini, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, M. MT. IPM, selaku Kepala Prodi Teknik Informatika yang senantiasa memberikan solusi di setiap permasalahan mahasiswanya.
4. Dr. M. Faisal, M.T., selaku dosen pembimbing pertama, yang senantiasa tidak pernah lelah memberikan arahan, dukungan, dan motivasi untuk menyelesaikan skripsi ini.
5. Fatchurrohman, M.Kom, yang telah membimbing dan memberikan semangat serta motivasi kepada penulis agar segera menyelesaikan skripsi

6. Seluruh dosen Teknik Informatika yang telah memberikan ilmu dan pengalaman yang berharga selama masa perkuliahan.
7. Seluruh staf Teknik Informatika yang telah membantu dalam hal administrasi.
8. Kedua orang tua penulis, Ibu Wahyu Setyorini dan Bapak Sucipto, terimakasih sudah menjadi sumber semangat atas nasihat Bapak Ibu dan selalu mendukung penulis dalam setiap keadaan, dan tidak berhenti memberikan doa dan kasih sayang kepada penulis, dan selalu menanyakan skripsi agar diselesaikan, semoga ini semua dapat memberikan obat bagi rasa kecewa kalian
9. Kepada saudara penulis yang selalu memantau perkembangan skripsi, dan selalu memberikan perhatian serta semangat terhadap penulis untuk segera menyelesaikan skripsi.
10. Semua pihak yang ikut membantu dalam menyelesaikan Skripsi ini yang tidak dapat penulis sebutkan satu persatu

Penulis menyadari bahwa dalam penyusunan Skripsi ini masih terdapat kekurangan, penulis berharap Skripsi ini dapat memberikan manfaat terhadap pembaca dan khususnya bagi penulis secara pribadi. Amin Ya Rabbal Alamin

*Wassalamu'alaikum Wr. Wb.*

Malang, 12 Desember 2023

Penulis

## DAFTAR ISI

**HALAMAN PENGAJUAN** ..... ii
**HALAMAN PERSETUJUAN** .....**Error! Bookmark not defined.**
**HALAMAN PENGESAHAN**.....**Error! Bookmark not defined.**
**PERNYATAAN KEASLIAN TULISAN**.....**Error! Bookmark not defined.**
**MOTTO** ..... vi
**HALAMAN PERSEMBAHAN** ..... vii
**KATA PENGANTAR**..... viii
**DAFTAR ISI**..... x
**DAFTAR GAMBAR**..... xii
**DAFTAR TABEL** ..... xiii
**ABSTRAK** ..... xiv
**ABSTRACT**..... xv
خلاصة ..... xvi
**BAB I PENDAHULUAN**..... 1
**1.1 Latar Belakang** .....1
**1.2 Pernyataan Masalah**.....6
**1.3 Tujuan Penelitian** .....6
**1.4 Manfaat Penelitian**.....6
**1.5 Batasan Masalah**.....7
**1.6 Sistematika Penulisan**.....7
**BAB II STUDI PUSTAKA** ..... 9
**2.1 Studi Pustaka** .....9
**2.2 Kajian Teoritis** .....15
**2.2.1 *Machine Learning***.....16
**2.2.2 Metode Klasifikasi**.....24
**2.2.3 *K-Means***.....33
**BAB III DESAIN DAN IMPLEMENTASI** ..... 39
**3.1 Rancangan Penelitian**.....39
**3.1.1 Desain *Interface*** .....40
**3.1.2 *Flowchart***.....46
**3.2 Sistematika *K-Means*** .....49
**BAB IV UJI COBA DAN HASIL**..... 64

**4.1** ***Interface System***.....65
**4.2** **Pengecekan Statistik Dataset *Input***.....69
**4.3** **Pengujian Sistem**.....80
**4.4** **Pembahasan** .....90
**4.5** **Integrasi Terhadap Pandangan Islam** .....92
**BAB V KESIMPULAN DAN SARAN** ..... 95
**5.1** **Kesimpulan**.....95
**5.2** **Saran**.....96
**DAFTAR PUSTAKA**
**LAMPIRAN**

## DAFTAR GAMBAR

Gambar 2.1 Contoh Grafis *Supervised Learning*. Sumber : (Russell, R. 2018) ... 22
Gambar 2.2 *Data Clustering* dengan *K-Means*. Sumber : (Parsian, M. 2015) ..... 38
Gambar 3.1 Desain *Input* Data Pelanggan ..... 42
Gambar 3.2 Desain *Interface* Algoritma *K-Means* ..... 44
Gambar 3.3 Desain *Output* Grafik & Tabel Klasifikasi *K-Means* ..... 45
Gambar 3.4 Alur *Flowchart K-Means* ..... 47
Gambar 4.1 Pemanggilan *Localhost Jupyter* dengan *Command Prompt* ..... 66
Gambar 4.2 Tampilan *User Files Jupyter Notebook* ..... 67
Gambar 4.3 Tampilan *File Program K-Means* ..... 68
Gambar 4.4 Informasi Atribut dan Tipe Data ..... 73
Gambar 4.5 *Diagram Pie* Data Pelanggan Kolom "*Gender*" ..... 75
Gambar 4.6 *Grafik Bar* Data Pelanggan dengan Kolom "*Age*" ..... 76
Gambar 4.7 *Grafik Bar* Data Pelanggan Kolom "*Annual Income*" ..... 77
Gambar 4.8 *Grafik Bar* Data Pelanggan denagn Kolom "*Spending Score*" ..... 78
Gambar 4.9 Hasil Program Pemetaan Data tentang Pelanggan ..... 79
Gambar 4.10 Grafik Optimasi nilai "K" dengan *Elbow Method* ..... 83
Gambar 4.11 Hasil Grafik *Elbow* dengan *Silhouette Score* pada *K-Means* ..... 84
Gambar 4.12 Hasil *Clustering* Data Pelanggan ..... 86
Gambar 4.13 *Diagram Pie* 5 *Cluster K-Means* Data Pelanggan ..... 87

# DAFTAR TABEL

Tabel 2.1 Hasil Penelitian Sebelumnya ..... 13
Tabel 2.2 Perbandingan Penelitian pada Metode *K-Means* ..... 14
Tabel 3.1 Data *Mall Customer* ..... 52
Tabel 3.2 Deskripsi Data *Mall Customer* ..... 52
Tabel 3.3 Skala Standar Data ..... 56
Tabel 3.4 Data *Centroid* 5 Cluster ..... 57
Tabel 3.5 Data *Mall Customer* Beserta Kelompok *Cluster* ..... 63
Tabel 4.1 Tampilan Dataset "*Mall_Customer.csv*" ..... 71
Tabel 4.2 Informasi Jumlah dan Satuan Penting Data ..... 74
Tabel 4.3 Data Pelanggan *Cluster* 1 ..... 88
Tabel 4.4 Data Pelanggan *Cluster* 2 ..... 88
Tabel 4.5 Data Pelanggan *Cluster* 3 ..... 89
Tabel 4.6 Data Pelanggan *Cluster* 4 ..... 89
Tabel 4.7 Data Pelanggan *Cluster* 5 ..... 90

# ABSTRAK

Cahyono, Ghifari Dwi. 2023. **Implementasi Algoritma *K-Means* Pada Klasifikasi Pengeluaran Belanja Pelanggan *Mall*.** Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. M. Faisal, M.T. (II) Fatchurrohman, M.Kom

Kata Kunci: Mall, Pelanggan, *K-Means, Cluster, Elbow Method, Silhouette Score*

Proses bisnis dicapai oleh perusahaan yang memiliki visi jangka panjang dan strategi yang matang. Dengan melakukan upaya khusus untuk memuaskan pelanggan, perusahaan dapat menghasilkan keuntungan jangka panjang, menghemat biaya pemasaran, dan menghindari risiko kehilangan pangsa pasar. Pelanggan yang merasa puas dengan layanan yang diberikan cenderung kembali menggunakan layanan tersebut dan merekomendasikan kepada orang lain, sehingga dapat meningkatkan loyalitas pelanggan dan keuntungan bagi perusahaan *mall*. Meskipun pelanggan adalah aset berharga bagi setiap mall, ada beberapa pelanggan yang dapat menjadi beban dan merugikan mall. Pelanggan seperti ini dapat menyebabkan biaya tambahan bagi *mall* dalam hal penggantian produk atau pelayanan ulang. *Auditor internal* bertanggung jawab untuk memastikan bahwa perusahaan *mall* mengoperasikan bisnisnya dengan cara yang optimal dan sesuai dengan tujuan strategis yang telah ditetapkan. Oleh karena itu, *K-Means* dapat menjadi solusi metode dalam mengatasi permasalahan pelanggan *mall* karena *K-Means* adalah algoritma *data mining* untuk mengelompokkan atau mengklasifikasikan "N" objek berdasarkan atribut atau fitur mereka menjadi "K" kelompok (yang disebut *cluster*) dengan meminimalkan jumlah kuadrat jarak antara data dan *centroid cluster* yang sesuai. Dalam proses algoritma K-Means, dilakukan pelatihan dataset pelanggan *mall* untuk mendapatkan nilai "k" *cluster* yang optimal, selanjutnya dilakukan pengujian dataset dengan alur proses algoritma *K-Means* dan mengukur akurasi "k" terbaik dengan visualisasi *Silhouette Score*. Hasil pengujian algoritma *K-Means* dalam metode klasifikasi pengeluaran belanja pelanggan mall dapat meningkatkan tingkat akurasi sebesar 0.476 dengan Skala Standar dari perhitungan *Silhouette Score*. Kemudian pada *Elbow Method*, penelitian menunjukkan hasil "K" terbaik jatuh kepada k = 3 atau *Cluster* terbaik yaitu *Cluster* 4 jenis "*Sensible*" dengan jumlah pelanggan mencapai 22 orang. Dalam akurasi yang dihimpun dari perhitungan tersebut dinyatakan pemrograman sistem aplikasi yang dikembangkan dapat melengkapi algoritma *K-Means* dengan sempurna.

# ABSTRACT

Cahyono, Ghifari Dwi. 2023. **Implementation of K-Means Algorithm on Mall Customer Shopping Expenditure Classification**. Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang. Supervisor: (I) Dr. M. Faisal, M.T. (II) Fatchurrohman, M.Kom

Business processes are achieved by companies that have a long-term vision and a well-thought-out strategy. By making special efforts to satisfy customers, companies can generate long-term profits, save marketing costs, and avoid the risk of losing market share. Customers who are satisfied with the service provided will tend to return to use the service and recommend it to others, thereby increasing customer loyalty and profits for the mall company. Although customers are valuable assets for every mall, some customers can become a burden and harm the mall. Such customers can cause additional costs for the mall in terms of product replacement or re-service. The internal auditor is responsible for ensuring that the mall company operates its business in an optimized manner and by the strategic objectives that have been set. Therefore, *K-Means* can be a solution method in overcoming mall customer problems because *K-Means* is a *data mining* algorithm to group or classify "N" objects based on their attributes or features into "K" groups (called clusters) by minimizing the sum of squared distances between the data and the *centroid* of the corresponding *cluster*. In the *K-Means* algorithm process, the mall customer dataset training is carried out to obtain the optimal *cluster* "k" value, then testing the dataset with the *K-Means* algorithm process flow and measuring the best "k" accuracy with *Silhouette Score* visualization. The results of testing the K-Means algorithm in the classification method of mall customer spending can increase the accuracy rate by 0.476 with the Standard Scale of the *Silhouette Score* calculation. Then in the *Elbow Method*, the research shows the best "K" results fall to k = 3 or the best *Cluster* is *Cluster* 4 type "*Sensible*" with the number of customers reaching 22 people. In the accuracy gathered from the calculation, it is stated that the programming of the developed application system can complement the *K-Means* algorithm perfectly.

**Key Words**: Mall, Customer, *K-Means, Cluster, Elbow Method, Silhouette Score*

# خلاصة

كاهيونو ، غيفاري دوي. 2023. تنفيذ خوارزمية K-Means على تصنيف نفقات التسوق لعملاء المول. اطروحه. برنامج دراسة هندسة المعلوماتية ، كلية العلوم والتكنولوجيا ، جامعة مولانا مالك إبراهيم الإسلامية الحكومية ، مالانج. المشرف: (I) د. م. فيصل، م. ت. (2) فتح الرحمن، م. كوم.

الكلمات الدالة: مول ، عميل ، K-Means ، كتلة ، طريقة الكوع ، درجة صورة ظلية.

يتم تحقيق العمليات التجارية من قبل الشركات التي لديها رؤية طويلة الأجل واستراتيجية ناضجة. من خلال بذل جهود خاصة لإرضاء العملاء ، يمكن للشركات تحقيق أرباح طويلة الأجل ، وتوفير تكاليف التسويق ، وتجنب مخاطر فقدان حصتها في السوق. يميل العملاء الراضون عن الخدمات المقدمة إلى العودة لاستخدام الخدمة والتوصية بها للآخرين ، وذلك لزيادة ولاء العملاء والأرباح لشركات المول. على الرغم من أن العملاء هم أصول قيمة لكل مركز تجاري ، إلا أن هناك بعض العملاء الذين يمكن أن يصبحوا عبئا ويلحقون الضرر بالمركز التجاري. يمكن للعملاء مثل هذا أن يتسببوا في تكاليف إضافية للمركز التجاري في حالة استبدال المنتج أو إعادة الخدمة. المدققون الداخليون مسؤولون عن ضمان قيام شركات مراكز التسوق بتشغيل أعمالها بطريقة مثالية ووفقا للأهداف الاستراتيجية المحددة لذلك ، يمكن أن تكون K-Means طريقة حل للتغلب على مشاكل عملاء المركز التجاري لأن K-Means هي خوارزمية استخراج بيانات لتجميع أو تصنيف كائنات "N" بناء على سماتها أو ميزاتها في مجموعات "K" (تسمى المجموعات) عن طريق تقليل مجموع المسافات التربيعية بين البيانات ومجموعة centroid المقابلة. في عملية خوارزمية K-Means ، تم إجراء تدريب على مجموعة بيانات عملاء المركز التجاري للحصول على القيمة المثلى للمجموعة "k" ، ثم تم إجراء اختبار مجموعة البيانات باستخدام تدفق عملية خوارزمية K-Means وقياس أفضل دقة "k" باستخدام تصور Silhouette Score. يمكن أن تؤدي نتائج اختبار خوارزمية K-Means في طريقة تصنيف نفقات التسوق لعملاء المركز التجاري إلى زيادة معدل الدقة بمقدار 0.476 باستخدام المقياس القياسي من حساب Silhouette Score. ثم في طريقة الكوع ، تظهر الأبحاث أن أفضل النتائج "K" تقع إلى k = 3 أو أفضل مجموعة ، وهي المجموعة 4 من النوع "المعقول" مع وصول عدد العملاء إلى 22 شخصا. في الدقة التي تم جمعها من هذه الحسابات ، يذكر أن برمجة نظام التطبيق التي تم تطويرها يمكن أن تكمل خوارزمية K-Means بشكل مثالي.
.

# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Proses bisnis dengan pemasaran produk saat ini sangat ramai dalam pengawasan pertumbuhan ekonomi. Dengan komoditas perlokasi, dimana pendapatan setiap orang juga bermacam – macam. Proses bisnis dicapai oleh perusahaan yang memiliki visi jangka panjang dan strategi yang matang. Dalam status jumlah toko ritel yang dikemukakan oleh *Euromonitor*, jumlah toko ritel di Indonesia sebanyak 3,98 juta unit (Bayu, D., 2020). Data ini menunjukkan penurunan yang sudah stabil karena sebelumnya mempunyai rentang yang kecil sehingga tidak mempengaruhi penurunan perekonomian. Mereka menyadari bahwa dalam dunia bisnis yang kompetitif, mereka harus terus beradaptasi dengan perubahan pasar dan mencari peluang baru untuk tumbuh. Salah satu cara untuk mempertahankan atau memperluas posisi mereka adalah dengan melakukan inovasi. Inovasi dapat melibatkan pengembangan produk baru, peningkatan proses produksi, atau bahkan menciptakan pasar baru. Perusahaan - perusahaan ini menyadari bahwa tanpa inovasi, mereka akan tertinggal oleh pesaingnya dan kemungkinan besar akan kehilangan pangsa pasar.

Solusi pada inovasi perusahaan tersebut terkait dengan Wahyu milik Allah SWT terkandung pada Q.S. Al-Baqarah/2 ayat 254 yang berbunyi :

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اَنْفِقُوْا مِمَّا رَزَقْنٰكُمْ مِّنْ قَبْلِ اَنْ يَّأْتِيَ يَوْمٌ لَّا بَيْعٌ فِيْهِ وَلَا خُلَّةٌ وَّلَا شَفَاعَةٌۗ وَالْكٰفِرُوْنَ هُمُ الظّٰلِمُوْنَ

*"Wahai orang-orang yang beriman, infakkanlah sebagian dari rezeki yang telah Kami anugerahkan kepadamu sebelum datang hari (Kiamat) yang tidak ada (lagi) jual beli padanya (hari itu), tidak ada juga persahabatan yang akrab, dan tidak ada pula syafaat. Orang-orang kafir itulah orang-orang zalim." (Q.S. Al-Baqarah : 254)*

Pada ayat tersebut mengingatkan umat manusia bahwa rezeki yang mereka peroleh diperlukan untuk kegiatan silaturahmi sesama manusia termasuk jual beli. Karena akan datang hari dimana sudah binasa seluruh umat manusia dan tidak ada pertolongan pada siapapun sehingga untuk menghargai waktu dan menjalin tali silaturahmi dengan sebagian rezeki perolehannya.

Selain dari ayat *Mushaf*, dari Hadits Riwayat Ibnu Majah tentang jual beli atau transaksi dilakukan dengan sukarela dan tidak ada paksaan. Maka *hadist* tersebut menyatakan :

إِنَّمَا الْبَيْعُ عَنْ تَرَاضٍ

*"Sesungguhnya jual beli (harus) atas dasar saling ridha (suka sama suka)." (HR. Ibnu Majah no. 2185, dan dishahihkan oleh Syaikh Muhammad Nashiruddin Al Albani dalam Irwa' al-Ghalil 5/125.)*

Mempertahankan pelanggan merupakan hal penting bagi setiap perusahaan mall. Dengan melakukan upaya khusus untuk memuaskan pelanggan, perusahaan dapat menghasilkan keuntungan jangka panjang, menghemat biaya pemasaran, dan menghindari risiko kehilangan pangsa pasar. Oleh karena itu, perusahaan *mall* harus memberikan perhatian yang serius terhadap pelanggan mereka agar tetap loyal dan puas. Pelanggan yang merasa puas dengan layanan yang diberikan cenderung kembali menggunakan layanan tersebut dan merekomendasikan kepada

orang lain, sehingga dapat meningkatkan loyalitas pelanggan dan keuntungan bagi perusahaan. Maka penting bagi perusahaan untuk meningkatkan kualitas pelayanan agar dapat meningkatkan kepuasan pelanggan (Kastanya, J., 2023).

Meskipun pelanggan adalah aset berharga bagi setiap perusahaan, ada beberapa pelanggan yang dapat menjadi beban dan merugikan perusahaan. Salah satu contoh pelanggan yang tidak menguntungkan adalah pelanggan yang sering mengajukan komplain atau mengembalikan barang. Pelanggan seperti ini dapat menyebabkan biaya tambahan bagi perusahaan dalam hal penggantian produk atau pelayanan ulang. Bagi pemilik produk, itu dapat merusak reputasi perusahaan dengan memberikan ulasan negatif di media sosial atau *platform review online*. Selain itu, ada juga pelanggan yang hanya membeli produk atau menggunakan jasa pada saat diskon besar-besaran atau penjualan musiman. Mereka tidak loyal terhadap merek atau perusahaan tertentu dan hanya mencari harga termurah. Pelanggan semacam ini cenderung tidak memberikan kontribusi signifikan terhadap pendapatan jangka panjang perusahaan. Untuk itu ada misi penting dari auditor internal perusahaan adalah melakukan pemeriksaan apakah proses bisnis saat ini secara efisien dan konsisten mempertahankan pelanggan yang diinginkan. *Auditor internal* bertanggung jawab untuk memastikan bahwa perusahaan mengoperasikan bisnisnya dengan cara yang optimal dan sesuai dengan tujuan strategis yang telah ditetapkan. Pemeriksaan yang dilakukan oleh *auditor internal* melibatkan analisis mendalam terhadap setiap aspek proses bisnis, mulai dari pengelolaan persediaan, produksi, pemasaran, hingga layanan pelanggan. Tujuannya adalah untuk mengidentifikasi kelemahan atau ketidaksempurnaan

dalam sistem yang mungkin menghambat efisiensi operasional dan kemampuan perusahaan dalam mempertahankan pelanggan.

Untuk kemudahan dalam identifikasi pelanggan pada permasalahan yang dibahas, klasifikasi merupakan hal penting dalam pengendalian presentasi pengeluaran belanja dari pelanggan. Identifikasi pelanggan adalah langkah awal yang harus dilakukan untuk memahami kebutuhan dan preferensi mereka. Dengan mengklasifikasikan pelanggan berdasarkan karakteristik tertentu, perusahaan dapat lebih mudah mengidentifikasi dan memahami kelompok-kelompok pelanggan yang berbeda. Manfaat utama dari klasifikasi pelanggan adalah kemudahan dalam mengidentifikasi masalah atau permasalahan yang dihadapi oleh kelompok-kelompok tertentu. Dengan mengetahui karakteristik dan preferensi pelanggan, perusahaan dapat dengan cepat mengetahui apakah ada masalah umum yang dialami oleh kelompok tersebut. Misalnya, jika sekelompok pelanggan memiliki tingkat kepuasan yang rendah, perusahaan dapat melakukan analisis lebih lanjut untuk menemukan penyebabnya dan mencari solusi yang tepat. Dalam pengendalian presentasi pengeluaran belanja dari pelanggan, klasifikasi juga membantu dalam mengatur anggaran secara efisien. Dengan mengetahui jumlah dan jenis produk atau layanan yang dibeli oleh setiap kelompok pelanggan, perusahaan dapat mengalokasikan sumber daya dengan lebih baik. Misalnya, jika sekelompok pelanggan cenderung membeli produk tertentu secara reguler, perusahaan dapat mengatur persediaan dan produksi dengan lebih efisien. Setiap metode klasifikasi memiliki kelebihan dan kekurangan masing-masing, sehingga pemilihan metode klasifikasi yang tepat sangat penting untuk menghasilkan model

yang akurat dan efektif dalam memecahkan masalah klasifikasi (Yudianto, M.R.A. dkk, 2020).

Untuk saat ini, klasifikasi pelanggan yang tergantung pada pengeluaran pendapat masih belum merata. Hal ini dapat dilihat dari perbedaan yang signifikan antara kelompok pelanggan dengan tingkat pengeluaran yang tinggi dan rendah. Ketidakmerataan dalam klasifikasi pelanggan berdasarkan pengeluaran pendapatan juga berdampak pada kesenjangan sosial dan ekonomi dalam masyarakat. Kelompok pelanggan dengan pengeluaran rendah cenderung terjebak dalam lingkaran kemiskinan karena sulit untuk meningkatkan taraf hidup mereka tanpa akses ke produk dan layanan berkualitas. Hingga dalam statistik upah yang dicapai masih belum sesuai. Di saat yang lain, manajemen dalam pemeliharaan keuangan pada perusahaan *mall* juga terkendala berkat ketidakmerataan klasifikasi pelanggan. Perusahaan butuh penyesuaian bagaimana klasifikasi pengeluaran belanja dapat menghasilkan data tentang pernyataan pelanggan terhadap jumlah pengeluaran belanja agar sesuai dengan identifikasi kelompok pelanggan tersebut.

Maka dari itu, untuk menemukan solusi permasalahan ini dibutuhkan sistem kecerdasan klasifikasi yang bisa mengidentifikasi jenis pelanggan sesuai pengeluaran belanja dan dapat digunakan pada seluruh identifikasi lainnya. Dengan adanya sistem kecerdasan tersebut diharapkan dapat menjadi indikasi yang tepat untuk mengatur perilaku pelanggan dalam perputaran kebutuhan taraf hidup. Disini, penulis menerapkan metode *k-means* dikarenakan metode tersebut sangat efektif dalam keputusan *segmentasi* dalam penerapan strategi dalam penyesuaian pengeluaran belanja tersebut. Sementara itu, *k-means* dapat menjadi solusi metode

dalam mengatasi *data mining* karena *k-means* adalah algoritma yang intuitif dan mudah dipahami yang dapat mengklasifikasikan data yang belum diklasifikasikan (yang disebut sebagai *input query*) berdasarkan kesamaannya atau jaraknya dengan data dalam dataset pelatihan. Data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama, sedangkan data dengan karakteristik yang berbeda dikelompokkan dalam *cluster* yang berbeda. *K-Means* dapat digunakan untuk berbagai aplikasi dalam *data mining*, seperti *clustering* karakter dan lain - lain. Sehingga dalam mencapai hasil yang efektif pada penelitian tersebut, pengolahan metode yang dianjurkan adalah survei data pada pelanggan untuk optimasi data yang terbukti. Bila telah memenuhi anjuran, maka proses klasifikasi dapat digunakan sebagaimana penelitian sebelumnya (Putra, B.Y. dkk, 2023).

## 1.2 Pernyataan Masalah

Pada pokok permasalah tersebut terdapat rumusan masalah sebagai berikut:

"Bagaimana penggunaan algoritma *k-means clustering* dalam menentukan klasifikasi pengeluaran belanja pada pelanggan *mall*?"

## 1.3 Tujuan Penelitian

Menerapkan algoritma *k-means clustering* dari sisi klasifikasi pengeluaran belanja pada pelanggan *mall.*

## 1.4 Manfaat Penelitian

Sistem ini dapat dimanfaatkan oleh perusahaan dalam perhitungan rasio keuntungan dari pengeluaran pengunjung.

## 1.5 Batasan Masalah

Penelitian ini memiliki batasan yang dilakukan sebagai berikut :

1. Data sampel berupa pelanggan *mall* tercatat pada 200 *customer*.
2. Perhitungan algoritma *K-Means* berdasarkan pengeluaran belanja dengan mata uang *dollar* ($).

## 1.6 Sistematika Penulisan

Dapat mengetahui urutan juga susunan bagian-bagian penting pada penelitian yang dilakukan. Pada penelitian ini terdiri dari lima bab antara lain yaitu:

**BAB I: PENDAHULUAN**

Berisi latar belakang dari masalah terkait penelitian, tujuan dan manfaat dari penelitian, batasan masalah, metodologi dan sistematika penulisan laporan penelitian.

**BAB II: STUDI PUSTAKA**

Berisi tentang penjelasan tentang penelitian atau teori serta data-data sebelumnya yang terkait tentang penelitian yang mencakup studi dari statistik dataset pelanggan *mall*, penelitian pada metode *k-means* serta teori-teori dari penelitian terdahulu yang dapat membantu penelitian ini.

**BAB III: DESAIN DAN IMPLEMENTASI**

Berisi tentang pembentukan desain dan proses sistem manual tentang metode *k-means* yang telah dirancang pada sistem yang telah dibuat, dan membahas secara rinci terhadap proses dan langkah-langkah implementasi yang dilakukan.

**BAB IV: UJI COBA DAN HASIL**

Berisi tentang hasil pengujian dan eksperimen tentang metode *k-means* yang telah dirancang pada sistem yang telah buat, dan membahas secara rinci terhadap hasil dan proses yang dilakukan.

**BAB V: KESIMPULAN DAN SARAN**

Berisi kesimpulan dan saran dari penelitian yang telah dilakukan, untuk tujuan agar dapat dikembangkan lagi pada penelitian selanjutnya.

# BAB II

# STUDI PUSTAKA

Dalam menerapkan *machine learning* yang berkaitan dengan populasi, terutama pada kasus karakteristik pelanggan *mall* ternyata ada banyak permasalahan dalam perhitungan yang sempurna. Karena untuk perhitungan yang terkait di bidang statistik membutuhkan tahap – tahap dengan skala yang luas. Perhitungan bisa saja berbeda tergantung produk perusahaan seperti potongan harga, jumlah stok yang tersisa, jenis produk yang dibutuhkan, dan lain sebagainya. Untuk itu, diprioritaskan pada pengunjung *mall* yang tentu saja ada berbagai produk yang tersedia. Tentu dalam menganalisis tempat tersebut disertai metode klasifikasi yang akurat serta dengan mengetahui jumlah pengeluaran belanja tersebut.

*Machine learning* adalah salah satu cabang kecerdasan buatan yang memungkinkan komputer untuk belajar dari data dan pengalaman tanpa harus diprogram secara eksplisit. Dalam konteks populasi, dapat dikatakan pada data pengunjung harus diperhatikan ciri – ciri dari setiap data untuk menjadi acuan tipe data/ atribut. Acuan tersebut sebenarnya pada klasifikasi juga dibutuhkan untuk menunjang keberhasilan dalam pengolahan data.

## 2.1 Studi Pustaka

Penelitian yang dilakukan ini akan merujuk ke sumber para peneliti dengan obyek yang sama. Dalam penelitian terkait, kebanyakan pembahasan yang dikemukakan bersifat teoritis yang menyatakan fakta – fakta penelitian ini didasarkan dengan pembuktian metode dengan sistem pengaplikasian yang bisa

digunakan sebagai media dari metode tersebut. Pembuktian metode merupakan langkah penting dalam penelitian karena hal ini menjamin bahwa hasil penelitian dapat dipercaya dan diandalkan. Metode tersebut harus memiliki sistem pengaplikasian yang jelas sehingga dapat digunakan sebagai media dalam proses pengumpulan data. Selain itu, pembuktian metode juga melibatkan uji coba atau eksperimen untuk memastikan bahwa metode tersebut efektif dalam menghasilkan data yang akurat. Hal ini penting agar hasil penelitian dapat diinterpretasikan dengan benar dan memberikan kontribusi signifikan terhadap bidang ilmu tertentu.

Namun demikian, tidak semua pembahasan dalam penelitian bersifat teoritis. Ada juga beberapa studi empiris yang lebih fokus pada analisis data daripada membahas teori-teori dasar. Meskipun demikian, baik pendekatan teoritis maupun empiris memiliki kelebihan dan kekurangan masing-masing. Oleh karena itu, penting bagi para peneliti untuk memilih metode yang sesuai dengan tujuan penelitian mereka dan mempertimbangkan keterbatasan yang ada.

Dalam kesimpulannya, pembahasan dalam penelitian terkait umumnya bersifat teoritis yang didasarkan pada pembuktian metode dengan sistem pengaplikasian yang bisa digunakan sebagai media dari metode tersebut. Pembuktian metode ini penting untuk memastikan validitas dan reliabilitas hasil penelitian. Namun, tidak semua pembahasan dalam penelitian bersifat teoritis, ada juga studi empiris yang lebih fokus pada analisis data. Penting bagi para peneliti untuk memilih metode yang sesuai dengan tujuan penelitian mereka dan mempertimbangkan keterbatasan yang ada. Adapun sistem – sistem yang sesuai

tema teknologi informasi yaitu aplikasi yang menerapkan operasi metode dalam mengolah obyek terbatas berdasarkan apa yang diteliti.

(Safitri, Cholissodin, dan Muflikhah 2018) melakukan penelitian sebelumnya, "Pemetaan Potensi Pelanggan Sebagai Strategi Promosi Pakaian Menggunakan Algoritma *K-Means Clustering*". Ini disebabkan oleh rencana promosi yang tidak jelas, yang dapat menyebabkan kerugian jika tidak dilakukan dengan benar. Menurut penelitian ini, tujuan dari penelitian ini adalah untuk membagi pelanggan potensial berdasarkan wilayah atau kecamatan; kelompok 1 memiliki 3 kecamatan, kelompok 2 memiliki 7 kecamatan, dan kelompok 3 memiliki 13 distrik. Hasil tersebut diperkuat oleh pengujian software *rapidminer*, yang menguji akurasi data berdasarkan hasil perhitungan dari 23 data.

"Algoritma *K-Means* Dengan Metode *Elbow* Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia" adalah penelitian sebelumnya yang dilakukan oleh (Rina Yuliana Sari, Hardian Oktavianto, dan Henny Wahyu Sulistyo 2018). Dengan menggunakan metode *clustering*, penelitian ini bertujuan untuk membantu pemerintah mengidentifikasi masalah dan mempertimbangkan pengambilan kebijakan pada wilayah kabupaten/kota di provinsi Jawa Tengah berdasarkan variabel - variabel IPM. Penelitian ini menggunakan metode *k-means*, yang merupakan algoritma yang berfungsi dengan baik untuk menganalisis data yang sangat besar.

(Augusto Luis Ballardini 2018) melakukan penelitian sebelumnya dalam "Tutorial tentang *Clustering Particel Swarm Optimization*". Penelitian ini

mengusulkan penggunaan *particle swarm optimization* untuk memperbaiki cluster yang dibuat oleh algoritma *k-means*. Hasil penelitian ini menunjukkan bahwa algoritma *PSO* lebih baik daripada algoritma *k-means*. Ini karena algoritma *PSO* bekerja lebih lambat tetapi memiliki kesalahan kuantisasi yang lebih rendah, sedangkan algoritma *k-means* bekerja lebih cepat tetapi memiliki kesalahan kuantisasi yang lebih tinggi.

"Apakah Pelanggan Ditawari Diskon yang Sesuai? Studi Eksplorasi Menggunakan Teknik *Clustering* dalam Audit Internal" adalah penelitian sebelumnya oleh (Jun Dai, Paul Byrnes, dan Miklos Vasarhely 2019). Menurut penelitian ini, tujuan penerapan metode *k-means* untuk menentukan kebijakan diskon konsumen. Teknik *clustering* digunakan untuk mengidentifikasi karakteristik pelanggan untuk mencapai tujuan penelitian ini. Setelah mengurangi dimensi, menyimpan, dan normalisasi data, algoritma *k-means* digunakan untuk membuat kelompok berdasarkan enam atribut utama. Hasilnya adalah tujuh kategori yang sangat baik untuk pelanggan kartu kredit, dimulai dengan yang paling disukai hingga yang paling tidak disukai. Oleh karena itu, kebijakan diskon yang lebih masuk akal dan relevan harus dibuat untuk membedakan harga untuk berbagai jenis pelanggan.

Penelitian tentang "*Clustering* Pengunjung *Mall* Menggunakan Metode *K-Means* dan *Particle Swarm Optimization*" Oleh Teuku Muhammad Dista dan Ferian Fauzi Abdulloh (2022) yaitu penelitian yang dilakukan untuk mengklaster pengunjung *mall* menggunakan metode *k-means* dan *particle swarm optimization*. Hasilnya yaitu penggunaan metode *particle swarm optimization* (*PSO*) untuk

mengoptimasi hasil clustering yang dihasilkan oleh metode *k-means* dapat meningkatkan akurasi hasil *clustering*. Evaluasi menggunakan *davies bouldin index* (*DBI*) menunjukkan bahwa hasil clustering yang dihasilkan oleh metode *k-means* yang dioptimasi dengan *PSO* lebih baik dibandingkan dengan yang belum dioptimasi, meskipun kinerja algoritmanya lebih lambat. Penelitian ini juga berhasil mengidentifikasi 5 karakteristik pengunjung *mall* yang dapat digunakan untuk membantu pihak manajemen dan kreatif marketing dalam meningkatkan pendapatan *mall*.

Tabel 2.1 Hasil penelitian sebelumnya

| Nama | Judul | Hasil Analisis |
|---|---|---|
| Mardalius dan Tika Christy | Pemetaan Potensi Pelanggan Sebagai Strategi Promosi Pakaian Menggunakan Algoritma *K-Means Clustering* | Dapat membagi pelanggan potensial berdasarkan wilayah atau kecamatan; kelompok 1 memiliki 3 kecamatan, kelompok 2 memiliki 7 kecamatan, dan kelompok 3 memiliki 13 distrik dengan perhitungan 23 data |
| Rina Yuliana Sari dkk | Algoritma *K-Means* Dengan Metode *Elbow* Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia | Mengidentifikasi masalah dan mempertimbangkan pengambilan kebijakan pada wilayah kabupaten/kota di provinsi Jawa Tengah berdasarkan variabel variabel IPM |
| Augusto Luis Ballardini | Tutorial tentang *Particle Swarm Optimization Clustering* | Algoritma *PSO* bekerja lebih lambat tetapi memiliki kesalahan kuantisasi yang lebih rendah, sedangkan algoritma *k-means* bekerja lebih cepat tetapi memiliki kesalahan kuantisasi yang lebih tinggi. |
| Jun Dai dkk | Apakah Pelanggan Ditawari Diskon yang Sesuai? Studi Eksplorasi Menggunakan | Mengidentifikasi karakteristik pelanggan untuk mencapai tujuan penelitian ini yaitu mengurangi dimensi, menyimpan, dan normalisasi data, dan membuat kelompok berdasarkan enam atribut utama. Hasilnya |

| Nama | Judul | Hasil Analisis |
|---|---|---|
| | Teknik *Clustering* dalam Audit Internal | adalah tujuh kategori yang sangat baik untuk pelanggan kartu kredit |
| Teuku Muhammad Dista dan Ferian Fauzi Abdulloh | *Clustering* Pengunjung *Mall* Menggunakan Metode *K-Means* dan *Particle Swarm Optimization* | Penggunaan metode *particle swarm optimization* (*PSO*) untuk mengoptimasi hasil *clustering* yang dihasilkan oleh metode *k-means* dapat meningkatkan akurasi hasil clustering. Evaluasi menggunakan *davies bouldin index* (*DBI*) menunjukkan bahwa hasil clustering yang dihasilkan oleh metode *k-means* yang dioptimasi dengan *PSO* lebih baik dibandingkan dengan yang belum dioptimasi, meskipun kinerja algoritmanya lebih lambat. |

Tabel 2.1 merupakan hasil perumusan jurnal – jurnal pada peneliti sebelumnya untuk meringkaskan sebagian penjelasan agar dapat diidentifikasi bagaimana penggunaan dan analisis algoritma *k-means* tersebut. Dari kelima jenis jurnal yang disampaikan adakalanya bisa mengatasi keterbatasan penelitian seperti tempat survei dilaksanakan, objek yang dipakai, hasil perhitungan program, dan sebagainya.

Tabel 2.2 Perbandingan Penelitian pada Metode *K-Means*

| Penelitian Terdahulu | | Penelitian Kelanjutan | |
|---|---|---|---|
| Mardalius dan Tika Christy | Pembagian data potensial | Penelitian ini | Pembagian dengan data penghasilan dan pengeluaran |
| Rina Yuliana Sari dkk | Penentuan kebijakan klasifikasi wilayah | | Penentuan klasifikasi pengeluaran pelanggan |
| Augusto Luis Ballardini | Perbandingan 2 algoritma dengan kesalahan kuitansi | | Penggunaan 1 algoritma untuk kinerja kuitansi |

| Penelitian Terdahulu | | Penelitian Kelanjutan | |
|---|---|---|---|
| Jun Dai dkk | Melakukan normalisasi data berdasarkan 6 atribut | | Penyetaraan data dengan nilai *cluster* |
| Teuku Muhammad Dista dan Ferian Fauzi Abdulloh | Perbandingan optimasi 2 algoritma | | Optimasi dengan *elbow* |

Tabel 2.2 adalah perbandingan antara penelitian terdahulu dengan penelitian pada artikel yang dibahas ini. Perbandingan yang dimaksud yaitu bagaimana penelitian yang terdahulu menggunakan algoritma *k-means* dengan objek yang sudah didapatkan. Sehingga hasil tersebut akan ditindaklanjuti perkembangan objek yang diteliti dan hasil dari perkembangan projek tersebut.

## 2.2 Kajian Teoritis

Menilik sumber yang didapatkan dari pencarian artikel dari pendapat para ilmuwan, beberapa teori tentang suatu penelitian ini membahas bagaimana metode *k-means* bekerja dengan mengandalkan sistem informasi dengan pengolahan data mencakup di *platform* komputer. Dasar – dasar teori seperti definisi, langkah – langkah, serta hubugan antar teori – teori tersebut akan diperjelaskan secara teliti. Penjelasan sumber – sumber yang didapatkan digunakan sebagai perbandingan hasil penelitian ini dengan penelitian dari sumber – sumber pilihan untuk kelayakan percobaan metode tersebut. Dari perihal tersebut, *machine learning*, metode klasifikasi, dan *k-means* yang menjadi sorotan untuk mengidentifikasi tujuan pembuatan proyek hasil penelitian yang dilakukan dalam artikel ini.

### 2.2.1 *Machine Learning*

(Russell, R. 2018) *Machine Learning* mengacu pada penggunaan algoritma dan model statistik oleh sistem komputer untuk melakukan tugas tertentu tanpa menggunakan instruksi eksplisit. Hal ini melibatkan kemampuan mesin untuk belajar dan berkembang dari pengalaman, sehingga memungkinkan mereka untuk membuat prediksi atau keputusan berdasarkan data. Tujuan dari pembelajaran mesin adalah untuk mengembangkan sistem komputer yang dapat belajar dari data, mengidentifikasi pola, dan membuat keputusan atau prediksi tanpa instruksi eksplisit, sehingga memungkinkan mesin untuk meningkatkan kinerjanya dari waktu ke waktu dan menghasilkan hasil yang lebih akurat. Data tersebut diolah dengan sedemikian mungkin dapat menimbulkan pernyataan yang efektif dan sangat mudah dipahami.

(Dista, T.M. 2022) Kita sering melihat alat-alat *machine learning* di berbagai area di sekitar kita, misalnya di *facebook*, pembelajaran mesin membantu kita mengidentifikasi diri kita dan teman-teman kita, bahkan di *youtube* dapat merekomendasikan video berdasarkan apa yang kita suka. Analis data sering menggunakan pembelajaran yang diawasi untuk memecahkan masalah seperti klasifikasi dan regresi, yang berarti bahwa data ini berisi sinyal objektif yang ingin Anda laporkan di masa mendatang, misalnya dalam industri perbankan, analisis data dapat digunakan untuk mengklasifikasikan pelanggan menjadi kelompok risiko tinggi atau rendah berdasarkan riwayat kredit mereka.

(Satinet, C. 2022) *Machine Learning* telah digunakan dalam LCA (*Life Cycle Assessment*) untuk memperkirakan nilai faktor karakterisasi dampak

lingkungan, melakukan analisis sensitivitas, atau mengembangkan kerangka kerja untuk memprediksi dampak lingkungan dari produk di seluruh siklus hidupnya. Pendapat tentang perkiraan, dampak yang dikeluarkan saat analisis dan siklus hidup bisa diprediksi sesuai dengan metode yang sudah dikategorikan oleh peneliti. Bila peneliti mempelajari faktor karakterisasi pada *machine learning*, implementasi dataset bisa saja menjadi sangat dibutuhkan.

*Machine learning* memproses data dengan menggunakan algoritma untuk menganalisis dan belajar dari data input untuk membuat prediksi atau keputusan. Proses ini melibatkan pelatihan model pada kumpulan data berlabel dalam pembelajaran yang diawasi, di mana algoritma belajar memetakan data input ke output yang sesuai. Dalam pembelajaran tanpa pengawasan, algoritma mengidentifikasi pola atau struktur dalam data yang tidak berlabel untuk mengekstrak wawasan yang bermakna. Pembelajaran penguatan melibatkan agen yang belajar untuk membuat keputusan dengan berinteraksi dengan lingkungan dan menerima hadiah atau hukuman berdasarkan tindakannya. Pembelajaran *batch* membutuhkan pelatihan sistem dengan menggunakan semua data yang tersedia sekaligus, sementara pembelajaran *online* memungkinkan sistem untuk belajar secara bertahap dengan diberi contoh data satu per satu.

Sistem *machine learning* juga dapat menangani data dalam jumlah besar dan beradaptasi dengan perubahan, sehingga cocok untuk tugas-tugas seperti pengelompokan, pembelajaran aturan asosiasi, visualisasi, dan pengurangan dimensi. Pengujian dan validasi penting untuk memastikan bahwa model dapat menggeneralisasi dengan baik untuk kasus-kasus baru, dan untuk mengidentifikasi

dan mengatasi masalah seperti *overfitting*. Efektivitas pembelajaran mesin juga bergantung pada kualitas dan relevansi data pelatihan, serta pemilihan dan ekstraksi fitur yang berguna melalui rekayasa fitur.

Proses *machine learning* biasanya dibagi menjadi dua set: set pelatihan dan set pengujian. Set pelatihan digunakan untuk melatih model, sedangkan set pengujian digunakan untuk mengevaluasi kinerja model dan kemampuannya untuk menggeneralisasi data baru yang belum pernah dilihat. Proses ini membantu mengidentifikasi masalah seperti *overfitting*, di mana model berkinerja baik pada data pelatihan tetapi buruk pada data baru, dan *underfitting*, di mana model tidak dapat menangkap pola yang mendasari data. Singkatnya, proses data adalah komponen fundamental dari *machine learning*, karena menjadi dasar untuk mengajarkan model bagaimana membuat prediksi atau keputusan berdasarkan data masukan.

*Data training*, juga dikenal sebagai set pelatihan, mengacu pada bagian dari kumpulan data yang digunakan untuk melatih model pembelajaran mesin. Data ini terdiri dari data masukan bersama dengan keluaran yang benar, atau "label", dalam kasus pembelajaran yang diawasi. *Data training* digunakan untuk mengajarkan model bagaimana membuat prediksi atau keputusan dengan mempelajari pola dan hubungan di dalam data. Kualitas dan relevansi data pelatihan sangat penting untuk keefektifan model. Penting untuk memastikan bahwa data pelatihan bebas dari kesalahan dan pencilan, dan mengandung cukup banyak fitur yang relevan untuk dipelajari oleh model. Selain itu, rekayasa fitur, yang melibatkan pemilihan,

ekstraksi, dan pembuatan fitur baru berdasarkan data, memainkan peran penting dalam mempersiapkan data pelatihan untuk model *machine learning*.

*Data testing*, juga dikenal sebagai set pengujian, adalah bagian dari dataset yang digunakan untuk mengevaluasi kinerja model *machine learning*. Dataset ini terpisah dari dataset pelatihan dan digunakan untuk menilai seberapa baik model dapat menggeneralisasi data baru yang belum pernah dilihat sebelumnya. Set pengujian sangat penting untuk mengidentifikasi masalah seperti *overfitting*, di mana model berkinerja baik pada data pelatihan tetapi buruk pada data baru, dan *underfitting*, di mana model tidak dapat menangkap pola yang mendasari data. Dengan mengevaluasi model pada set pengujian, dimungkinkan untuk mengukur kesalahan generalisasinya, yang mengindikasikan seberapa baik model tersebut diharapkan bekerja pada data baru yang tidak terlihat. Proses ini membantu memastikan bahwa model tersebut efektif dan dapat diandalkan untuk membuat prediksi atau keputusan dalam aplikasi dunia nyata.

Sudah menjadi praktik umum untuk membagi dataset menjadi set pelatihan dan set pengujian, dengan set pelatihan digunakan untuk melatih model dan set pengujian digunakan untuk mengevaluasi kinerjanya. Set pengujian harus mewakili data yang diharapkan akan ditemui oleh model dalam praktiknya, dan tidak boleh digunakan untuk pelatihan untuk menghindari bias pada hasil evaluasi. Set pengujian adalah komponen penting dari proses pembelajaran mesin, karena menyediakan sarana untuk menilai kinerja model dan kemampuannya untuk menggeneralisasi ke kasus-kasus baru.

Peneliti menerapkan *machine learning* dengan langkah-langkah sebagai berikut:

1) Pengumpulan Data: Peneliti mengumpulkan data yang relevan dan representatif untuk masalah yang ingin diselesaikan. Data ini digunakan sebagai *training set* untuk melatih model *machine learning.*
2) Pemrosesan Data: Data yang dikumpulkan kemudian diproses untuk membersihkan data dari kesalahan, *outlier,* dan fitur yang tidak relevan. Proses ini juga melibatkan *feature engineering*, yaitu pemilihan, ekstraksi, dan penciptaan fitur-fitur baru berdasarkan data.
3) Pembagian Data: Data dibagi menjadi dua set, yaitu *training set* dan *testing set*. *Training set* digunakan untuk melatih model, sedangkan *testing set* digunakan untuk mengevaluasi kinerja model.
4) Pemilihan Model: Peneliti memilih model *machine learning* yang sesuai dengan masalah yang ingin diselesaikan, seperti regresi, klasifikasi, atau pengelompokan.
5) Pelatihan Model: Model *machine learning* dilatih menggunakan *training set* untuk belajar pola dan hubungan dalam data.
*6)* Evaluasi Model: Model dievaluasi menggunakan *testing set* untuk mengukur kinerjanya, seperti *generalization error,* dan untuk mengidentifikasi masalah seperti *overfitting* atau *underfitting.*
7) Penyesuaian Model: Jika diperlukan, model disesuaikan dan dilatih ulang dengan menggunakan teknik seperti *cross-validation* untuk meningkatkan kinerjanya.

8) Implementasi Model: Model yang telah dilatih dan dievaluasi kemudian diimplementasikan untuk digunakan dalam aplikasi nyata.

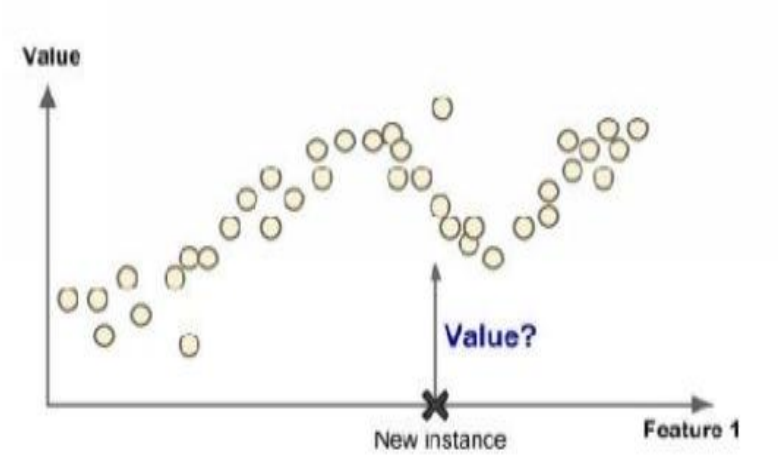
Proses ini memerlukan pemahaman yang mendalam tentang konsep-konsep *machine learning* dan pemrograman, serta penggunaan algoritma dan teknik yang sesuai dengan masalah yang ingin diselesaikan .

Metode ini memungkinkan para ahli data untuk mengidentifikasi pola dan tren dalam set data yang besar. Dengan menggunakan teknik ini, para ahli data dapat mengidentifikasi pola dan tren dalam set data besar dan membuat prediksi akurat tentang hasil masa depan. Selain itu, implementasi juga melibatkan pengumpulan data dalam sebuah set pelatihan, memberikan set tersebut ke algoritma pembelajaran, dan mendapatkan *output* atau prediksi. Proses ini juga melibatkan pemilihan fitur-fitur yang berguna, ekstraksi fitur, dan penciptaan fitur baru berdasarkan data (*feature engineering*). Setelah itu, model yang dihasilkan dapat diuji dengan menggunakan data uji untuk memastikan bahwa model tersebut dapat melakukan generalisasi dengan baik pada kasus-kasus baru (Russell, R. 2018).

(Hyde, K. 2019) *Machine Learning* secara garis besar dapat dipilah menjadi dua kategori, *Supervised Learning* dan *Unsupervised Learning*.

1) *Supervised Learning* : *Supervised Learning* melibatkan algoritma yang menggunakan variabel input untuk memprediksi klasifikasi target (yaitu variabel dependen), yang dapat berupa kategorikal atau kontinu. Jenis pembelajaran mesin ini melibatkan pelatihan model pada kumpulan data berlabel, di mana data *input* dan *output* yang sesuai telah diketahui. Dalam pembelajaran yang diawasi,

data yang dimasukkan ke dalam algoritma, bersama dengan solusi yang diinginkan, disebut sebagai "label". Beberapa algoritma yang diawasi yang penting termasuk *k-nearest neighbors, linear regression, neural networks, support vector machines, logistic regression, decision trees,* dan *random forests* (Russell, R. 2018).




Gambar 2.1 Contoh Grafis *Supervised Learning*. Sumber : (Russell, R. 2018)

Gambar 2.1 di atas adalah contoh grafik *supervised learning* yang baik dimana untuk hal ini karena program ini telah dilatih dengan banyak iterasi yang selaras pada saat yang sama dengan kelasnya. Contoh lainnya adalah memprediksi nilai numerik seperti harga sebuah flat, dengan memberikan sekumpulan fitur (lokasi, jumlah kamar, fasilitas) yang disebut prediktor; jenis tugas ini disebut regresi. Anda harus ingat bahwa beberapa algoritma regresi juga dapat digunakan untuk klasifikasi, dan sebaliknya. Algoritma yang paling penting - *K-nearest neighbor - Linear Regression - Neural Networks - Support Vector Machines - Logistic Regression - Decision Trees,* dan *Random Forests*.

2) *Unsupervised Learning* : Sebaliknya pada *unsupervised learning*, penggunanya tidak selalu memiliki label atau target khusus untuk diprediksi, misalnya *clustering*, berdasarkan model matematisnya, algoritma pada *unsupervised learning* tidak memiliki target variabel. *unsupervised learning* melibatkan kerja dengan data tanpa label, di mana algoritma belajar untuk mengidentifikasi pola atau struktur di dalam data. Dalam *unsupervised learning,* algoritma bekerja dengan mencari pola atau struktur dalam data secara mandiri. Algoritma ini tidak diberikan informasi tentang apa yang harus dicari atau diidentifikasi dalam data tersebut. Sebagai gantinya, algoritma menggunakan teknik-teknik seperti *clustering*, reduksi dimensi, dan asosiasi untuk menemukan pola-pola yang ada dalam data. Tanpa adanya target variabel, algoritma pada *unsupervised learning* dapat mengeksplorasi dan menggali informasi baru dari data secara mandiri. Algoritma ini mampu mengenali kelompok-kelompok atau kategori - kategori yang ada dalam data tanpa bantuan manusia. Namun demikian, kekurangan dari *unsupervised learning* adalah kurangnya interpretasi hasil yang ditemukan oleh algoritma. Karena tidak ada target variabel yang digunakan sebagai acuan, hasil dari *unsupervised learning* seringkali sulit untuk diinterpretasikan secara langsung (Hidayat, S. 2018).

   Jenis-jenis pembelajaran tanpa pengawasan meliputi:

   Pengelompokan: *K-Means,* Analisis Klaster Hirarkis, Pembelajaran aturan asosiasi: *Eclat, Apriori*, Visualisasi dan pengurangan dimensi: *PCA Kernel, T-Distribusi, PCA.*

Algoritma pembelajaran tanpa pengawasan digunakan untuk tugas-tugas seperti pengelompokan, pembelajaran aturan asosiasi, dan visualisasi dan pengurangan dimensi. Dalam sistem pembelajaran mesin jenis ini, Anda dapat menebak bahwa data tidak berlabel. Algoritma tanpa pengawasan yang paling penting - Pengelompokan : *K-Means*, Analisis Klaster Hirarkis - Pembelajaran aturan asosiasi : *Eclat, Apriori* - Visualisasi dan reduksi dimensi : *PCA Kernel, T-Distribusi, PCA.*

### 2.2.2 Metode Klasifikasi

Klasifikasi adalah masalah mengidentifikasi ke dalam kategori mana suatu observasi baru termasuk, berdasarkan pada kumpulan data latihan yang berisi observasi yang keanggotaan kategorinya diketahui. Banyak masalah dunia nyata dapat dimodelkan sebagai masalah klasifikasi, seperti mengelompokkan *email* ke dalam kelas "*spam*" atau "*non-spam*", secara otomatis mengelompokkan kategori - kategori (misalnya, "olahraga" dan "hiburan") dari berita yang akan datang, dan memberikan diagnosis kepada pasien berdasarkan karakteristik yang diamati (jenis kelamin, tekanan darah, kehadiran atau ketiadaan gejala tertentu, dan lain - lain.). Proses umum klasifikasi data terdiri dari dua fase, yaitu fase pelatihan dan fase prediksi. Misalkan pada penelitian sebelumnya, metode klasifikasi didasarkan pada konstruksi kamus berdasarkan fitur-fitur yang dipilih, yang kemudian digunakan untuk mengkategorikan ulasan film ke dalam sentimen positif atau negatif (Pratiwi, A.I. 2018).

Metode klasifikasi memiliki beberapa manfaat, antara lain:

1) Meningkatkan efektivitas pembelajaran: Metode *semi-supervised learning* dan *transfer learning* menggunakan data tambahan, baik berupa data berlabel maupun tidak berlabel, untuk meningkatkan kualitas pembelajaran.
2) Mengurangi *overfitting*: Proses klasifikasi memerlukan pemilihan fitur yang tepat untuk meminimalkan *overfitting,* terutama saat menggunakan data latihan yang terbatas.
3) Beragam aplikasi: Metode klasifikasi dapat diterapkan dalam berbagai domain, seperti pemasaran target pelanggan, diagnosis penyakit medis, deteksi peristiwa, analisis data multimedia, analisis data biologis, kategorisasi dokumen, dan analisis jaringan sosial.
4) Meningkatkan hasil: *Meta-algoritma* dalam klasifikasi dapat digunakan untuk menggabungkan hasil dari beberapa model pembelajaran, sehingga menghasilkan hasil yang lebih andal dan *robust*.
5) Penyesuaian dengan berbagai jenis data: Metode klasifikasi dapat diadaptasi untuk berbagai jenis data, termasuk data teks, multimedia, data jaringan, data deret waktu, dan data probabilitas .

Dengan manfaat-manfaat tersebut, metode klasifikasi menjadi penting dalam berbagai aplikasi dan penelitian di bidang *data mining* dan *machine learning* (Aggarwal, C.C. 2015).

Klasifikasi *data mining* adalah proses menemukan kesamaan karakteristik dalam suatu kelompok atau kelas. Ini adalah salah satu metode yang paling umum digunakan dalam *data mining*. Metode klasifikasi digunakan untuk memperkirakan kelas suatu objek yang labelnya tidak diketahui. Dalam dunia yang semakin

terhubung ini, jumlah data yang dihasilkan setiap hari semakin meningkat dengan cepat. Oleh karena itu, penting bagi kita untuk dapat mengklasifikasikan dan memahami data tersebut agar dapat mengambil keputusan yang tepat. Namun, perlu diingat bahwa proses klasifikasi juga memiliki batasan dan tantangan tersendiri. Salah satu tantangannya adalah memilih algoritma yang tepat untuk jenis data tertentu. Selain itu, interpretasi hasil juga bisa menjadi sulit jika terdapat banyak variabel atau atribut dalam dataset.

Fase pelatihan pada klasifikasi merupakan tahap di mana model klasifikasi dibangun dari data latihan. Pada fase ini, data dianalisis menjadi kumpulan fitur berdasarkan model pembentukan fitur, seperti model ruang vektor untuk data teks. Fitur-fitur ini dapat berupa kategori, *ordinal*, nilai *integer*, atau nilai *real*, dan beberapa algoritma memerlukan data diskrit untuk bekerja. Setelah mewakili data melalui fitur-fitur yang diekstraksi, algoritma pembelajaran menggunakan informasi label dan data itu sendiri untuk mempelajari fungsi pemetaan dari fitur ke label. Ini dilakukan untuk membangun model klasifikasi yang dapat memprediksi label dari data yang belum terlihat.

Selama fase pelatihan, penting untuk melakukan seleksi fitur yang tepat karena fitur yang tidak relevan dapat mempengaruhi akurasi klasifikasi dan menyebabkan *overfitting*, terutama saat menggunakan data latihan yang terbatas. Oleh karena itu, pemilihan fitur yang tepat sangat penting untuk memastikan kualitas model klasifikasi yang dihasilkan. Dengan demikian, fase pelatihan pada klasifikasi melibatkan pembentukan model klasifikasi dari data latihan dengan mempelajari hubungan antara fitur-fitur dan label-label yang terkait.

Fase pengujian data pada klasifikasi merupakan tahap di mana model klasifikasi yang telah dibangun dari data latihan digunakan untuk memprediksi label dari data uji yang belum terlihat. Pada fase ini, data diwakili oleh kumpulan fitur yang diekstraksi dalam proses pelatihan, dan kemudian model klasifikasi yang telah dipelajari dari fase pelatihan digunakan untuk memprediksi label dari data uji berdasarkan fitur-fiturnya .

Ada dua cara umum di mana hasil dari algoritma klasifikasi dapat disajikan untuk data uji. Pertama, label diskrit dapat diberikan untuk data uji, yang berarti bahwa data uji akan diberi label berdasarkan prediksi model klasifikasi. Kedua, skor numerik dapat diberikan untuk setiap kombinasi label kelas dan data uji, yang memungkinkan perbandingan relatif dari kecenderungan berbeda dari data uji untuk menjadi bagian dari kelas tertentu .

Selain itu, ada metode klasifikasi yang disebut sebagai metode pembelajaran berbasis instansi, di mana fase konstruksi model pelatihan sering kali dihilangkan. Pada metode ini, data uji langsung terkait dengan instansi pelatihan untuk membuat model klasifikasi. Metode ini disebut sebagai metode pembelajaran malas, karena mereka menunggu pengetahuan tentang instansi uji untuk membuat model yang dioptimalkan secara lokal, yang spesifik untuk instansi uji. Pada metode klasifikasi antara *rule based* dan *random forest* banyak sekali perbandingannya karena performa pengklasifikasi *rule based* dibandingkan dengan pengklasifikasi tradisional seperti *random forest* (Lee, S. 2022).

Proses klasifikasi dimulai dengan pengumpulan dan pemilihan data yang relevan. Kemudian, algoritma klasifikasi digunakan untuk menganalisis pola-pola

dalam data tersebut. Algoritma ini mencari kesamaan karakteristik antara objek-objek dalam kelompok atau kelas tertentu. Misalnya, jika kita ingin mengklasifikasikan email sebagai spam atau bukan spam, algoritma mencari pola-pola seperti kata-kata tertentu atau tautan yang sering muncul dalam email spam. Hasil dari proses klasifikasi ini sangat berharga karena dapat membantu kita membuat keputusan yang lebih baik. Misalnya, hasil klasifikasi dapat digunakan untuk memprediksi perilaku pelanggan atau mengidentifikasi penipuan keuangan. Ada beberapa metode klasifikasi yang umum digunakan dalam data mining, antara lain regresi logistik, *naïve bayes, decision tree, random forest, k-nearest neighbor*, dan jaringan syaraf tiruan. Proses klasifikasi dilakukan dengan cara belajar dari data yang sudah ada dan kemudian mengklasifikasikan data baru, dengan hasil dari metode klasifikasi berupa kategorikal (*nominal* atau *ordinal*).

Proses klasifikasi data umumnya melibatkan beberapa langkah, yaitu:

1) Seleksi Fitur: Langkah pertama dari hampir semua algoritma klasifikasi adalah seleksi fitur. Pemilihan fitur yang tepat sangat penting karena fitur yang tidak relevan dapat mempengaruhi akurasi klasifikasi dan menyebabkan *overfitting*, terutama saat menggunakan data latihan yang terbatas.
2) Pembentukan Fitur: Pada fase pelatihan, data dianalisis menjadi kumpulan fitur berdasarkan model pembentukan fitur, seperti model ruang vektor untuk data teks. Fitur-fitur ini dapat berupa kategori, *ordinal*, nilai *integer*, atau nilai *real*, dan beberapa algoritma memerlukan data diskrit untuk bekerja.
3) Pembelajaran: Pada fase pelatihan, algoritma pembelajaran menggunakan informasi label dan data itu sendiri untuk mempelajari fungsi pemetaan dari fitur

ke label. Ini dilakukan untuk membangun model klasifikasi yang dapat memprediksi label dari data yang belum terlihat.

4) Klasifikasi: Pada fase prediksi, model klasifikasi yang telah dipelajari dari fase pelatihan digunakan untuk memprediksi label dari data baru berdasarkan fitur-fiturnya.
5) Evaluasi: Setelah proses klasifikasi, model yang telah dibangun dievaluasi untuk mengukur kinerjanya. Evaluasi ini dapat dilakukan dengan menggunakan metrik seperti akurasi, presisi, *recall*, atau *F1-score*.

Proses klasifikasi ini sering kali disebut sebagai pembelajaran terawasi karena pengelompokan data baru didasarkan pada pengetahuan sebelumnya yang terdapat dalam data latihan.

Metode klasifikasi dapat dikategorikan berdasarkan beberapa pendekatan, yaitu:

1) Berbasis Teknik: Klasifikasi dapat menggunakan berbagai jenis teknik seperti pohon keputusan, metode berbasis aturan, jaringan saraf tiruan, metode *support vector machine*, metode *k-nearest neighbor*, dan metode probabilistik. Teknik berbasis pada metode klasifikasi mencakup berbagai pendekatan yang digunakan untuk membangun model klasifikasi dari data.

   Pohon keputusan adalah teknik yang membagi data berdasarkan serangkaian keputusan yang diambil berdasarkan fitur-fitur data. Metode berbasis aturan juga menggunakan aturan-aturan logika untuk mengklasifikasikan data. Jaringan saraf tiruan adalah model matematika yang terinspirasi dari cara kerja otak manusia dan dapat digunakan untuk mempelajari pola-pola kompleks dalam

data. Metode SVM (*Support Vector Machine*) adalah teknik yang mencari *hyperplane* terbaik untuk memisahkan data ke dalam kelas-kelas yang berbeda. Metode tetangga terdekat, atau *k-nearest neighbor*, mengklasifikasikan data berdasarkan kesamaan dengan tetangga terdekatnya dalam ruang fitur.

2) Berbasis Tipe Data : Klasifikasi juga dapat dikategorikan berdasarkan jenis data yang digunakan, seperti data teks, multimedia, data yang tidak pasti, deret waktu, urutan diskrit, dan data jaringan. Kategori *data-type centered* pada metode klasifikasi mencakup berbagai jenis data yang digunakan dalam proses klasifikasi. Beberapa contoh dari jenis data yang termasuk dalam kategori ini meliputi data teks, multimedia, data yang tidak pasti, deret waktu, urutan diskrit, dan data jaringan.

   Data teks, misalnya, sering kali dihadapi dalam klasifikasi dan memiliki karakteristik unik seperti dimensi yang tinggi dan kejarangan. Representasi data teks sering menggunakan model "*bag-of-words*", di mana informasi urutan antar kata tidak digunakan. Tantangan utama dalam klasifikasi teks adalah dimensi yang sangat tinggi dan kejarangannya. Sebuah *leksikon teks* biasanya memiliki ratusan ribu kata, namun sebuah dokumen biasanya hanya mengandung sedikit kata. Oleh karena itu, sebagian besar nilai atribut adalah nol, dan frekuensinya relatif kecil. Banyak kata umum mungkin sangat bising dan tidak terlalu diskriminatif untuk proses klasifikasi .

   Selain itu, data multimedia, seperti gambar dan video, juga digunakan dalam klasifikasi. Data multimedia sering kali memiliki dimensi yang tinggi dan

memerlukan teknik klasifikasi khusus yang dapat menangani representasi data yang kompleks.

3) Variasi pada Analisis Klasifikasi: Terdapat berbagai variasi pada masalah klasifikasi, seperti pembelajaran kelas langka, pembelajaran transfer, pembelajaran semi-terawasi, atau pembelajaran aktif. Kategori *variations on classification analysis* pada metode klasifikasi mencakup berbagai variasi dari permasalahan klasifikasi standar yang ada, yang menangani skenario yang lebih menantang. Beberapa variasi tersebut meliputi pembelajaran kelas langka (*rare class learning*), pembelajaran transfer (*transfer learning*), pembelajaran semi-terawasi (*semi-supervised learning*), atau pembelajaran aktif (*active learning*). Selain itu, variasi lain dari klasifikasi, seperti analisis *ensemble*, juga dapat digunakan untuk meningkatkan efektivitas algoritma klasifikasi .

   Pembelajaran kelas langka berkaitan dengan kasus di mana kelas target yang jarang muncul dalam data. Pembelajaran transfer berkaitan dengan mentransfer pengetahuan dari satu domain ke domain lain. Pembelajaran semi-terawasi melibatkan penggunaan sebagian data yang berlabel dan sebagian tidak berlabel dalam proses pembelajaran. Sedangkan pembelajaran aktif melibatkan intervensi manusia untuk memperbaiki hasil klasifikasi.

   Variasi ini berkaitan erat dengan masalah model *ensemble*, yang bertujuan untuk mengurangi bias dan varians dalam model klasifikasi. Model *ensemble* menggunakan beberapa model untuk meningkatkan keandalan hasil klasifikasi dengan menggabungkan hasil dari beberapa model pelatihan secara berurutan atau independen.

Metode klasifikasi digunakan di berbagai bidang seperti ilmu komputer, sains dan teknik, pemerintahan, penegakan hukum, kedokteran, olahraga, dan masih banyak lagi yang lainnya. Metode ini melibatkan pembelajaran dari data yang ada dan kemudian mengklasifikasikan data baru, menjadikannya alat yang berharga untuk membuat prediksi dan keputusan berdasarkan data historis. Proses klasifikasi melibatkan pembuatan model atau fungsi yang menggambarkan dan membedakan kelas data, yang bertujuan untuk digunakan dalam pengambilan keputusan di masa depan.

Penerapan klasifikasi memiliki banyak aplikasi dalam kehidupan sehari-hari. Salah satu contoh penerapannya adalah dalam pemasaran target pelanggan. Dalam pemasaran, klasifikasi digunakan untuk memprediksi minat beli pelanggan berdasarkan data fitur yang dimiliki oleh pelanggan. Contohnya, data fitur seperti riwayat pembelian, preferensi produk, atau perilaku online dapat digunakan untuk memprediksi minat beli pelanggan. Selain itu, klasifikasi juga digunakan dalam diagnosis penyakit medis. Data medis seperti riwayat kesehatan, hasil tes, dan gejala pasien dapat digunakan untuk memprediksi kemungkinan pasien mengalami penyakit tertentu di masa depan. Hal ini dapat membantu dokter dalam membuat diagnosis yang lebih akurat. Penerapan klasifikasi juga dapat ditemukan dalam analisis data multimedia, seperti klasifikasi foto, video, atau audio. Misalnya, klasifikasi foto digunakan dalam aplikasi pengenalan wajah atau klasifikasi objek dalam gambar. Klasifikasi ini membantu dalam mengelompokkan dan mengidentifikasi data multimedia secara efisien.

Selain itu, klasifikasi juga digunakan dalam analisis data biologis, seperti dalam prediksi sifat-sifat urutan biologis atau dalam analisis jaringan biologis. Contoh lainnya adalah dalam kategorisasi dokumen dan filterisasi, yang digunakan dalam aplikasi seperti layanan berita untuk mengelompokkan dan menyaring dokumen secara otomatis. Dengan demikian, penerapan klasifikasi memiliki dampak yang signifikan dalam berbagai aspek kehidupan sehari-hari, membantu dalam pengambilan keputusan yang lebih baik dan efisien

### 2.2.3 *K-Means*

*K-means clustering* adalah algoritma *data mining* untuk mengelompokkan atau mengklasifikasikan "N" objek berdasarkan atribut atau fitur mereka menjadi "K" kelompok (yang disebut *cluster*) dengan meminimalkan jumlah kuadrat jarak antara data dan *centroid cluster* yang sesuai. "K" adalah bilangan bulat positif (K = 2, 3, 4, ...). Jika kita menerapkan *K-Means* pada sekumpulan "N" objek, maka hasilnya menjadi "K" kelompok yang saling lepas (menambahkan semua objek dalam kelompok "K" menghasilkan "N" objek) (Parsian, M. 2015).

Manfaat dari algoritma *k-means clustering* adalah untuk mengelompokkan data menjadi kelompok-kelompok yang saling lepas berdasarkan atribut atau fitur yang dimiliki oleh data tersebut. Hal ini dapat membantu dalam memahami pola dan hubungan dalam data, serta memudahkan dalam pengambilan keputusan. Contoh aplikasi dari algoritma *k-means clustering* adalah dalam bidang pemasaran untuk mengelompokkan pelanggan berdasarkan perilaku pembelian mereka, atau

dalam bidang asuransi untuk mengidentifikasi kelompok pemegang polis asuransi kendaraan dengan biaya klaim rata-rata yang tinggi.

*K-means clustering* adalah salah satu algoritma *data mining* yang paling umum digunakan dan efektif untuk mengelompokkan data. Namun, keefektifan algoritma ini tergantung pada banyak faktor, seperti jumlah data, jumlah *cluster*, dan kualitas data. Oleh karena itu, sebelum menerapkan algoritma *k-means*, penting untuk mempertimbangkan faktor-faktor tersebut dan melakukan evaluasi kinerja untuk memastikan keefektifan algoritma.

Algoritma juga berupaya menemukan grup dalam data, dan jumlah grup mewakili K variabel. Penentuan nilai "K" pada algoritma *k-means clustering* sangat penting karena dapat mempengaruhi hasil *clustering* yang dihasilkan. Namun, tidak ada rumus atau metode pasti untuk menentukan nilai "K" yang optimal. Beberapa cara yang dapat digunakan untuk menentukan nilai "K" adalah sebagai berikut (Parsian, M. 2015):

1) Metode *Elbow*: Metode ini melibatkan plot jumlah *cluster* (K) terhadap nilai fungsi objektif (misalnya, SSE atau *Sum of Squared Errors*). Kemudian, nilai K yang optimal adalah titik di mana penurunan SSE mulai melambat dan membentuk seperti siku (*elbow*).
2) Metode *Silhouette*: Metode ini melibatkan perhitungan koefisien *silhouette* untuk setiap titik data dalam setiap *cluster*. Koefisien *silhouette* mengukur seberapa baik titik data cocok dengan *cluster* tertentu. Kemudian, nilai "K" yang optimal adalah nilai yang menghasilkan koefisien *silhouette* tertinggi.

3) Metode *Gap Statistic*: Metode ini melibatkan perbandingan *sum of squared errors* aktual dengan *sum of squared errors* yang dihasilkan dari data acak. Kemudian, nilai "K" yang optimal adalah nilai yang menghasilkan perbedaan SSE aktual dan SSE acak yang paling besar.
4) Metode *Domain Knowledge*: Metode ini melibatkan pengetahuan domain ahli untuk menentukan nilai "K" yang optimal berdasarkan tujuan *clustering* dan karakteristik data.

Namun, dalam prakteknya, penentuan nilai "K" seringkali merupakan proses iteratif yang melibatkan beberapa metode di atas dan percobaan dengan nilai "K" yang berbeda untuk memastikan hasil yang optimal.

Algoritma pengelompokan *k-means* memproses pengelompokan dalam *multi-dimensi* dengan secara iteratif menetapkan titik-titik data ke pusat-pusat klaster terdekat dan memperbarui pusat-pusat tersebut berdasarkan rata-rata titik-titik di setiap *cluster*. Algoritma ini bekerja sebagai berikut:

1) Inisialisasi: Pilih "K" *centroid cluster* awal secara acak dari sampel "n" titik dalam ruang *d-dimensi.*
2) Penugasan: Hitung jarak dari setiap titik dalam kumpulan *input* ke masing-masing "K" pusat dan tetapkan setiap titik ke pusat *cluster* tertentu yang jaraknya paling dekat.
3) Perbarui pusat-pusat: Hitung ulang posisi "K" pusat berdasarkan rata-rata dari titik-titik di setiap *cluster*.
4) Ulangi: Langkah 2 dan 3 diulangi sampai pusat *cluster* tidak lagi berubah (atau hanya berubah sedikit).

Algoritma melakukan iterasi hingga tidak ada perubahan pada pusat-pusatnya, di mana pada saat itu jumlah *cluster* yang diinginkan tercapai.

Solusi *mapreduce* untuk pengelompokan *k-means* juga bekerja dalam ruang *multi-dimensi*, di mana setiap iterasi algoritma disusun sebagai satu pekerjaan *mapreduce*, secara berulang meningkatkan partisi data ke dalam "K" *cluster*. Fungsi *map()* dalam pekerjaan *mapreduce* mengklasifikasikan data dengan menetapkan setiap titik ke pusat terdekat, dengan mempertimbangkan sifat *multi-dimensi* data.

Oleh karena itu, algoritma *k-means* mampu memproses pengelompokan dalam ruang *multi-dimensi* dengan memperbarui pusat-pusat *cluster* secara iteratif dan menetapkan titik-titik data ke pusat-pusat terdekat, yang pada akhirnya menemukan *cluster* yang optimal dalam data *multi-dimensi*. Apabila nilai "K" sudah mencapai keinginan, maka hitung jarak data dari pusat dengan menggunakan jarak *euclidean. Euclidean distance* adalah salah satu fungsi jarak yang digunakan dalam algoritma *k-means clustering* untuk menghitung jarak antara dua titik dalam ruang *n-dimensi*. Dalam konteks *k-means clustering, euclidean distance* digunakan untuk menghitung jarak antara setiap titik data dengan *centroid* dari setiap *cluster*. Jarak *euclidean* antara dua titik dalam ruang dua dimensi dapat dihitung dengan menggunakan rumus (2.1) berikut:

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(y_i - x_i)x^2} \qquad (2.1)$$

Keterangan rumus :

d(x,y) = Jarak

xi = Data *training*

yi = Data *testing*

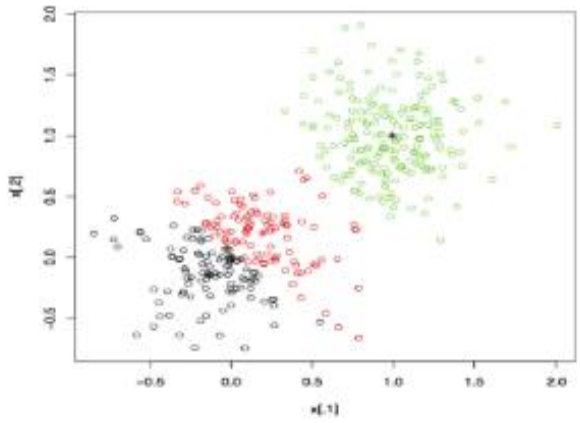
i = Variabel data

n = Dimensi data

Dimana $x_1$ dan $y_1$ adalah koordinat titik pertama, dan $x_2$ dan $y_2$ adalah koordinat titik kedua. Dalam ruang *n-dimensi*, rumus ini dapat diperluas dengan menambahkan koordinat tambahan. Dalam algoritma *k-means clustering*, titik data akan diatributkan ke *cluster* yang memiliki *centroid* terdekat berdasarkan jarak *euclidean*.

Pada awalnya, nilai *centroid* pada algoritma *k-means clustering* ditentukan secara acak dari data *input*. Salah satu cara umum untuk menentukan *nilai centroid* adalah dengan memilih "K" titik secara acak dari data input sebagai *centroid* awal. Setelah nilai *centroid* awal ditentukan, algoritma *k-means clustering* menghitung jarak antara setiap titik data dengan setiap *centroid*, dan kemudian menetapkan setiap titik data ke dalam *cluster* yang memiliki *centroid* terdekat. Setelah semua titik data ditetapkan ke dalam *cluster*, nilai *centroid* baru dihitung dengan mengambil rata-rata dari semua titik data dalam setiap *cluster*. Proses ini diulang sampai nilai *centroid* tidak berubah atau berubah sangat sedikit. Dalam beberapa kasus, algoritma *k-means clustering* dapat menghasilkan hasil yang berbeda-beda tergantung pada nilai *centroid* awal yang dipilih. Oleh karena itu, penting untuk

melakukan beberapa percobaan dengan nilai *centroid* yang berbeda untuk memastikan hasil yang optimal.




Gambar 2.2 Data *Clustering* dengan *K-Means*. Sumber : (Parsian, M. 2015)

Gambar 2.2 menampilkan akhir dari pengujian *cluster* yang berada pada titik – titik "K" *centroid.* Hasil akhir dari algoritma ini adalah kumpulan "K" *centroid* dan penugasan setiap objek ke salah satu dari "K" *cluster. Cluster - cluster* ini dapat digunakan untuk berbagai tujuan, seperti mengidentifikasi pola dalam data, mengelompokkan objek-objek yang serupa, atau mengurangi dimensi data (Parsian, M. 2015).

# BAB III

# DESAIN DAN IMPLEMENTASI

## 3.1 Rancangan Penelitian

Untuk pembuatan rancangan penelitian, maka terdapat banyak sekali peta cakupan yang membahas proses penelitian yang dijabarkan dengan alur grafik. Peta cakupan ini dilengkapi penjelasan pada grafik – grafik yang memuat poin – poin penting pada setiap kotak grafik sesuai penjelasan kotak tersebut. Ini menjadi solusi untuk memudahkan dalam penyelesaian penelitian yang memakai suatu model matematika atau kecerdasan buatan yang membutuhkan langkah – langkah yang rumit dalam pembacaan perhitungan data. Peta cakupan ini berfungsi sebagai panduan bagi peneliti dalam merencanakan dan melaksanakan penelitian dengan baik.

Dalam pembuatan peta cakupan ini, alur grafik digunakan untuk menggambarkan hubungan antara setiap tahapan dalam proses penelitian. Alur grafik ini membantu memvisualisasikan urutan langkah-langkah yang harus dilakukan oleh peneliti. Dengan adanya peta cakupan yang dijabarkan dengan alur grafik, peneliti dapat memiliki panduan yang jelas dalam merancang dan melaksanakan penelitian. Peta cakupan ini membantu menghindari kesalahan dan memastikan bahwa setiap tahapan dalam proses penelitian telah dilakukan dengan baik. Dengan demikian, hasil penelitian yang diperoleh lebih valid dan dapat dipercaya.

Dalam penelitian yang dibahas ini, maka pembuatan rancangan penelitian dimulai dengan pengumpulan dataset yang menggambarkan tabel sesuai laporan, pemetaan proses pemodelan *k-means* dengan *flowchart*, dan terakhir yaitu sistematika perhitungan prediksi dengan model *k-means* sesuai aturannya. Dalam penelitian yang dibahas ini, pembuatan rancangan penelitian dimulai dengan pengumpulan dataset yang menggambarkan tabel. Pengumpulan dataset merupakan langkah awal yang penting dalam proses penelitian, karena data yang akurat dan representatif akan menjadi dasar untuk analisis dan kesimpulan yang valid.

Dalam kesimpulannya, pembuatan rancangan penelitian dimulai dengan pengumpulan dataset yang menggambarkan tabel. Pengumpulan dataset ini merupakan langkah awal yang penting dalam proses penelitian dan harus dilakukan dengan hati-hati untuk memastikan data yang akurat dan representatif. Dengan memiliki dataset yang baik, peneliti dapat melanjutkan analisis data dan menyimpulkan hasil penelitian secara valid. Agar dapat disimak secara rinci, penulis menjabarkan rancangan tersebut pada konteks yang dibuat dalam bab selanjutnya.

### 3.1.1 Desain *Interface*

Salah satu aspek penting dalam pembuatan media klasifikasi algoritma *k-means* yaitu mengetahui dasar – dasar *interface*. Pada proses seluruh *machine learning* berdasarkan penelitian dengan proses yang sama yaitu *input*, proses, dan *output*.

Tahap pertama adalah *input*, dimana data atau informasi yang diperlukan untuk melakukan proses *machine learning* diinputkan ke dalam sistem. Data ini

dapat berupa angka, teks, gambar, atau bahkan suara. Pentingnya data yang berkualitas sangatlah besar karena hasil akhir dari proses *machine learning* sangat bergantung pada kualitas data yang digunakan.

Setelah data diinputkan, tahap selanjutnya adalah proses. Pada tahap ini, algoritma *machine learning* menganalisis dan memproses data tersebut untuk mencari pola atau hubungan tertentu di antara variabel - variabel yang ada. Proses ini melibatkan penggunaan berbagai teknik statistik dan matematika untuk menghasilkan model prediksi atau klasifikasi.

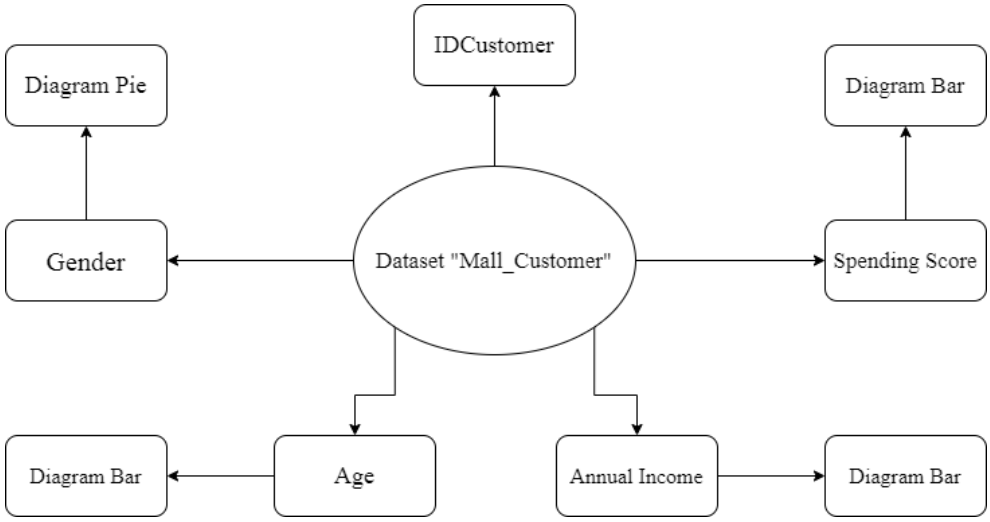
Terakhir adalah tahap *output*. Setelah melalui proses analisis dan pemodelan, hasil dari *machine learning* ditampilkan sebagai *output. Output* ini bisa berupa prediksi nilai numerik atau kategori tertentu berdasarkan data *input* yang diberikan sebelumnya.

Secara keseluruhan, proses *machine learning* didasarkan pada penelitian ilmiah dengan menggunakan metodologi yang sama yaitu *input-proses-output*. Penelitian dalam bidang machine learning bertujuan untuk mengembangkan algoritma dan model prediktif yang dapat digunakan untuk memecahkan masalah kompleks secara otomatis. Dalam era digital saat ini, kemampuan mesin untuk belajar dan beradaptasi dari data menjadi semakin penting dalam berbagai bidang, seperti pengenalan wajah, analisis risiko keuangan, dan pengenalan suara. Mesin dapat belajar dari data yang ada dan meningkatkan kemampuannya dalam mengenali wajah secara akurat.

Dalam era digital saat ini, pemanfaatan teknologi ini akan terus berkembang dan memberikan manfaat besar bagi masyarakat. Mengetahui pada proses – proses

ini, peneliti akan menjabarkan dan memberikan penggambaran desain apa yang dibutuhkan pada masing – masing faktor *interface*.

#### 3.1.1.1 *Input*



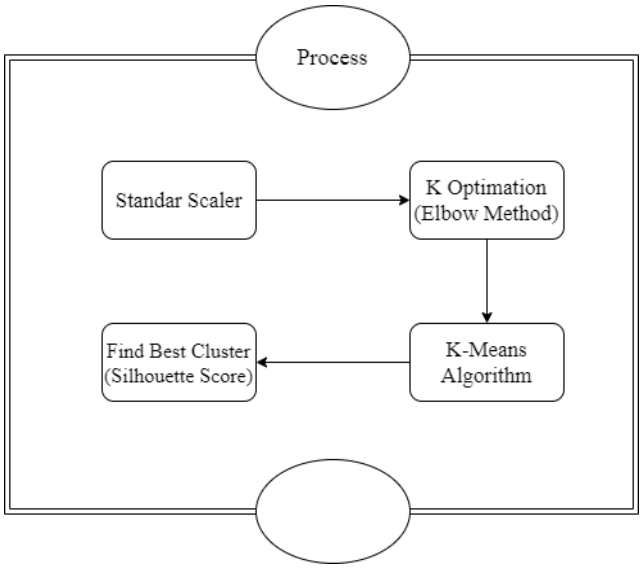

Gambar 3.1 Desain *Input* Data Pelanggan

Untuk desain dari input program pada gambar 3.1, berupa program yang dapat mengambil data dari *file* tersebut. Apakah itu *file teks*, *file csv*, atau mungkin *file excel*? Setelah itu, kita perlu memutuskan bagaimana cara membaca data dari *file* tersebut. Selain itu, penting juga untuk mempertimbangkan validasi data yang diambil dari *file*. Kita harus memastikan bahwa data yang dibaca sesuai dengan format yang diharapkan dan tidak ada kesalahan saat mengambil nilai-nilai tertentu. Terakhir, tetapi tidak kalah pentingnya adalah menangani kesalahan saat membaca atau mengambil data dari *file*. Kita harus menyediakan penanganan kesalahan yang tepat agar program tidak berhenti secara tiba-tiba jika terjadi masalah dengan *input*.

Dalam pemecahan dataset yaitu kolom “ID” berupa nomor urutan pada data. Kolom ini hanya mengurutkan data pelanggan yang datang pertama kali lalu dilanjutkan pelanggan seterusnya. Kolom “*Gender*” berisi tentang jenis kelamin dari pelanggan dengan penjelasan atribut berupa 2 objek saja yaitu laki – laki dan perempuan. Kolom “*Age*” berisikan hitungan umur dari pelanggan karena pelanggan itu manusia. Kolom “*Annual Income*” berisikan jumlah penghasilan pelanggan. Kolom terakhir yaitu “*Spending Score*” berisikan persentase pengeluaran belanja pelanggan yang berdasarkan penghasilan yang diperoleh.

Pada masing – masing kolom diberlakukan pengecekan statistik. Awal seluruh dataset memberikan info dan deskripsi tentang satuan penting terkait dengan perhitungan dari algoritma *k-means*. Untuk kolom “*Gender*” diberlakukan informasi statistik dengan *diagram pie* yang mempresentasikan persentase jumlah dataset antara laki – laki dan perempuan. Untuk kolom “*Age*”, ”*Annual Income*”, dan “*Spending Score*” diperoleh data statistik berupa *diagram bar* untuk mengetahui tingkatan masing – masing data berdasarkan hitungan angka yang tersedia.

### 3.1.1.2 *Proses*




Gambar 3.2 Desain *Interface* Algoritma *K-Means*

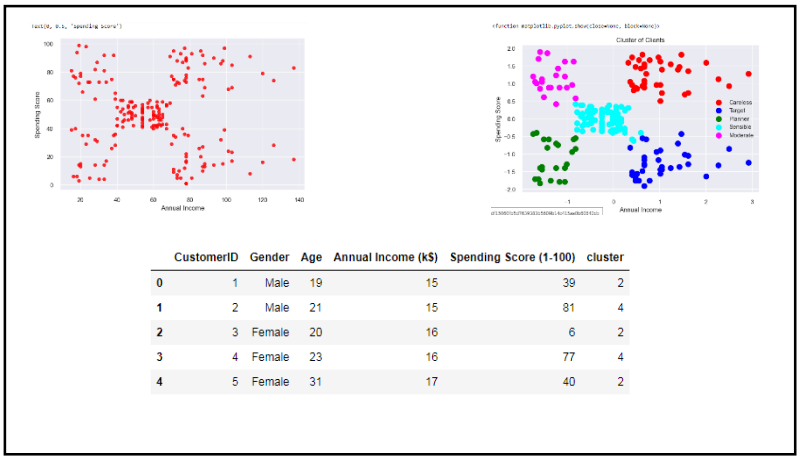
Untuk desain gambar 3.2 ini, program menerima input, antara lain "*N Cluster*" untuk jumlah *clustering* yang dibuat dengan format angka (*integer*) dan "K" untuk menentukan jumlah *cluster* saat proses pelatihan data dengan *k-means* dengan format angka jenis koordinat (*x,y*). *Clustering* adalah teknik dalam analisis data yang digunakan untuk mengelompokkan objek-objek serupa ke dalam kelompok-kelompok tertentu. Dalam konteks ini, program menggunakan algoritma *k-means* untuk melakukan *clustering*. Algoritma ini bekerja dengan membagi data menjadi beberapa kelompok berdasarkan jarak terdekat antara titik-titik data.

Input "*n cluster*" adalah angka *integer* yang menentukan jumlah kelompok atau *cluster* yang ingin dibuat oleh program. Angka ini harus sesuai dengan kebutuhan analisis data yang dilakukan. Semakin besar angka tersebut, semakin banyak kelompok terbentuk. *Input* "k" adalah angka *integer* yang menentukan nilai "K" saat proses pelatihan data dengan algoritma *k-means*. Nilai "K" merupakan

jumlah pusat atau *centroid* awal yang dipilih secara acak oleh algoritma untuk memulai proses *clustering*.

Petunjuk untuk proses data yaitu proses pengubahan data "*Annual Income*" dan "*Spending Score*" menjadi skala standar, penentuan nilai "k" optimal dengan metode *elbow*, perhitungan klasifikasi algoritma *k-means*, dan terakhir penentuan *cluster* terbaik menggunakan *silhouette score*.

### 3.1.1.3 *Output*




| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) | cluster |
|---|---|---|---|---|---|---|
| 0 | 1 | Male | 19 | 15 | 39 | 2 |
| 1 | 2 | Male | 21 | 15 | 81 | 4 |
| 2 | 3 | Female | 20 | 16 | 6 | 2 |
| 3 | 4 | Female | 23 | 16 | 77 | 4 |
| 4 | 5 | Female | 31 | 17 | 40 | 2 |

Gambar 3.3 Desain *Output* Grafik & Tabel Klasifikasi *K-Means*

Untuk desain gambar 3.3 yang terakhir yaitu *output* pada program *k-means* melibatkan beberapa tampilan. Dimulai dengan tabel klasifikasi yang menambahkan satu kolom yaitu *cluster* untuk menginisial klasifikasi data ke klaster yang ditentukan pada nomor *cluster*. Berikutnya yaitu grafik titik koordinasi pada 2 gambar tentang perbandingan antar data yang masih mentah berada di sebelah kiri dengan data yang sudah diklasifikasi dengan *k-means* yang berada di sebelah kanan. Salah satu tampilan *output* yang umum adalah visualisasi hasil *clustering*.

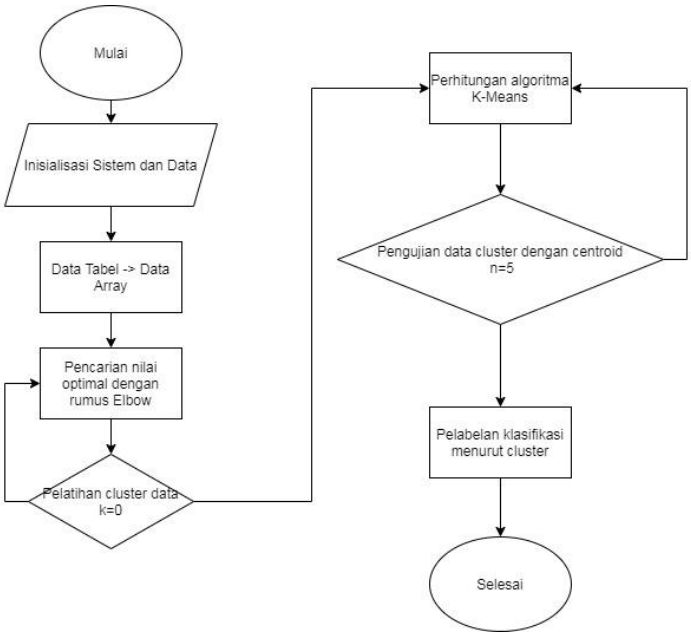
Dalam tampilan ini, data ditampilkan dalam bentuk grafik atau plot dengan setiap kelompok diberi warna atau simbol yang berbeda. Visualisasi ini membantu kita memahami bagaimana data terbagi menjadi kelompok-kelompok berdasarkan atribut-atribut tertentu.

Selain itu, *output* juga dapat mencakup tabel atau daftar yang menunjukkan pusat - pusat *cluster* dan anggota - anggota dari setiap kelompok. Pusat *cluster* adalah titik - titik representatif dari masing-masing kelompok, sedangkan anggota - anggota adalah data individu yang termasuk dalam suatu kelompok tertentu.

Tampilan *output* lainnya mungkin mencakup metrik evaluasi seperti *SSE (Sum of Squared Errors)* atau *silhouette score*. *SSE* mengukur sejauh mana titik-titik data dalam suatu kelompok mendekati pusat *cluster* mereka sendiri, sementara *silhouette score* menggambarkan seberapa baik setiap titik data cocok dengan kelompoknya dibandingkan dengan kelompok lainnya.

### 3.1.2 *Flowchart*

*Flowchart* menunjukkan seberapa lama dalam mengklasifikasikan sebuah pernyataan yang dikelola oleh penelitian saat menargetkan pelanggan *mall*. *Flowchart* menunjukkan kualitas *mind mapping* dalam pembuatan sistem klasifikasi yang dimanfaatkan oleh platform tersendiri. Mind mapping adalah teknik pemetaan pikiran yang menghubungkan ide - ide dan informasi dengan cara visual. Dalam pembuatan sistem klasifikasi, *mind mapping* dapat membantu dalam mengorganisir dan menyusun informasi secara logis. *Flowchart* merupakan salah satu bentuk visualisasi dari *mind mapping* ini.




Gambar 3.4 Alur *Flowchart K-Means*

Untuk gambar 3.4 dan penjelasan sebelumnya, kita dapat melihat langkah-langkah dalam proses pembuatan sistem klasifikasi dengan jelas dan terstruktur. *Flowchart* adalah representasi grafis dari alur kerja atau proses yang digambarkan dalam bentuk simbol-simbol dan panah-panah yang menghubungkannya.

Pertama-tama, langkah pertama dalam pembuatan sistem klasifikasi adalah mengidentifikasi tujuan dari sistem tersebut. Tujuan ini akan menjadi panduan untuk menentukan jenis klasifikasi yang digunakan dan data apa yang perlu dikumpulkan. Setelah itu, langkah berikutnya adalah mengumpulkan data yang diperlukan untuk melakukan klasifikasi. Data ini bisa berupa informasi tentang objek atau subjek yang akan diklasifikasikan. Kemudian, langkah selanjutnya adalah merancang algoritma atau metode untuk melakukan klasifikasi. Algoritma

ini harus dipilih dengan hati-hati agar dapat memberikan hasil yang akurat dan relevan.

Setelah algoritma dirancang, maka langkah berikutnya adalah mengimplementasikannya ke dalam kode program. Kode program ini menjalankan algoritma dan menghasilkan output berupa hasil klasifikasi. Terakhir, setelah kode program selesai diimplementasikan, maka dilakukan pengujian untuk memastikan bahwa sistem klasifikasi bekerja dengan baik dan memberikan hasil yang sesuai dengan harapan.

Dengan menggunakan *flowchart*, semua langkah-langkah tersebut dapat dijelaskan secara visual sehingga memudahkan pemahaman dan koordinasi antara tim pengembang. Selain itu, *flowchart* juga membantu dalam mendokumentasikan proses pembuatan sistem klasifikasi sehingga memudahkan dalam pengembangan dan pemeliharaan sistem di masa depan.

Secara keseluruhan, *flowchart* adalah alat yang sangat berguna dalam proses pembuatan sistem klasifikasi. Dengan menggunakan *flowchart*, kita dapat melihat langkah - langkah secara jelas dan terstruktur, sehingga memudahkan dalam mengembangkan sistem klasifikasi yang efektif dan efisien. Setiap langkah diwakili oleh simbol-simbol tertentu seperti kotak atau panah, sehingga memudahkan kita untuk memahami urutan dan hubungan antar langkah tersebut. Selain itu, *flowchart* juga memungkinkan kita untuk melihat bagaimana setiap langkah saling terkait dan berpengaruh satu sama lain. Kita dapat melihat apakah ada kemungkinan kesalahan atau kekurangan dalam sistem klasifikasi yang sedang dibuat, serta mencari solusi

atau perbaikan yang tepat. Kemungkinan dalam penelusuran alur *flowchart* juga disertai perhitungan manual seperti pada bab selanjutnya.

## 3.2 Sistematika *K-Means*

Metode yang digunakan adalah algoritma *k-means*, yaitu metode analisis yang mengumpulkan data berdasarkan pusat *cluster* terdekat (*centroid*) data. *K-means* merupakan upaya teknik pengelompokan non-superordinat membagi data yang ada menjadi satu atau lebih *cluster*. Teknik ini sering digunakan dalam analisis data dan *machine learning* untuk mengidentifikasi pola atau kelompok dalam dataset.

Metode *k-means* bekerja dengan cara mengelompokkan data ke dalam "k" *cluster*, di mana setiap *cluster* memiliki pusatnya sendiri. Pusat *cluster* ini disebut *centroid*, dan tujuan dari metode *k-means* adalah untuk meminimalkan jarak antara setiap titik data dengan *centroid* dari *cluster* yang sesuai.

Proses pengelompokan dimulai dengan menentukan jumlah *cluster* yang diinginkan. Kemudian, titik-titik awal dipilih secara acak sebagai *centroid* awal. Setelah itu, setiap titik data diberikan kepada *cluster* terdekat berdasarkan jarak *euclidean* antara titik tersebut dan *centroid cluster*. Selanjutnya, *centroid* baru dihitung berdasarkan rata-rata dari semua titik data dalam *cluster* tersebut. Proses ini diulang sampai tidak ada perubahan lagi pada posisi *centroid* atau jumlah iterasi tertentu telah mencapai batas.

Keuntungan dari metode *k-means* adalah kemampuannya untuk mengelompokkan data secara cepat dan efisien. Namun, metode ini juga memiliki

beberapa kelemahan seperti sensitivitas terhadap inisialisasi awal dan rentan terhadap *outlier.*

Secara keseluruhan, *k-means* merupakan salah satu teknik pengelompokan yang populer karena kemampuannya dalam membagi data menjadi kelompok-kelompok yang relevan. Dengan menggunakan metode ini, kita dapat mengidentifikasi pola atau kelompok dalam dataset yang dapat digunakan untuk berbagai tujuan seperti segmentasi pelanggan, analisis pasar, dan pengenalan pola. Untuk implementasinya, anda bisa menggunakan media pemrograman dari *library jupyter notebook.* Pengoperasian algoritma *k-means* adalah sebagai berikut:

1) Tentukan nilai "k" sebagai banyaknya klasifikasi yang ingin dibuat
2) Inisialisasi "k" sebagai pusat objek yang dihasilkan secara acak
3) Hitung jarak setiap titik data ke setiap *centroid* menggunakan persamaan *euclidean*
4) Kelompokkan setiap bagian data berdasarkan jarak antara data dan pusat objeknya
5) Carilah posisi baru pusat objek (k)
6) Kembali ke langkah 3 jika posisi *centroid* baru dan *centroid* lama tidak sama

```
import numpy as np
import random

def euclidean_distance(point1, point2):
    return np.sqrt(np.sum((point1 - point2) ** 2))

def find_closest_cluster(data_point, cluster_centroids):
    distances = [euclidean_distance(data_point, centroid) for
centroid in cluster_centroids]
    closest_cluster_index = np.argmin(distances)
    return closest_cluster_index

def k_means_clustering(data, k, num_iterations):
    # Step 1: Choose the number of clusters.
```

```
    num_clusters = k

    # Step 2: Initialize the cluster centroids randomly.
    num_features = len(data[0])
    centroids = [[random.random() for _ in range(num_features)]
for _ in range(num_clusters)]

    # Create a list to store the centroid of each cluster.
    centroid_centroid_distances = []

    for iteration in range(num_iterations):
        # Create a list to store the centroid of each cluster.
        centroid_centroid_distances = []

        # Step 3: Assign each data point to the closest cluster
centroid.
        clusters = [[] for _ in range(num_clusters)]
        for data_point in data:
            closest_cluster_index                                =
find_closest_cluster(data_point, centroids)
            clusters[closest_cluster_index].append(data_point)

        # Step 4: Update the cluster centroids by computing the
mean of all data points assigned to that cluster.
        for cluster_index, cluster in enumerate(clusters):
            centroids[cluster_index] = np.mean(cluster, axis=0)

        # Step 5: Calculate the distance between each pair of
cluster centroids.
        for i in range(num_clusters):
            for j in range(i+1, num_clusters):

centroid_centroid_distances.append(euclidean_distance(centroids[
i], centroids[j]))

        # Check if the cluster assignments do not change
significantly.
        if    len(centroid_centroid_distances)    ==    0    or
np.sum(np.array(centroid_centroid_distances) > 0.1) == 0:
            break

    return clusters, centroids
```

Kode Sumber 3.1 Rumus Algoritma Program *K-Means*

Untuk penerapan rumus *k-means* jenis klasifikasi bisa dilihat dalam kode program 3.1 diatas. Secara keseluruhan menggambarkan proses klasifikasi *k-means* yang digunakan pada kode *library*. Dataset pelanggan didapat dari situs web "*Kaggle.com"* yang berisi data pelanggan *mall* seperti ini:

Tabel 3.1 Data *Mall Customer*

| ID Customer | Jenis Kelamin | Umur | Penghasilan(k$/tahun) | Skor Pengeluaran (1-100) |
|---|---|---|---|---|
| 1 | Laki - Laki | 19 | 15 | 39 |
| 2 | Laki - Laki | 21 | 15 | 81 |
| 3 | Perempuan | 20 | 16 | 6 |
| 4 | Perempuan | 23 | 16 | 77 |
| 5 | Perempuan | 30 | 17 | 40 |
| 6 | Perempuan | 22 | 17 | 76 |
| 7 | Perempuan | 35 | 18 | 6 |
| 8 | Perempuan | 23 | 18 | 94 |
| 9 | Laki - Laki | 64 | 19 | 3 |
| 10 | Perempuan | 30 | 19 | 72 |

Dari tabel 3.1 diatas, data yang ditampilkan yaitu ID pelanggan, jenis kelamin, umur, penghasilan pelanggan tiap tahun dengan kurs *dollar*, dan skor pengeluaran belanja pelanggan dengan *range* 1 sampai 100. Sekarang diputuskan variabel dataset dari tabel yang ditampilkan menjadi seperti ini :

Tabel 3.2 Deskripsi Data *Mall Customer*

| | CustomerID | Umur | Penghasilan (k$/tahun) | Skor Pengeluaran (1-100) |
|---|---|---|---|---|
| **Jumlah Data** | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| **Rata - rata** | 100.500000 | 38.850000 | 60.560000 | 50.200000 |
| **Standar Deviasi** | 57.879185 | 13.969007 | 26.264721 | 25.823522 |
| **Data min** | 1.000000 | 18.000000 | 15.000000 | 1.000000 |
| **Data 25%** | 50.750000 | 28.750000 | 41.500000 | 34.750000 |
| **Data 50%** | 100.500000 | 36.000000 | 61.500000 | 50.000000 |
| **Data 75 %** | 150.250000 | 49.000000 | 78.000000 | 73.000000 |
| **Data max** | 200.000000 | 70.000000 | 137.000000 | 99.000000 |

Pada dataset yang digunakan dari tabel 3.2, jumlah data sebanyak 200 data, dan usia pergantian minimal 18 tahun sedangkan maksimal 70 tahun, dan omset atau pendapatan tahunan minimal 15.000 dengan maksimal $137.000, dan pada

variabel skor pengeluaran setiap pengunjung minimal 1 sedangkan maksimal 99. Dataset yang berisi rata – rata dan standar deviasi sangat penting untuk mendapatkan perhitungan selanjutnya. Rata-rata memberikan gambaran tentang nilai pusat dari dataset. Dengan mengetahui rata-ratanya, kita dapat memahami karakteristik keseluruhan data tersebut.

Standar deviasi mengukur seberapa jauh data tersebar dari rata-ratanya. Semakin besar standar deviasinya, semakin besar variasi datanya. Standar deviasi juga membantu dalam menentukan apakah suatu data tergolong *homogen* atau *heterogen*. Jika standar deviasinya kecil, artinya data cenderung *homogen* dan terkumpul di sekitar nilai rata-ratanya. Namun jika standar deviasinya besar, artinya data cenderung *heterogen* dan tersebar lebih luas.

Dalam analisis statistik lanjutan seperti uji hipotesis atau regresi linier, menggunakan dataset yang memiliki informasi tentang rata-rata dan standar deviasi sangat penting. Kedua ukuran ini membantu dalam memperoleh hasil yang akurat dan dapat dipercaya serta memudahkan interpretasi hasil analisis statistik tersebut.

Dalam kesimpulan, dataset yang berisi rata-rata dan standar deviasi sangat penting untuk mendapatkan perhitungan selanjutnya. Rata-rata memberikan gambaran tentang nilai pusat data, sedangkan standar deviasi mengukur sejauh mana data tersebar dari rata-ratanya. Kedua ukuran ini membantu dalam analisis statistik lanjutan dan memastikan hasil yang akurat serta dapat dipercaya.

Sistematika *k-means* dimulai dengan penentuan *cluster* dari pelanggan menjadi 5 *cluster* antara lain :

C1 : pelanggan “ceroboh”

C2 : pelanggan “yang ditargetkan”

C3 : pelanggan “perencana”

C4 : pelanggan “yang masuk akal”

C5 : pelanggan “normal”

Pemilihan nilai *cluster* digunakan pada penerapan identifikasi pada algoritma *k-means*. Tujuan dari algoritma ini adalah untuk mengelompokkan data menjadi beberapa kelompok berdasarkan kesamaan karakteristik.

Salah satu langkah kunci dalam algoritma *k-means* adalah menentukan jumlah *cluster* yang optimal. Jumlah *cluster* yang tidak tepat dapat menghasilkan hasil yang tidak akurat atau tidak bermakna. Oleh karena itu, pemilihan nilai *cluster* harus dilakukan dengan hati-hati. Ada beberapa metode yang dapat digunakan untuk memilih nilai *cluster* yang tepat. Salah satunya adalah menggunakan metode *elbow* atau siku tangan. Metode ini melibatkan plot jumlah *cluster* terhadap variansi total dalam kelompok. Jika plot membentuk lengkungan seperti siku tangan, maka titik di mana lengkungan tersebut terjadi merupakan jumlah *cluster* optimal.

Metode lainnya adalah menggunakan indeks validasi eksternal seperti indeks *dunn* atau indeks *davies-bouldin*. Indeks ini mengukur kualitas partisi dengan mempertimbangkan jarak antara pusat *cluster* dan jarak antara *cluster* itu sendiri. Selain itu, pemilihan nilai *cluster* juga dapat dilakukan secara empiris berdasarkan pengetahuan domain atau pengalaman praktis. Namun, pendekatan ini mungkin kurang objektif dan rentan terhadap bias subjektif.

Dalam kesimpulannya, pemilihan nilai *cluster* pada penerapan identifikasi pada algoritma *k-means* merupakan langkah penting untuk mendapatkan hasil yang

akurat dan bermakna. Metode *elbow*, indeks validasi eksternal, dan pendekatan empiris dapat digunakan untuk memilih nilai *cluster* yang tepat. Untuk selanjunya mengubah dataset yang bilangan diatas 0 menjadi bilangan dengan skala antara -2 dan 2 dengan *standard scaler* dengan rumus (3.1) sebagai berikut :

$$x = \frac{x_i - x_{mean}}{standard\ deviation} \qquad (3.1)$$

Untuk perhitungan masing – masing data sebagai berikut di bawah ini.

Ct_penghasilan1 = (15 – 60.56) / 26.264721 = -1.73899919

Ct_penghasilan2 = (15 – 60.56) / 26.264721 = -1.73899919

Ct_penghasilan3 = (16 – 60.56) / 26.264721 = -1.70082976

Ct_penghasilan4 = (16 – 60.56) / 26.264721 = -1.70082976

Ct_penghasilan5 = (17 – 60.56) / 26.264721 = -1.66266033

Ct_pengeluaran1 = (39 – 50.2) / 25.823522 = -0.43480148

Ct_pengeluaran2 = (81 – 50.2) / 25.823522 = 1.19570407

Ct_pengeluaran3 = (6 – 50.2) / 25.823522 = -1.71591298

Ct_pengeluaran4 = (77 – 50.2) / 25.823522 = 1.04041783

Ct_pengeluaran5 = (40 – 50.2) / 25.823522 = -0.39597992

Ct_penghasilan6 = (17 – 60.56) / 26.264721 = -1.658498485

Ct_penghasilan7 = (18 – 60.56) / 26.264721 = -1.620424599

Ct_penghasilan8 = (18 – 60.56) / 26.264721 = -1.620424599

Ct_penghasilan9 = (19 – 60.56) / 26.264721 = -1.582350713

Ct_penghasilan10 = (19 – 60.56) / 26.264721 = -1.582350713

Ct_pengeluaran6 = (76 – 50.2) / 25.823522 = 0.999089125

Ct_pengeluaran7 = (6 – 50.2) / 25.823522 = -1.711617803

Ct_pengeluaran8 = (94 – 50.2) / 25.823522 = 1.696128049

Ct_pengeluaran9 = (3 – 50.2) / 25.823522 = -1.827790957

Ct_pengeluaran10 = (72 – 50.2) / 25.823522 = 0.844191586

Tabel 3.3 Skala Standar Data

| Id | Skala standar "*Annual Income*" | Skala standar "*Spending Score*" |
|---|---|---|
| 1 | -1.73899919 | -0.43480148 |
| 2 | -1.73899919 | 1.19570407 |
| 3 | -1.70082976 | -1.71591298 |
| 4 | -1.70082976 | 1.04041783 |
| 5 | -1.66266033 | -0.39597992 |
| 6 | -1.658498485 | 0.999089125 |
| 7 | -1.620424599 | -1.711617803 |
| 8 | -1.620424599 | 1.696128049 |
| 9 | -1.582350713 | -1.827790957 |
| 10 | -1.582350713 | 0.844191586 |

Tabel 3.3 menunjukkan hasil – hasil perhitungan manual pada perhitungan *standard scaler* sehingga menghasilkan nilai *standard scaler* 10 data awal.

```
# Randomly initialize the centroids
centroids = [[random.randint(0, 10), random.randint(0, 10)] for
_ in range(5)]


# Function to calculate the distance between two points
def distance(point1, point2):
    return np.sqrt(np.sum((point1 - point2) ** 2))


# Apply the K-means algorithm
for iteration in range(100): # limit the number of iterations to
avoid getting stuck in a local minimum
```

```
    clusters = [[] for _ in range(5)]


    # Assign each data point to the closest centroid
    for point in data:
        closest_centroid_index = np.argmin([distance(point,
centroid) for centroid in centroids])
        clusters[closest_centroid_index].append(point)


    # Recalculate the centroids
    centroids = [np.mean(cluster, axis=0) for cluster in
clusters]
```

Kode Sumber 3.2 Baris Program Perhitungan Data *Centroid*

Sekarang pada kode program 3.2 dengan nilai K = 5 perhitungan data *centroid* dilakukan dengan media pemrograman. Dalam perhitungan data *centroid*, seringkali memiliki banyak perubahan yang terjadi sehingga hasil tersebut benar – benar acak.

Tabel 3.4 Data *Centroid* 5 *Cluster*

| ***Cluster*** | ***Centroid "Annual Income"*** | ***Centroid "Spending Score"*** |
|---|---|---|
| C1 | -0.20091257 | -0.02645617 |
| C2 | 0.99158305 | 1.23950275 |
| C3 | 1.05500302 | -1.28443907 |
| C4 | -1.32954532 | 1.13217788 |
| C5 | -1.30751869 | -1.13696536 |

Maka tabel 3.4 merupakan nilai terbaik yang melalui perhitungan dan penyempurnaan akurasi dari proses pencarian data *centroid*. Setelah pertimbangan

penentuan data *centroid*, dilakukan perhitungn jarak data ke *cluster* dengan rumus *euclidian distance* sebagai berikut :

Data Id 1 :

(d1, c1) = $\sqrt{(-1.73899919 - (-0.20091257))^2 + (-0.43480148 - (-0.02645617))^2}$ = 1.591369329

(d1, c2) = $\sqrt{(-1.73899919 - 0.99158305)^2 + (-0.43480148 - 1.23950275)^2}$ = 3.203025761

(d1, c3) = $\sqrt{(-1.73899919 - 1.05500302)^2 + (-0.43480148 - (-1.28443907))^2}$ = 2.920330869

(d1, c4) = $\sqrt{(-1.73899919 - (-1.32954532))^2 + (-0.43480148 - 1.13217788)^2}$ = 1.619591549

(d1, c5) = $\sqrt{(-1.73899919 - (-1.30751869))^2 + (-0.43480148 - (-1.13696536))^2}$ = 0.824141697

Data Id 2 :

(d2, c1) = $\sqrt{(-1.73899919 - (-0.20091257))^2 + (1.19570407 - (-0.02645617))^2}$ = 1.964532032

(d2, c2) = $\sqrt{(-1.73899919 - 0.99158305)^2 + (1.19570407 - 1.23950275)^2}$ = 2.730933484

(d2, c3) = $\sqrt{(-1.73899919 - 1.05500302)^2 + (1.19570407 - (-1.28443907))^2}$ = 3.735981577

(d2, c4) = $\sqrt{(-1.73899919 - (-1.32954532))^2 + (1.19570407 - 1.13217788)^2}$ = 0.414352565

(d2, c5) = $\sqrt{(-1.73899919 - (-1.30751869))^2 + (1.19570407 - (-1.13696536))^2}$ = 2.372239889

Data Id 3 :

(d3, c1) = $\sqrt{(-1.70082976 - (-0.20091257))^2 + (-1.71591298 - (-0.02645617))^2}$ = 2.259206916

(d3, c2) = $\sqrt{(-1.70082976 - 0.99158305)^2 + (-1.71591298 - 1.23950275)^2}$ = 3.997945582

(d3, c3) = $\sqrt{(-1.70082976 - 1.05500302)^2 + (-1.71591298 - (-1.28443907))^2}$ = 2.78940568

(d3, c4) = $\sqrt{(-1.70082976 - (-1.32954532))^2 + (-1.71591298 - 1.13217788)^2}$ = 2.872189702

(d3, c5) = $\sqrt{(-1.70082976 - (-1.30751869))^2 + (-1.71591298 - (-1.13696536))^2}$ = 0.699909955

Data Id 4 :

(d4, c1) = $\sqrt{(-1.70082976 - (-0.20091257))^2 + (1.04041783 - (-0.02645617))^2}$ = 1.840644373

(d4, c2) = $\sqrt{(-1.70082976 - 0.99158305)^2 + (1.04041783 - 1.23950275)^2}$ = 2.699763239

(d4, c3) = $\sqrt{(-1.70082976 - 1.05500302)^2 + (1.04041783 - (-1.28443907))^2}$ = 3.605492188

(d4, c4) = $\sqrt{(-1.70082976 - (-1.32954532))^2 + (1.04041783 - 1.13217788)^2}$ = 0.382455281

(d4, c5) = $\sqrt{(-1.70082976 - (-1.30751869))^2 + (1.04041783 - (-1.13696536))^2}$ = 2.212620879

Data Id 5 :

(d5, c1) = $\sqrt{(-1.66266033 - (-0.20091257))^2 + (-0.39597992 - (-0.02645617))^2}$ = 1.507731513

(d5, c2) = $\sqrt{(-1.66266033 - 0.99158305)^2 + (-0.39597992 - 1.23950275)^2}$ = 3.11766122

(d5, c3) = $\sqrt{(-1.66266033 - 1.05500302)^2 + (-0.39597992 - (-1.28443907))^2}$ = 2.85920509

(d5, c4) = $\sqrt{(-1.66266033 - (-1.32954532))^2 + (-0.39597992 - 1.13217788)^2}$ = 1.564043437

(d5, c5) = $\sqrt{(-1.66266033 - (-1.30751869))^2 + (-0.39597992 - (-1.13696536))^2}$ = 0.82169642

Data Id 6 :

(d6, c1) = $\sqrt{(-1.658498485 - (-0.20091257))^2 + (0.999089125 - (-0.02645617))^2}$ = 1.782217678

(d6, c2) = $\sqrt{(-1.658498485 - 0.99158305)^2 + (0.999089125 - 1.23950275)^2}$ = 2.660964271

(d6, c3) = $\sqrt{(-1.658498485 - 1.05500302)^2 + (0.999089125 - (-1.28443907))^2}$ = 3.546490016

(d6, c4) = $\sqrt{(-1.658498485 - (-1.32954532))^2 + (0.999089125 - 1.13217788)^2}$ = 0.354856029

(d6, c5) = $\sqrt{(-1.658498485 - (-1.30751869))^2 + (0.999089125 - (-1.13696536))^2} = 2.164697572$

Data Id 7 :

(d7, c1) = $\sqrt{(-1.620424599 - (-0.20091257))^2 + (-1.711617803 - (-0.02645617))^2} = 2.203357468$

(d7, c2) = $\sqrt{(-1.620424599 - 0.99158305)^2 + (-1.711617803 - 1.23950275)^2} = 3.941027338$

(d7, c3) = $\sqrt{(-1.620424599 - 1.05500302)^2 + (-1.711617803 - (-1.28443907))^2} = 2.709316263$

(d7, c4) = $\sqrt{(-1.620424599 - (-1.32954532))^2 + (-1.711617803 - 1.13217788)^2} = 2.858633352$

(d7, c5) = $\sqrt{(-1.620424599 - (-1.30751869))^2 + (-1.711617803 - (-1.13696536))^2} = 0.654320669$

Data Id 8 :

(d8, c1) = $\sqrt{(-1.620424599 - (-0.20091257))^2 + (1.696128049 - (-0.02645617))^2} = 2.232109046$

(d8, c2) = $\sqrt{(-1.620424599 - 0.99158305)^2 + (1.696128049 - 1.23950275)^2} = 2.651620377$

(d8, c3) = $\sqrt{(-1.620424599 - 1.05500302)^2 + (1.696128049 - (-1.28443907))^2} = 4.005208271$

(d8, c4) = $\sqrt{(-1.620424599 - (-1.32954532))^2 + (1.696128049 - 1.13217788)^2} = 0.634547514$

(d8, c5) = √(-1.620424599 – (-1.30751869))$^2$ + (1.696128049 – (-1.13696536))$^2$ = 2.850320749

Data Id 9 :

(d9, c1) = √(-1.582350713 – (-0.20091257))$^2$ + (-1.827790957 – (-0.02645617))$^2$ = 2.270061311

(d9, c2) = √(-1.582350713 – 0.99158305)$^2$ + (-1.827790957 – 1.23950275)$^2$ = 4.004176033

(d9, c3) = √(-1.582350713 – 1.05500302)$^2$ + (-1.827790957 – (-1.28443907))$^2$ = 2.692743208

(d9, c4) = √(-1.582350713 – (-1.32954532))$^2$ + (-1.827790957 – 1.13217788)$^2$ = 2.970745038

(d9, c5) = √(-1.582350713 – (-1.30751869))$^2$ + (-1.827790957 – (-1.13696536))$^2$ = 0.743486817

Data Id 10 :

(d10, c1) = √(-1.582350713 – (-0.20091257))$^2$ + (0.844191586 – (-0.02645617))$^2$ = 1.63291116

(d10, c2) = √(-1.582350713 – 0.99158305)$^2$ + (0.844191586 – 1.23950275)$^2$ = 2.604113272

(d10, c3) = √(-1.582350713 – 1.05500302)$^2$ + (0.844191586 – (-1.28443907))$^2$ = 3.389203916

(d10, c4) = √(-1.582350713 – (-1.32954532))$^2$ + (0.844191586 – 1.13217788)$^2$ = 0.383205783

(d10, c5) = √(-1.582350713 – (-1.30751869))2 + (0.844191586 – (-1.13696536))2 = 2.000128867

Akhirnya dengan jarak data pada nilai *centroid* didapatkan. Pada masing – masing jarak tersebut diambil nilai terkecil karena nilai terkecil mengindiasikan jarak pada data saling berdekatan.

Tabel 3.5 Data *Mall Customer* Beserta Kelompok *Cluster*

| ID | Jenis Kelamin | Umur | Penghasilan (k$/tahun) | Skor Pengeluaran (1-100) | Kelompok Cluster |
|---|---|---|---|---|---|
| 1 | Laki - Laki | 19 | 15 | 39 | C5 |
| 2 | Laki - Laki | 21 | 15 | 81 | C4 |
| 3 | Perempuan | 20 | 16 | 6 | C5 |
| 4 | Perempuan | 23 | 16 | 77 | C4 |
| 5 | Perempuan | 31 | 17 | 40 | C5 |
| 6 | Perempuan | 22 | 17 | 76 | C4 |
| 7 | Perempuan | 35 | 18 | 6 | C5 |
| 8 | Perempuan | 23 | 18 | 94 | C4 |
| 9 | Laki - Laki | 64 | 19 | 3 | C5 |
| 10 | Perempuan | 30 | 19 | 72 | C4 |

Dari tabel 3.5, proses penyesuaian jarak *euclidian* memberikan data sebuah label sebagai tanda *cluster* yang ke berapa. Akhirnya terbentuk 5 *cluster* dari perhitungan manual algoritma *k-means* hingga selanjutnya ke tahap penelitian dengan media pemrograman.

# BAB IV

# UJI COBA DAN HASIL

Dalam Bab 4 ini, penelitian memasuki tahap implementasi suatu sistem ke media perangkat yang bisa memproses analisa perhitungan manual dari bab sebelumnya. Implementasi ini merupakan langkah penting dalam proses penelitian karena melibatkan penggunaan teknologi untuk mempermudah dan meningkatkan efisiensi dalam melakukan analisis. Pada bab sebelumnya, telah dilakukan analisis perhitungan secara manual untuk mengumpulkan data dan informasi yang diperlukan. Namun, metode ini memiliki beberapa keterbatasan seperti waktu yang dibutuhkan yang cukup lama dan potensi kesalahan manusia. Oleh karena itu, implementasi sistem ke media perangkat menjadi solusi yang tepat untuk mengatasi masalah tersebut.

Dengan menggunakan media perangkat seperti komputer atau *smartphone*, proses analisis dapat dilakukan dengan lebih cepat dan akurat. Sistem yang diimplementasikan dapat membantu dalam mengolah data secara otomatis sehingga mengurangi risiko kesalahan manusia. Selain itu, sistem juga dapat menyimpan data dengan baik sehingga memudahkan akses dan pengolahan data di masa depan. Selama tahap implementasi ini, saat ini dilakukan uji coba terhadap sistem yang telah dirancang. Uji coba ini bertujuan untuk mengevaluasi kinerja sistem serta menemukan potensi masalah atau kekurangan yang perlu diperbaiki. Hasil dari uji coba tersebut digunakan sebagai dasar untuk melakukan penyempurnaan pada sistem agar sesuai dengan kebutuhan penelitian.

Dalam kesimpulan, implementasi suatu sistem ke media perangkat merupakan langkah penting dalam penelitian ini. Dengan menggunakan teknologi, proses analisis dapat dilakukan dengan lebih efisien dan akurat. Tahap implementasi ini juga memberikan kesempatan untuk melakukan uji coba dan penyempurnaan sistem agar sesuai dengan kebutuhan penelitian. Tahap ini akan mengukur seberapa layaknya proses perhitungan lewat program tersebut, apakah masih ada kesalahan ataupun kekurangan dari sistem yang dijalankan oleh program tersebut. Lalu sistem yang dijalankan juga menjadi lebih singkat saat memasuki proses klasifikasi sistem. Ini menjadikan bab 4 sebagai tahap terakhir untuk menyelesaikan penelitian ini

## 4.1 *Interface System*

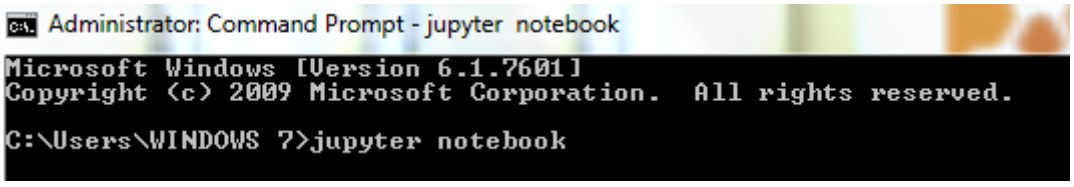
Kali ini, algoritma *k-means* dijalankan di *platform* "*Jupyter Notebook*". Media ini sangat mudah diaplikasikan dalam beberapa algoritma lain karena tidak lain yaitu bahasa pemrograman *python*. *Jupyter notebook* adalah sebuah lingkungan pengembangan yang sangat populer di kalangan ilmuwan data dan peneliti. Dalam beberapa tahun terakhir, *Jupyter notebook* telah menjadi alat yang sangat mudah diaplikasikan dalam berbagai algoritma lain, terutama karena menggunakan bahasa pemrograman *python*.

Bahasa pemrograman *python* telah menjadi bahasa yang paling banyak digunakan dalam dunia ilmu data dan kecerdasan buatan. Kelebihan utama *python* adalah sintaksisnya yang sederhana dan mudah dipahami, sehingga memungkinkan

para pengguna untuk dengan cepat mengimplementasikan algoritma - algoritma kompleks.

Dalam *jupyter notebook*, kita dapat menulis kode *python* secara interaktif dan menjalankannya secara langsung. Hal ini memungkinkan para pengguna untuk melihat hasil dari setiap langkah dalam proses komputasi, serta melakukan eksperimen dengan parameter-parameter yang berbeda. Selain itu, *jupyter notebook* juga mendukung visualisasi data yang kaya dan interaktif. Kita dapat membuat grafik atau plot langsung di dalam notebook, sehingga memudahkan kita untuk menganalisis dan memvisualisasikan hasil dari algoritma - algoritma yang telah diterapkan.

Dalam beberapa kasus, *jupyter notebook* juga digunakan sebagai dokumentasi interaktif untuk proyek - proyek ilmu data atau kecerdasan buatan. Para pengguna dapat menuliskan catatan-catatan atau penjelasan mengenai kode-kode yang telah ditulis, serta menyertakan hasil-hasil analisis atau eksperimen sebagai bagian dari dokumentasi tersebut.



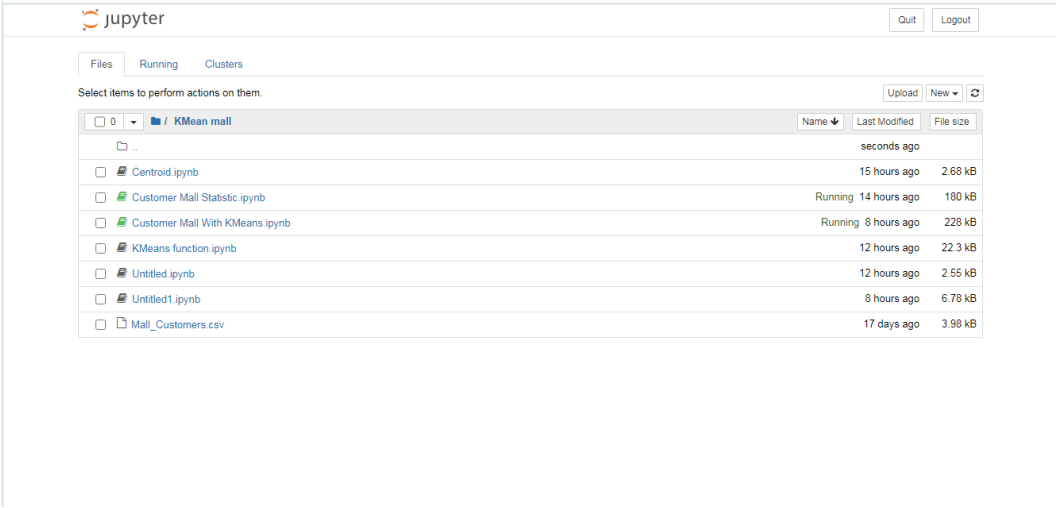

Gambar 4.1 Pemanggilan *Localhost Jupyetr* denagn *Command Prompt*

Gambar 4.1 merupakan panel *command prompt* untuk membuka platform "*Jupyter Notebook*". Dalam menjalankan platform "*Jupyter Notebook*", terlebih dahulu masuk ke *command prompt* pada "*Windows"* seperti ini. Lalu *command*

*prompt* mengarahkan media ke aplikasi *website* dengan alamat *host* yaitu *http://localhost:8888/tree/*. Disitu pembuatan program berlangsung dengan pembacaan *file extensi IPYNB* (*IPython Notebook*).


Gambar 4.2 Tampilan *User Files Jupyter Notebook*

Gambar 4.2 merupakan tampilan platform "*Jupyter Notebook"* dengan segala akses program berlangsung. Akses tersebut memungkinkan *file* jenis *python* dapat diakses. Hingga saat ini, penerapan pemrograman *python* masih ditulis kebanyakan dari media ini. Halaman media mirip seperti halaman database yang diambil dalam penyimpanan komputer dengan *user admin* sendiri. Ada beberapa *folder* yang dapat dibuka tetapi hanya bisa mengakses file jenis *python*. Kemudian pada halaman ini, file program *python* dibuka untuk menghubungkan program serta dataset yang tersedia.




Gambar 4.3 Tampilan File Program *K-Means*

Pada gambar 4.3 diatas, sudah tertera beberapa *file* untuk proses algoritma *k-means*. Awal program yaitu "*Customer Mall Statistic*" untuk menjelaskan dataset pelanggan *mall* yang dilengkapi dengan beberapa gambar grafik dan diagram. Grafik pada data menjelaskan peta dengan atribut bilangan. Pada diagram menjelaskan persentase data dengan atribut objek agar dapat diidentifikasi kelompok data yang berdasarkan objek tersebut.

Program selanjutnya yaitu "*Customer Mall With KMeans*" menjelaskan tentang proses pengolahan dataset dengan beberapa fungsi dan *library* yang dibutuhkan. Setiap proses mewakili langkah – langkah penunjang perubahan dataset. Untuk data disiapkan *file csv* yang berisikan tabel dari pelanggan yang sudah ditata sesuai atribut yang ditampilkan. *File* program ini sudah di tambahkan dengan *library* yang mencakup kebutuhan program pada *python*. Adapaun *library* yang disebutkan yaitu :

1) *Panda*
2) *Numpy*
3) *Matplotlib*
4) *Seaborn*
5) *Scikit Learn*
6) *YellowBrick*

## 4.2 Pengecekan Statistik Dataset *Input*

Sebelum proses pengujian sistem, dataset tidak langsung memasukkan tabel data hasil pembacaan *csv* ke dalam program Klasifikasi dulu. Dataset diidentifikasi setiap data, baris, maupun kolom dengan pencatatan informasi data. Mulai dari informasi atribut, informasi tipe data, informasi data minimal, informasi data maksimal, informasi data rata – rata, informasi standar deviasi, informasi jumlah data, informasi *diagram pie*, informasi *grafik bar*, dan lain – lain.

Pada awal pengenalan dataset dimulai dari pembacaan *library* khusus untuk pengenalan tabel data. *Library* ini terlebih dahulu diinisialisasikan sebagai berikut.

```
import warnings
warnings.filterwarnings('always')
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="darkgrid")

%matplotlib inline
```

Kode Sumber 4.1 Baris Program Inisialisasi *Library* Berhubungan denagn Tabel

Kode baris program 4.1 menjelaskan pengambilan *Library Pandas, Numpy, Matplotlib,* dan *Seaborn* sebagai fungsi tambahan dalam menampilkan dan memodifikasi tabel data pelanggan *mall* tersebut. Kode baris program *library* adalah kumpulan fungsi dan prosedur yang telah dikompilasi dan dapat digunakan kembali dalam berbagai aplikasi. Dalam konteks ini, kode baris program *library* dapat digunakan sebagai fungsi tambahan dalam menampilkan dan memodifikasi tabel data pelanggan di sebuah *mall*.

Salah satu manfaat utama dari menggunakan kode baris program *library* adalah efisiensi waktu dan upaya yang diperlukan untuk mengembangkan perangkat lunak. Dengan menggunakan kode baris program *library*, pengembang tidak perlu menulis ulang fungsi - fungsi dasar seperti menambahkan atau menghapus data pada tabel pelanggan. Mereka hanya perlu memanggil fungsi yang sudah ada di dalam kode baris program *library* tersebut.

Selain itu, penggunaan kode baris program *library* juga meningkatkan keamanan aplikasi. Dengan menggunakan fungsi-fungsi yang sudah teruji dan terpercaya dalam kode baris program *library*, risiko kesalahan pemrograman dapat diminimalisir. Hal ini sangat penting ketika bekerja dengan data pelanggan *mall* yang mungkin sensitif dan harus dijaga kerahasiaannya.

Dalam hal menampilkan tabel data pelanggan *mall*, kode baris program *library* dapat menyediakan fitur-fitur seperti *sorting* (pengurutan) berdasarkan kolom tertentu atau *filter* (penyaringan) berdasarkan kriteria tertentu. Fitur-fitur ini memudahkan pengguna untuk menemukan data yang mereka butuhkan dengan cepat dan efisien.

Selain itu, kode baris program *library* juga dapat digunakan untuk memodifikasi tabel data pelanggan. Misalnya, pengguna dapat menggunakan fungsi yang sudah ada di dalam kode baris program *library* untuk mengubah atau menghapus data pelanggan tertentu. Hal ini dapat memudahkan pengguna dalam melakukan tugas-tugas administratif terkait dengan manajemen pelanggan *mall.*

Dalam kesimpulan, penggunaan kode baris program *library* sebagai fungsi tambahan dalam menampilkan dan memodifikasi tabel data pelanggan *mall* sangatlah bermanfaat. Hal ini tidak hanya meningkatkan efisiensi dan keamanan aplikasi, tetapi juga memberikan fitur-fitur tambahan yang berguna bagi pengguna. Dengan demikian, implementasi kode baris program *library* menjadi suatu langkah yang bijak dalam mengembangkan perangkat lunak untuk manajemen pelanggan *mall.*

```
data = pd.read_csv("Mall_Customers.csv")
data.iloc[:]
```

Kode Sumber 4.2 Baris Program Pengambilan Data Pelanggan "*Mall_Customer.csv*"

Tabel 4.1 Tampilan Dataset "*Mall_Customer.csv*"

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 |
| 2 | Male | 21 | 15 | 81 |
| 3 | Female | 20 | 16 | 6 |
| 4 | Female | 23 | 16 | 77 |
| 5 | Female | 31 | 17 | 40 |
| 6 | Female | 22 | 17 | 76 |
| 7 | Female | 35 | 18 | 6 |
| 8 | Female | 23 | 18 | 94 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 9 | Male | 64 | 19 | 3 |
| 10 | Female | 30 | 19 | 72 |

Pada kode program 4.2 dan diatas menjelaskan bahwa untuk menampilkan data jenis *file csv* itu menggunakan perintah *read_csv* dengan bantuan *library pandas*. Maka hasil tersebut bisa menggunakan informasi tabel 4.1 yaitu dataset "*Mall_Customer.csv*". Penggunaan tabel sebagai alat visualisasi sangat penting. Tabel adalah cara yang efektif untuk menyajikan data secara terstruktur dan mudah dipahami. Dalam banyak situasi, hasil dari suatu penelitian atau analisis dapat digunakan dengan lebih baik menggunakan informasi tabel.

Pertama-tama, tabel memungkinkan kita untuk melihat pola dan tren dengan jelas. Dengan mengatur data dalam kolom dan baris, kita dapat dengan mudah mengidentifikasi perubahan seiring waktu atau perbandingan antara berbagai variabel. Misalnya, jika kita ingin melihat pertumbuhan populasi di beberapa negara selama 10 tahun terakhir, tabel membantu kita melihat apakah ada peningkatan atau penurunan yang signifikan.

Selain itu, tabel juga memungkinkan kita untuk melakukan perbandingan langsung antara data yang berbeda. Misalnya, jika kita ingin membandingkan tingkat pendidikan di beberapa negara, tabel memberi kami gambaran tentang persentase penduduk yang memiliki gelar sarjana atau lebih tinggi di setiap negara tersebut.

Selain itu, informasi dalam tabel juga dapat digunakan untuk membuat keputusan yang lebih baik. Misalnya, jika kita ingin membeli mobil baru dan ingin mengetahui harga rata-rata mobil dari beberapa merek terkenal, maka melihat informasi harga dalam bentuk tabel membantu kami membuat keputusan berdasarkan angka-angka tersebut.

Dalam kesimpulannya, hasil dari suatu penelitian atau analisis dapat digunakan dengan lebih baik menggunakan informasi tabel. Tabel memungkinkan kita untuk melihat pola dan tren, melakukan perbandingan langsung antara data yang berbeda, dan membuat keputusan yang lebih baik. Oleh karena itu, penting bagi kita untuk mengembangkan keterampilan dalam membaca dan menggunakan informasi dari tabel dengan efektif. Disini tabel ditampilkan dengan perintah *iloc* untuk membuat data muncul dengan aturan baris berapa dan kolom berapa.

```
data.info()
data.describe()
```

Kode Sumber 4.3 Baris Program Informasi Jumlah dan Satuan Penting Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Gambar 4.4 Informasi Atribut dan Tipe Data

Pada kode program 4.3 dan gambar 4.4 ini, data diperlihatkan informasi – informasi seputar data tersebut dari atribut dan tipe data pada baris pertama. Bisa dilihat bahwa selain "*Gender"*, jenis kolom memiliki tipe data jenis *Interger* berarti angka atau bilangan. Dikarenakan pada data tersebut konfirmasinya yaitu "*Age*" dengan keterangan berapa umur *customer* tersebut, "*Annual Income*" dengan keterangan berapa penghasilan *customer* tersebut, dan "*Spending Score*" dengan keterangan berapa persentase pengeluaran dari customer tersebut. Lalu untuk "*Gender*" dengan keterangan customer tersebut berjenis kelamin apa.

Tabel 4.2 Informasi Jumlah dan Satuan Penting Data

| | **CustomerID** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
|---|---|---|---|---|
| count | 200 | 200 | 200 | 200 |
| mean | 100.5 | 38.85 | 60.56 | 50.2 |
| std | 57.879185 | 13.969007 | 26.264721 | 25.823522 |
| min | 1 | 18 | 15 | 1 |
| 25% | 50.75 | 28.75 | 41.5 | 34.75 |
| 50% | 100.5 | 36 | 61.5 | 50 |
| 75% | 150.25 | 49 | 78 | 73 |
| max | 200 | 70 | 137 | 99 |

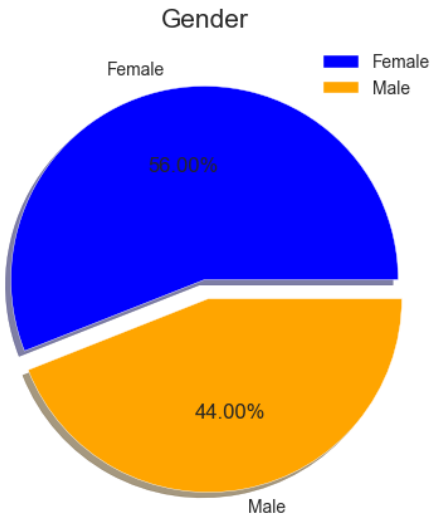
*Output* yang kedua memperlihatkan tabel 4.2 tentang data jenis satuan penting yaitu rata – rata, standar deviasi, jumlah data, data minimal, dan data maksimal. Satuan – satuan tersebut sangat penting untuk perhitungan metode klasifikasi algoritma *k-means.*

```
labels = ['Female', 'Male']
size = data['Gender'].value_counts()
colors = ['blue', 'orange']
explode = [0, 0.1]

plt.rcParams['figure.figsize'] = (8, 5)
plt.pie(size, colors = colors, explode = explode, labels = labels,
shadow = True, autopct = '%.2f%%')
plt.title('Gender', fontsize = 15)
plt.axis('off')
```

```
plt.legend()
plt.show()
```

Kode Sumber 4.4 Baris Program *Diagram Pie* Data Kolom "*Gender*"

Kode Sumber 4.4 menjelaskan proses pemrograman statistik dari kolom "*Gender*" dengan pengaturan standar. Maka saat dijalankan muncul seperti gambar dibawah.




Gambar 4.5 *Diagram Pie* Data Pelanggan Kolom "*Gender*"

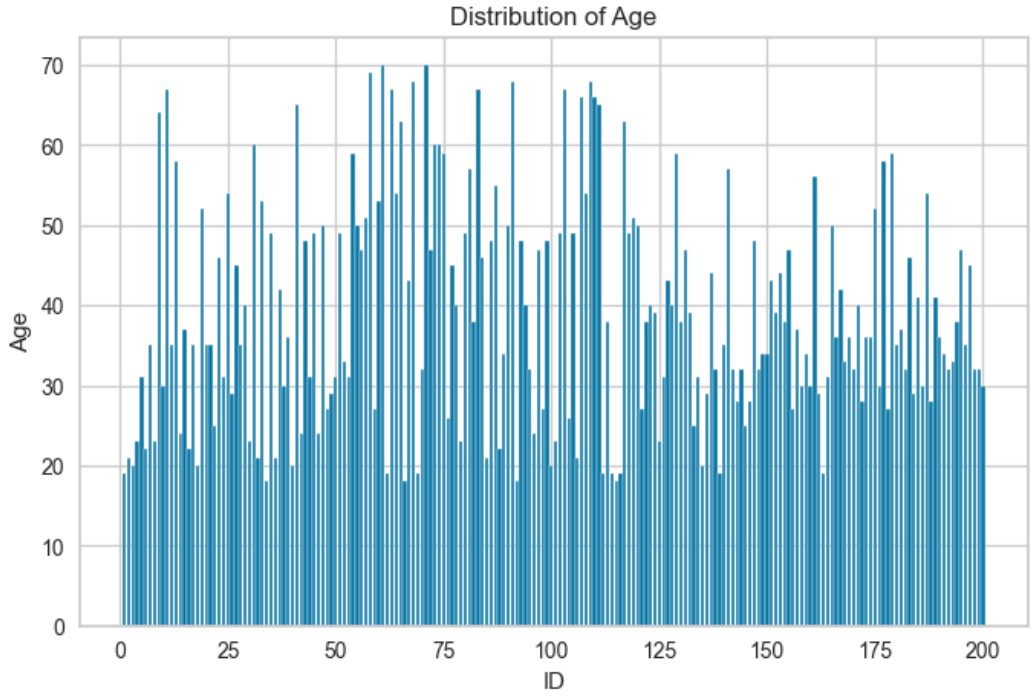
Gambar 4.5 merupakan panduan statistik kolom "*Gender*" dan menghasilkan format *Diagram Pie*. Diagram ini berukuran skala (8, 5) dengan laki – laki berwarna oranye dan perempuan berwarna biru. Bisa disaksikan bahwa perempuan mendominasi dataset dengan 56.00 % daripada laki – laki dengan 44.00 %.

```
# plot the graph
plt.rcParams['figure.figsize'] = (8, 5)
plt.bar(data['CustomerID'], data['Age'])
plt.xlabel('ID')
plt.ylabel('Age')
plt.title('Distribution of Age')
```

```
plt.show()
```

Kode Sumber 4.5 Baris Program Grafik Bar Data Kolom "*Age*"

Kode sumber 4.5 ini memprogramkan tampilan grafik bar pada kolom "*Age*". Hingga dijalankan akan muncul seperti gambar dibawah.




Gambar 4.6 Grafik Bar Data Pelanggan dengan Kolom "*Age*"

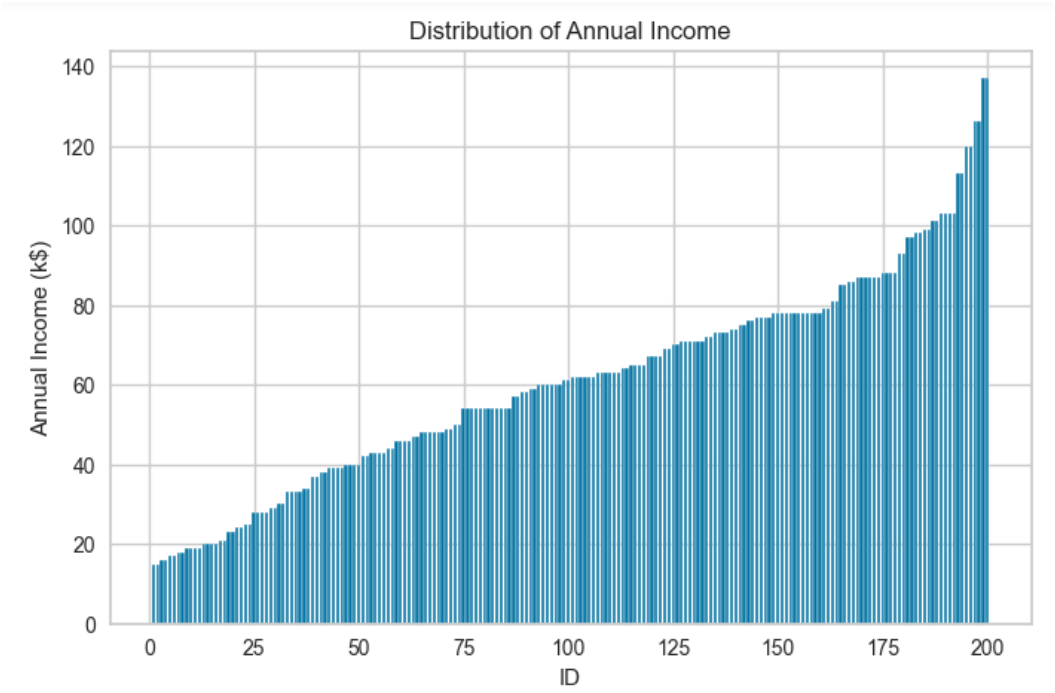
Gambar 4.6 yaitu pemuatan *grafik Bar* dari dataset kolom "*Age*". Kolom ini sangat bervariasi karena tiap pelanggan pasti memiliki umur yang berbeda. Selain itu, umur pada pelanggan minimal 17 tahun dan maksimal 70 tahun.

```
# plot the graph
plt.rcParams['figure.figsize'] = (8, 5)
plt.bar(data['CustomerID'], data['Annual Income (k$)'])
plt.xlabel('ID')
plt.ylabel('Annual Income (k$)')
plt.title('Distribution of Annual Income')
```

```
plt.show()
```

Kode Sumber 4.6 Baris Program Grafik Bar Data Kolom "*Annual Income*"

Kode sumber 4.6 ini memprogramkan tampilan grafik bar pada kolom "*Annual Income*". Hingga dijalankan akan muncul seperti gambar dibawah.




Gambar 4.7 Grafik Bar Data Pelanggan Kolom "*Annual Income*"

Gambar 4.7 lain ini berupa grafik data pelanggan dengan kolom "*Annual Income*". Pengenalan kolom tersebut mengindikasikan penghasilan dari para pelanggan maksimal sekitar 150 per ribu *dollar*. Program tersebut menghasilkan data grafik batang dengan kondisi naik. Dikarenakan pada kolom ini data diurutkan mulai penghasilan kecil ke penghasilan besar. Dilihat pada ID 70 sampai 80 grafik tersebut sejajar, yang mengindikasikan jumlah pelanggan dengan penghasilan pada gambar diatas tersebut paling banyak.

```
# plot the graph
plt.rcParams['figure.figsize'] = (8, 5)
plt.bar(data['CustomerID'], data['Spending Score (1-100)'])
plt.xlabel('ID')
plt.ylabel('Spending Score (1-100)')
plt.title('Distribution of Spending Score')
plt.show()
```

Kode Sumber 4.7 Baris Program Grafik Bar Data Kolom "*Spending Score*"

Kode sumber 4.7 ini memprogramkan tampilan grafik bar pada kolom "*Spending Score*". Hingga dijalankan akan muncul seperti gambar dibawah.


Gambar 4.8 Grafik Bar Data Pelanggan denagn Kolom "*Spending Score*"

Gambar 4.8 terakhir yaitu grafik data pelanggan dengan kolom "*Spending Score*". Pengenalan kolom tersebut mengindikasikan persentase pengeluaran pelanggan yang diambil dari penghasilan "*Annual Income*". Untuk tingkatan

pengeluaran tidak bisa diukur hanya pada grafik bar pada gambar diatas karena skor pengeluaran tersebut berhubungan dengan grafik bar penghasilan.

```
plt.figure(figsize=(8,5))
plt.scatter('Annual    Income    (k$)','Spending    Score    (1-
100)',data=data, s=30, color="red", alpha = 0.8)
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
```

Kode Sumber 4.8 Baris Program Pemetaan Data *Mall Customer*

Baris program 4.8 menjabarkan data dalam bentuk koordinat sesuai dengan nilai yang diterima. Nilai "x" menggambarkan penghasilan pelanggan dan nilai "y" menggambarkan skor pengeluaran. Hasil dari baris program seperti dibawah ini.


Gambar 4.9 Hasil Program Pemetaan Data tentang Pelanggan

Pada gambar 4.9 ini, terlihat pemetaan dataset pelanggan dengan koordinat masing – masing. Data tersebut terlihat sama dalam warnanya menandakan data ini masih mentah.

## 4.3 Pengujian Sistem

Pada sesi ini, data yang sudah diidentifikasi melalui berbagai proses akan melewati tahap pengujian dataset. Pengujian dataset merupakan langkah penting dalam analisis data yang bertujuan untuk memastikan keakuratan dan keandalan data yang telah dikumpulkan. Tahap pengujian dataset melibatkan beberapa metode dan teknik untuk menguji kualitas data. Salah satu teknik yang umum digunakan adalah validasi data, di mana data diperiksa untuk memastikan bahwa mereka sesuai dengan aturan dan format yang ditentukan sebelumnya. Misalnya, jika kita memiliki kolom tanggal dalam dataset, maka validasi dapat dilakukan untuk memastikan bahwa semua entri dalam kolom tersebut benar-benar tanggal.

Pengujian dataset juga melibatkan pemeriksaan *outlier* atau nilai ekstrim yang mungkin ada dalam data. *outlier* dapat mempengaruhi hasil analisis secara keseluruhan, oleh karena itu penting untuk mengidentifikasi dan menangani *outlier* dengan tepat. Pengujian dataset juga mencakup verifikasi konsistensi antara berbagai variabel atau atribut dalam dataset. Ini dilakukan untuk memastikan bahwa tidak ada inkonsistensi atau kontradiksi antara variabel - variabel tersebut.

Penjelasan lebih lanjut, tahap pengujian dataset adalah langkah penting dalam analisis data. Melalui pengujian ini, kita dapat memastikan keakuratan dan keandalan data yang telah dikumpulkan serta mengidentifikasi masalah potensial seperti outlier atau inkonsistensi antar variabel. Dengan melakukan pengujian

dataset secara cermat, kita dapat meningkatkan kualitas analisis data dan mengambil keputusan yang lebih baik berdasarkan informasi yang akurat. Di awal pengujian algoritma *k-means*, terlebih dahulu menjelaskan beberapa *library* yang tertera di baris program tersebut.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="darkgrid")

from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

from yellowbrick.cluster import KElbowVisualizer
from sklearn.metrics import silhouette_samples, silhouette_score

%matplotlib inline
```

Kode Sumber 4.9 Baris Program Inisialisasi Fungsi dan *Library* pada *Python*

Dari baris program 4.9, pengaplikasian *library* untuk mengaktifkan fitur pendukung yang dapat menerapkan sebuah tabel data seperti *array* dengan *library numerical python (Numpy),* menganalisis data pada *file* tabel dengan *library panda*, visualisasi hasil klasifikasi dengan *library matplotlib*, dan menerapkan pelatihan dan pengujian data dengan *library scikit learn*.

Untuk *library scikit learn*, proses klasifikasi yang diambil yaitu *k-means* sebagai fungsi *clustering* dan *standard scaler* sebagai fungsi *preprocessing*. Fungsi *clustering* menandakan bagaimana pengolahan dan penerapan program bahasa *python* bisa mengambil rumus model *k-means* dalam satu langkah *library*. Fungsi *preprocessing* menyatakan pengubahan standar nilai skala dari bilangan lebih dari nol (x > 0) diubah menjadi nilai yang berdekatan dengan 0. Karena standar nilai

pada *k-means* juga memanfaatkan nilai pada *standard scaler*, maka *library python* menyiapkan fungsi *standard scaler* dengan pemanggilan *transform*.

```
data = pd.read_csv("Mall_Customers.csv")
```

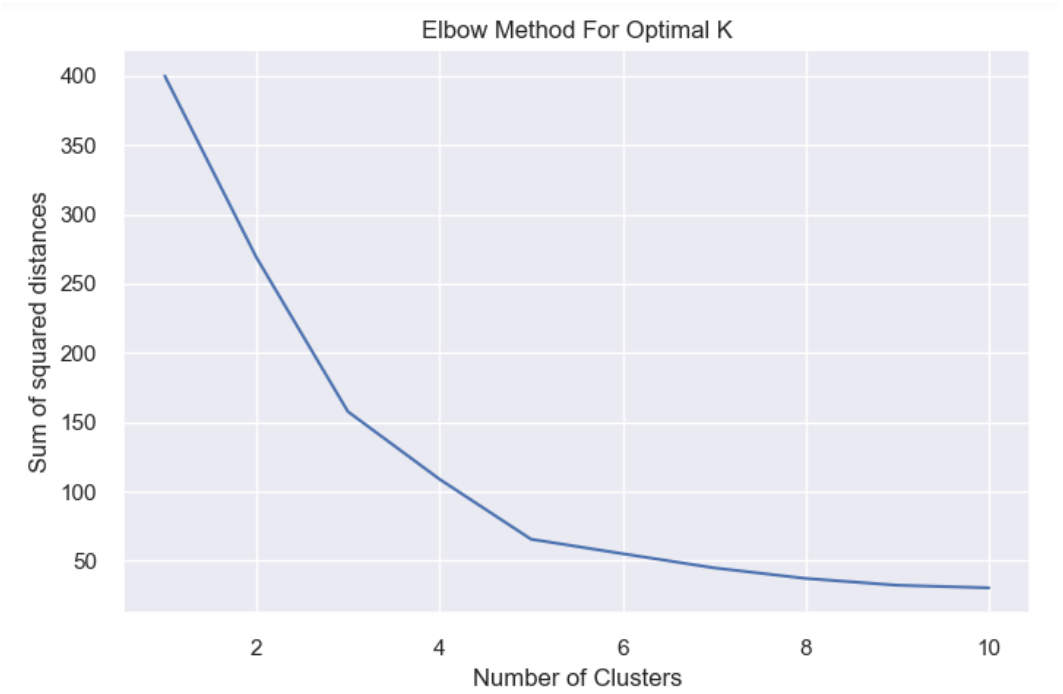
Kode Sumber 4.10 Pembacaan Dataset "*Mall_Customer*"

Seperti pada *file program* sebelumnya, dataset dibaca dengan bantuan perantara *library pandas* yang diperagakan pada baris program 4.10. Program tersebut menerima tabel dari *file* ke dalam program.

```
x= data.iloc[:,3:5]

x_array =  np.array(x)
scaler = StandardScaler()

x_scaled = scaler.fit_transform(x_array)
SSD =[]
K = range(1,11)

for k in K:
    km = KMeans(n_clusters = k)
    km = km.fit(x_scaled)
    SSD.append(km.inertia_)
plt.figure(figsize=(8,5))
plt.plot(K, SSD, 'bx-')
plt.xlabel('Number of Clusters')
plt.ylabel('Sum of squared distances')
plt.title('Elbow Method For Optimal K')
plt.show()
```

Kode Sumber 4.11 Baris Program Sakal Standar dan Optimasi Nilai K dengan *Elbow Method*

Pada baris program 4.11 tersebut merupakan proses optimasi *cluster* dari klasifikasi data pelanggan yang berjumlah 200 kustomer. Pada pembentukan *cluster*, program tersebut dibantu dengan *library scikit learn* dengan perintah *preprocessing* yaitu "*StandarScaler"*.



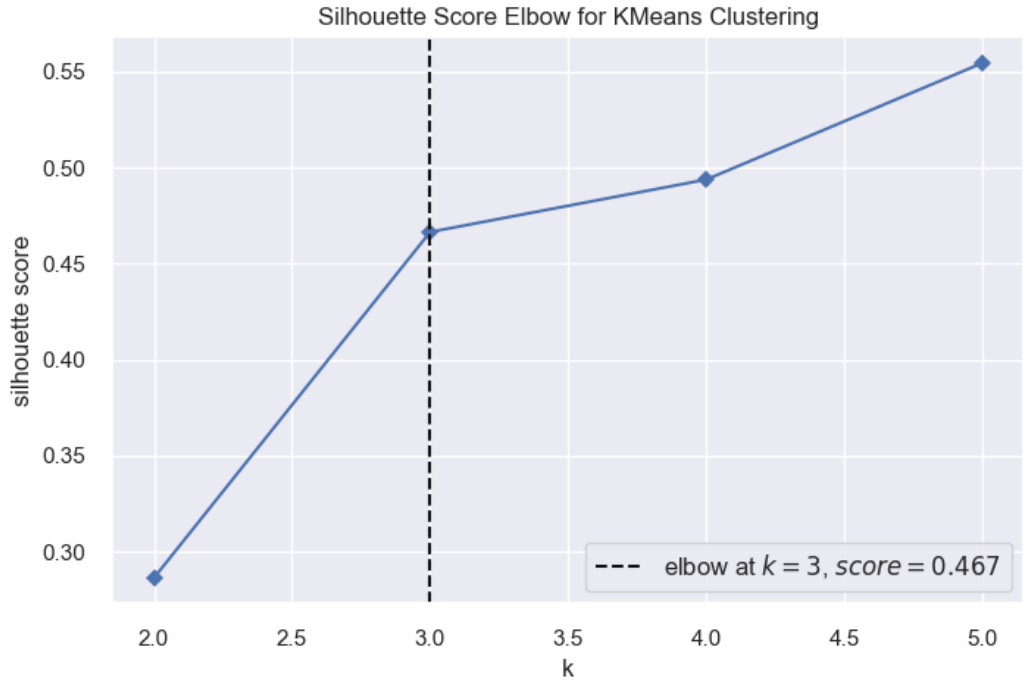

Gambar 4.10 Grafik Optimasi nilai "K" dengan *Elbow Method*

Pada gambar 4.10 menampilkan hasil dari perintah kode program tentang *elbow method*. Perintah tersebut menyeleksi *array* pelanggan dengan format (penghasilan, skor pengeluaran). Setelah itu program mempersiapkan inisialisasi setiap objek yang diambil dari beberapa materi seperti data tabel, *library*, fungsi, dan lain – lain dan juga fungsi dari program.

```
KMean= KMeans(n_clusters=5)
KMean.fit(x_scaled)
label=KMean.predict(x_scaled)

model = KMeans(random_state=123)

Visualizer        =        KElbowVisualizer(model,        k=(2,6),
metric='silhouette', timings=False)
plt.figure(figsize=(8,5))
Visualizer.fit(x_scaled)
Visualizer.poof()
```

Kode Sumber 4.12 Baris Program Proses Klasifikasi dan Pemilihan *Cluster* terbaik

Kode sumber 4.12 ini merupakan baris program paling penting yang dikerjakan yaitu bagian perhitungan algoritma *k-means* yang mengambil data – data yang dijabarkan pada masing – masing baris. Algoritma *k-means* diambil dari *library scikit learn* yang sudah lengkap dengan perintah – perintah dari *library* tersebut. Hasil dari *k-means* divisualisasikan dengan fungsi "K" dengan perbandingan *silhouette* dengan nilai pembatas antar *cluster*. Maka hasil program tersebut tercantum pada gambar dibawah.



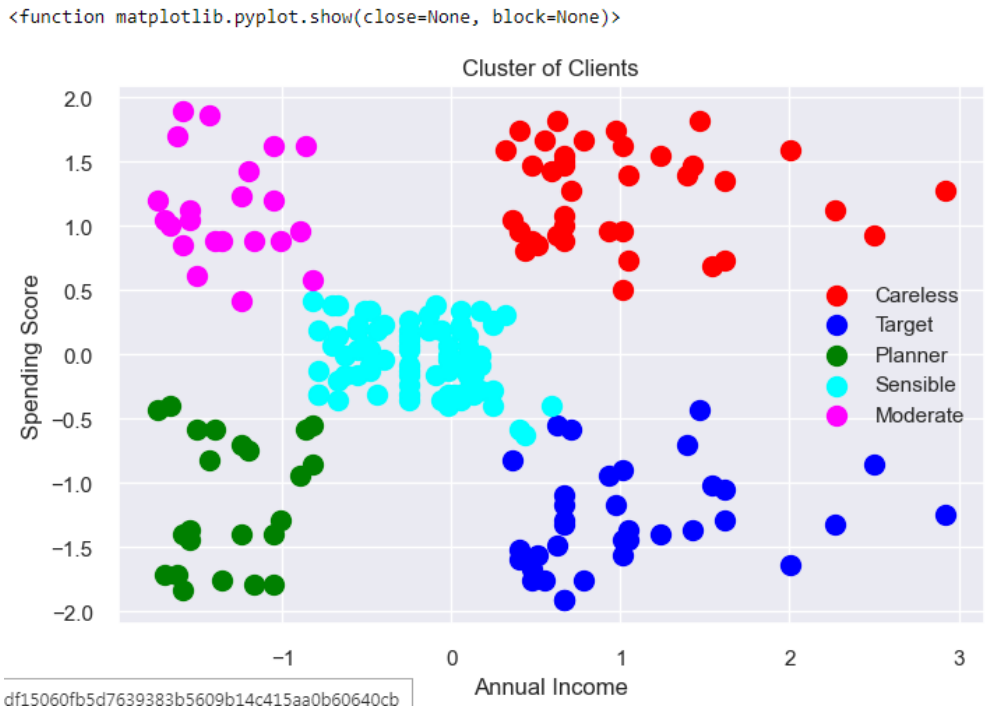

Gambar 4.11 Hasil Grafik *Elbow* dengan *Silhouette Score* pada *K-Means*

Gambar 4.11 menghasilkan data grafik *elbow* yang dihitung menggunakan *silhouette score* untuk melihat akurasi masing – masing *cluster*. Terlihat *cluster* dengan k = 3 adalah *cluster* yang mendapatkan akurasi paling tepat yaitu sekitar 0.467.

```
print(KMean.cluster_centers_)
print(KMean.labels_)
data["cluster"] = KMean.labels_
data.head()

plt.figure(figsize=(8,5))

plt.scatter(x_scaled[label==0, 0], x_scaled[label==0, 1], s=100,
c='red', label ='Careless')
plt.scatter(x_scaled[label==1, 0], x_scaled[label==1, 1], s=100,
c='blue', label ='Target')
plt.scatter(x_scaled[label==2, 0], x_scaled[label==2, 1], s=100,
c='green', label ='Planner')
plt.scatter(x_scaled[label==3, 0], x_scaled[label==3, 1], s=100,
c='cyan', label ='Sensible')
plt.scatter(x_scaled[label==4, 0], x_scaled[label==4, 1], s=100,
c='magenta', label ='Moderate')

plt.title('Cluster of Clients')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.legend()
plt.show
```

Kode Sumber 4.13 Baris Program Pemetaan Data setelah diproses *Clustering* oleh *K-Means*

Baris program 4.13 terakhir ini untuk menampilkan beberapa *output* klasifikasi algoritma *k-means* yang terdiri dari *data cluster*, *data centroid*, tabel *cluster* masing – masing data, dan grafik hasil klasifikasi dari 5 *cluster* tersebut. Untuk grafiknya, *cluster* mulai ditandai dengan label masing – masing. Ada *cluster "Careless", "Target", "Planar", "Sensible",* dan *"Moderate".* Maka hasil yang disebutkan menjadi seperti ini.




Gambar 4.12 Hasil *Clustering* Data Pelanggan

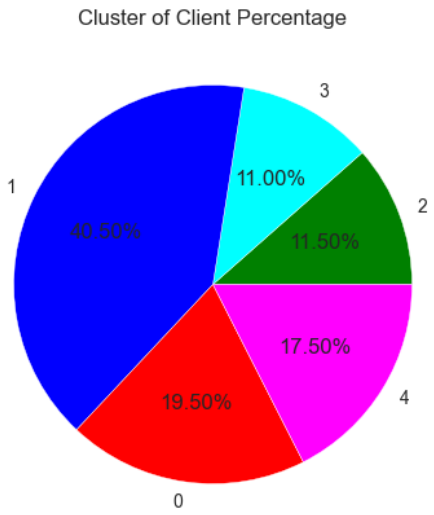
.

Gambar 4.12 menampilkan *output* algoritma *k-means* berupa pemetaan data yang telah melalui proses *clustering*. Terlihat beberapa data sudah dilabeli dan diwarnai sehingga bisa dikelompokkan sesuai cluster. Ini menunjukkan kalau *cluster* dari algoritma *k-means* telah memenuhi aturan.

```
clst = KMean.labels_
count_dict = {}
for value in clst:
    if value in count_dict:
        count_dict[value] += 1
    else:
        count_dict[value] = 1
labels = list(count_dict.keys())
sizes = list(count_dict.values())

colors = ['green', 'cyan', 'blue', 'red', 'magenta']

plt.pie(sizes, labels=labels, colors=colors, autopct='%.2f%%')
plt.title('Cluster of Client Percentage')
plt.show()
```

Kode Sumber 4.14 Program Pemetaan Dataset *Cluster* dengan Diagram Pie

Untuk pemetaan presentase *cluster k-means*, dibuatkan baris program 4.14 diatas dengan bantuan perulangan *value()* dan *library matplotlib* jenis *diagram pie.* Maka bisa dilihat persentase dari 5 jenis *cluster* berdasarkan jumlah setiap *cluster* tersebut.




Gambar 4.13 Diagram Pie 5 *Cluster K-Means* Data Pelanggan

Gambar 4.13 adalah hasil persentase dari program yang dibuat diatas dan memperlihatkan perbandingan persentase masing – masing jumlah pelanggan yang tergabung pada *cluster*. Terlihat *cluster* k = 1 memiliki persentase yang paling tinggi.

```
datac1 = (KMean.labels_ == 0)
datac2 = (KMean.labels_ == 1)
datac3 = (KMean.labels_ == 2)
datac4 = (KMean.labels_ == 3)
datac5 = (KMean.labels_ == 4)

data[datac1].iloc[:5]
data[datac2].iloc[:5]
data[datac3].iloc[:5]
data[datac4].iloc[:5]
data[datac5].iloc[:5]
```

Kode Sumber 4.15 Baris Program Tampilan Tabel 5 Cluster

Diakhir *file* pemrograman *k-means* disuguhkan baris program 4.15 yang menunjukkan penjelasan dan tampilan tabel setiap *cluster* yang dijabarkan pada gambar tabel dibawah ini.

Tabel 4.3 Data Pelanggan *Cluster* 1

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 2 | Male | 21 | 15 | 81 | 0 |
| 4 | Female | 23 | 16 | 77 | 0 |
| 6 | Female | 22 | 17 | 76 | 0 |
| 8 | Female | 23 | 18 | 94 | 0 |
| 10 | Female | 30 | 19 | 72 | 0 |

Tabel 4.3 berisikan dataset pelanggan dengan label *cluster* k = 0. Bisa dilihat, tabel ini masih mempertahankan urutan sesuai “ID”. Disini, diambil 5 data awal pelanggan karena nanti akan diteruskan pada lampiran.

Tabel 4.4 Data Pelanggan *Cluster* 2

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 86 | Male | 48 | 54 | 46 | 1 |
| 87 | Female | 55 | 57 | 58 | 1 |
| 88 | Female | 22 | 57 | 55 | 1 |
| 89 | Female | 34 | 58 | 60 | 1 |
| 90 | Female | 50 | 58 | 46 | 1 |

Tabel 4.4 berisikan dataset pelanggan dengan label *cluster* k = 1. Bisa dilihat, tabel ini masih mempertahankan urutan sesuai “ID”. Disini, diambil 5 data awal pelanggan karena nanti akan diteruskan pada lampiran.

Tabel 4.5 Data Pelanggan *Cluster* 3

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 124 | Male | 39 | 69 | 91 | 2 |
| 126 | Female | 31 | 70 | 77 | 2 |
| 128 | Male | 40 | 71 | 95 | 2 |
| 130 | Male | 38 | 71 | 75 | 2 |
| 132 | Male | 39 | 71 | 75 | 2 |

Tabel 4.5 berisikan dataset pelanggan dengan label *cluster* k = 2. Bisa dilihat, tabel ini masih mempertahankan urutan sesuai "ID". Disini, diambil 5 data awal pelanggan karena nanti akan diteruskan pada lampiran.

Tabel 4.6 Data Pelanggan *Cluster* 4

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 125 | Female | 23 | 70 | 29 | 3 |
| 129 | Male | 59 | 71 | 11 | 3 |
| 131 | Male | 47 | 71 | 9 | 3 |
| 135 | Male | 20 | 73 | 5 | 3 |
| 137 | Female | 44 | 73 | 7 | 3 |

Tabel 4.6 berisikan dataset pelanggan dengan label *cluster* k = 3. Bisa dilihat, tabel ini masih mempertahankan urutan sesuai "ID". Disini, diambil 5 data awal pelanggan karena nanti akan diteruskan pada lampiran.

Tabel 4.7 Data Pelanggan *Cluster* 5

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 | 4 |
| 3 | Female | 20 | 16 | 6 | 4 |
| 5 | Female | 31 | 17 | 40 | 4 |
| 7 | Female | 35 | 18 | 6 | 4 |
| 9 | Male | 64 | 19 | 3 | 4 |

Tabel 4.7 berisikan dataset pelanggan dengan label *cluster* k = 4. Bisa dilihat, tabel ini masih mempertahankan urutan sesuai "ID". Disini, diambil 5 data awal pelanggan karena nanti akan diteruskan pada lampiran.

## 4.4 Pembahasan

Hasil menunjukkan bagaimana grafik pada pelanggan yang tertera bisa diidentifikasi sesuai klasifikasi *cluster*. Penggunaan berbagai rumus seperti *elbow method, euclidian distance,* dan *silhouette score* menjadi titik tumpu bagi penyelesaian algoritma *k-means* bagi klasifikasi pengeluaran belanja dari pelanggan *mall*.

Dalam rumus *elbow method,* pencarian optimasi nilai "K" berbuah hasil yang baik dan terakurat. Melalui penemuan dan perhitungan letak tekukan, maka disitulah nilai "K" optimasi didapatkan. Dalam penggunaan nilai "K" tersebut dibuat untuk membagi setiap pelanggan sesuai sifat dan indentitas. Pada setiap identitas itu bisa ditelaah bagaimana pelanggan tersebut digolongkan berdasarkan

jumlah golongan tersebut hingga ditemukan jumlah *cluster* yang direkomendasikan oleh *elbow method* yaitu sebanyak 5 *cluster* (K = 5).

Dalam rumus *euclidian distance*, penentuan jarak data dengan masing – masing nilai *centroid cluster* menjelaskan data tersebut berada pada *cluster* yang mana terkait kedekatan data dengan *centroid*. Masing – masing data yang bersambungan dengan titik tumpu dari masing – masing *cluster* yang akhirnya bisa diidentifikasi data tersebut. Ini menggambarkan tentang hubungan pembeli terhadap golongan yang dibentuk melalui pendekatan. Untuk *cluster* yang disebutkan akan dijelaskan pada berikut ini:

*Cluster* 1 jenis "*Careless*", pelanggan yang dimiliki sekitar 39 orang

*Cluster* 2 jenis "*Target*", pelanggan yang dimiliki sekitar 81 orang

*Cluster* 3 jenis "*Planner*", pelanggan yang dimiliki sekitar 23 orang

*Cluster* 4 jenis "*Sensible*", pelanggan yang dimiliki sekitar 22 orang

*Cluster* 5 jenis "*Moderate*", pelanggan yang dimiliki sekitar 35 orang

Dalam rumus *silhouette score*, pencarian *cluster* terbaik dapat diprediksi melalui rumus yang dijabarkan tersebut. Hal ini bisa dipahami maksud masing – masing identifikasi *cluster*. Pada akhirnya hanya ada 1 *cluster* yang terbaik yang bisa ditiru dan dikunjungi. Ini menggambarkan bagaimana penjual menelaah kondisi pelanggan yang akan membeli produknya. Hingga akhir metode klasifikasi pengeluaran belanja pelanggan *mall* dapat meningkatkan tingkat akurasi sebesar 0.476 dengan skala standar dari perhitungan *silhouette score*. Kemudian pada *elbow method*, penelitian menunjukkan hasil "K" terbaik jatuh kepada k = 3 atau *cluster*

terbaik yaitu *cluster* 4 jenis "*Sensible*" dengan jumlah Pelanggan mencapai 22 orang.

Secara keseluruhan, hasil dari analisis klasifikasi *cluster* dan grafik yang dihasilkan memberikan wawasan yang berharga tentang pelanggan. Dengan memahami karakteristik dan preferensi mereka, perusahaan dapat mengoptimalkan strategi pemasaran mereka dan meningkatkan kepuasan pelanggan.

## 4.5 Integrasi Terhadap Pandangan Islam

Integrasi pandangan Islam dalam klasifikasi pelanggan dapat memberikan perspektif yang lebih holistik dan berkelanjutan. Dalam Islam, setiap individu dianggap sebagai makhluk unik yang memiliki kebutuhan fisik, emosional, sosial, dan spiritual. Oleh karena itu, penting bagi perusahaan untuk tidak hanya melihat pelanggan sebagai angka atau data statistik semata, tetapi juga sebagai manusia yang memiliki nilai-nilai moral dan etika.

Hubungan hasil penelitian terkait klasifikasi pelanggan dengan integrasi pandangan Islam adalah topik yang dibahas antara latar belakang dengan pembahasan implementasi. Dalam pembahasan menyimpulkan hubungan penjual dengan pelanggan yang dibentuk oleh golongan – golongan tertentu. Hingga akhirnya hanya satu golongan yang bisa memberikan kepuasan terbaik bagi penjual demi menjaga kondisi optimal dalam melakukan transaksi bisnis. Terbukti dari Q.S. Al-Baqarah/2 ayat 254 yang berbunyi:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اَنْفِقُوْا مِمَّا رَزَقْنٰكُمْ مِّنْ قَبْلِ اَنْ يَّأْتِيَ يَوْمٌ لَّا بَيْعٌ فِيْهِ وَلَا خُلَّةٌ وَّلَا شَفَاعَةٌۗ وَالْكٰفِرُوْنَ هُمُ الظّٰلِمُوْنَ

*"Wahai orang-orang yang beriman, infakkanlah sebagian dari rezeki yang telah Kami anugerahkan kepadamu sebelum datang hari (Kiamat) yang tidak ada (lagi) jual beli padanya (hari itu), tidak ada juga persahabatan yang akrab, dan tidak ada pula syafaat. Orang-orang kafir itulah orang-orang zalim."(Q.S. Al Baqarah : 254)*

*Tafsir Tahlili* dari *Nahdlatul Ulama* menyatakan, orang-orang yang beriman diperintahkan untuk menafkahkan sebagian dari kekayaan yang telah diberikan kepada mereka untuk kepentingan mereka sendiri dan keluarga mereka atau untuk kepentingan masyarakat umum. Mereka harus ingat bahwa hari kiamat akan terjadi, dan hari pembalasan akan datang. Tak ada lagi teman karib yang dapat membantu, dan tak ada lagi orang yang dapat menyelamatkan dan membantu. Anak cucu dan harta benda pun tidak dapat membantu. hanya jika mereka datang kepada Tuhan dengan hati yang bersih dan banyak amalan. Untuk hadits yang diriwayatkan oleh Ibnu Majah, Rasulullah SAW bersabda :

إِنَّمَا الْبَيْعُ عَنْ تَرَاضٍ

*"Sesungguhnya jual beli (harus) atas dasar saling ridha (suka sama suka)." (HR. Ibnu Majah no. 2185, dan dishahihkan oleh Syaikh Muhammad Nashiruddin Al Albani dalam Irwa' al-Ghalil 5/125.)*

Pada riwayat ini ditafsirkan, kaidah ini menjelaskan bahwa dipersyaratkan adanya saling ridha dalam setiap akad jual beli. Jika tidak ada keridhaan dari pelaku akad maka jual beli tersebut tidak sah. Oleh karena itu, apabila dalam akad terdapat unsur pemaksaan maka jual beli tersebut tidak sah, kecuali jika pemaksaan itu dilakukan dengan alasan yang benar, karena kata ( إِنَّمَا ) digunakan untuk pembatasan, ini menunjukkan harus ada keridhaan dalam akad jual beli.

Dalam kesimpulan, integrasi pandangan Islam dalam klasifikasi pelanggan adalah langkah penting dalam dunia bisnis modern. Hal ini tidak hanya

meningkatkan kepuasan pelanggan, tetapi juga membantu perusahaan untuk membangun hubungan yang adil dan saling menguntungkan dengan mereka. Adapau hubungan dengan latar belakang yang dari tafsiran *Al Baqarah* ayat 254 dapat menjadi ciri – ciri pelanggan dengan tolak ukur penghasilan dan skor pengeluaran, dimana pelanggan yang sering membagikan sebagian rezekinya dapat diketahui pada grafik klasifikasi dan tafsiran Hadits Riwayat Ibnu Majah yang mengetahui proses perolehan dataset pelanggan *mall* yang diikuti sebanyak 200 orang dengan pengeluaran yang direncanakan dan sah.

# BAB V

# KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Hasil pengujian algoritma *k-means* dalam metode klasifikasi pengeluaran belanja pelanggan *mall* dapat meningkatkan tingkat akurasi sebesar 0.476 dengan Skala Standar dari perhitungan *silhouette score*. Kemudian pada *elbow method*, penelitian menunjukkan hasil "K" terbaik jatuh kepada k = 3 atau *cluster* terbaik yaitu *cluster* 4 jenis "*Sensible*" dengan jumlah Pelanggan mencapai 22 orang.

Dalam akurasi yang dihimpun dari perhitungan tersebut dinyatakan pemrograman sistem aplikasi yang dikembangkan dapat melengkapi algoritma *k-means* dengan sempurna. Dalam pengujian tersebut, program algoritma *k-means* berhasil menghasilkan kelompok-kelompok data yang sesuai dengan karakteristiknya.

Keunggulan media aplikasi dalam melengkapi algoritma *k-means* terletak pada kemampuan bahasa pemrograman ini dalam memanipulasi dan menganalisis data secara efisien. Selain itu, tersedia juga banyak *library* dan modul tambahan dalam aplikasi yang dapat digunakan untuk meningkatkan kinerja algoritma *k-means.*

Dalam penelitian-penelitian sebelumnya, telah terbukti bahwa penggunaan aplikasi pemrograman dalam implementasi algoritma *k-means* memberikan hasil *clustering* yang lebih baik dibandingkan dengan bahasa pemrograman lainnya. Keunggulan sistem tersebut dalam memanipulasi dan menganalisis data, serta hasil

penelitian sebelumnya yang mendukung penggunaannya, menjadikan sistem ini sebagai pilihan yang tepat dalam implementasi algoritma *k-means*.

## 5.2 Saran

Penelitian ini masih mengalami penurunan dan masih memiliki keterbatasan dan kekurangan. Pada penyambutan kriteria klasifikasi yang dibandingkan dengan para peneliti yang dilakukan dengan data yang lain – lain seperti kepuasan pelanggan, penggunaan produk, klasifikasi karyawan *mall* terbaik, dan masih banyak lagi faktor – faktor yang membutuhkan proses metode klasifikasi.

# DAFTAR PUSTAKA


Monavia Ayu Rizaty (2023), "Jumlah Toko Retail di Indonesia Sebanyak 3,98 Juta pada 2022", https://dataindonesia.id/industri-perdagangan/detail/jumlah-toko-retail-di-indonesia-sebanyak-398-juta-pada-2022, diakses tanggal 13 November 2023

Kastanya, J., Sinay, P., & Kadbal, E.S. (2013). Pengaruh Kualitas Pelayanan terhadap Kepuasan Pelanggan Ramayana Mall Kota Sorong, *INNOVATIVE: Journal Of Social Science Research*, 3-4:5038–5052.

Yudianto, M.R.A., Kusrini, & Al Fatta, H. (2020). Analisis Pengaruh Tingkat Akurasi Klasifikasi Citra Wayang Dengan Algoritma Convolutional Neural Network, *Jurnal Teknologi Informasi*, 4-2.

Putra, B.Y., Azzahra, F.Y., & Erlanda, I.A. (2023). Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia, *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, 11-3.

Royal, S. (2020). Pemetaan Potensi Pelanggan Sebagai Strategi Promosi Pakaian Menggunakan Algoritma K-Means Clustering

Sari, R.Y., Oktavianto, H., Sulistyo, H.W. (2022). Klasterisasi Pengunjung Mall Menggunakan Algoritma K-Means Berdasarkan Pendapatan Dan Pengeluaran, *Jurnal Smart Teknologi*, 3-2.

Ballardini, A.L. (2018). A Tutorial on Particle Swarm Optimization Clustering, *Neural and Evolutionary Computing*.

Dai, J., Byrnes, P. dan Vasarhelyi, M. (2019). Are Customers Offered Appropriate Discounts? An Exploratory Study of Using Clustering Techniques in Internal Auditing, *Rutgers Stud. Account. Anal. Audit Anal. Financ. Ind.*, 59–69

Dista, T.M. dan Abdulloh, F.F. (2022). Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization, *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 6-3:1339-1348.

Parsian, M. (2015). Data Algorithm : Recipes For Scaling Up With Hadoop And Spark, *O'Reilly*

Russell, R. (2018). Machine Learning : Step-by-Step Guide To Implement Machine Learning Algorithms with Python,

Pratiwi, A.I dan Adiwijaya (2018). On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis, *Applied Computational Intelligence and Soft Computing*, 2018-5.

Aggarwal, C.C. (2015). Data Classification : Algorithms and Applications, *CRC Press Taylor & Francis Group*

Lee, S., Comuzzi, M., & Kwon, N. (2022). Exploring the Suitability of Rule-Based Classification to Provide Interpretability in Outcome-Based Process Predictive Monitoring, *Algorithms*, 15-187.

Hyde, K.K., Novack, M.N., LaHaye, N., Pelleriti, C.P., Anden, R., Dixon, D.R. & Linstead, E. (2019). Applications of Supervised Machine Learning in

Autism Spectrum Disorder Research: a Review, *Review Journal of Autism and Developmental Disorders*, 6:128–146.

Satinet, C. & Fouss, F. (2022). A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability, *Sustainability*, 14:1334.

Hidayat, S., Matsuoka, M., Baja, S. & Rampisela, D.A. (2018). Object-based image analysis for sago palm classification: Themost important features from high-resolution satellite imagery, *Remote Sens.*, 10-8.

# LAMPIRAN

# LAMPIRAN

## **Lampiran 1.** *Input* Dataset *"Mall_Customer.csv"*

Tabel. Tabel "Mall Customer.csv"

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 |
| 2 | Male | 21 | 15 | 81 |
| 3 | Female | 20 | 16 | 6 |
| 4 | Female | 23 | 16 | 77 |
| 5 | Female | 31 | 17 | 40 |
| 6 | Female | 22 | 17 | 76 |
| 7 | Female | 35 | 18 | 6 |
| 8 | Female | 23 | 18 | 94 |
| 9 | Male | 64 | 19 | 3 |
| 10 | Female | 30 | 19 | 72 |
| 11 | Male | 67 | 19 | 14 |
| 12 | Female | 35 | 19 | 99 |
| 13 | Female | 58 | 20 | 15 |
| 14 | Female | 24 | 20 | 77 |
| 15 | Male | 37 | 20 | 13 |
| 16 | Male | 22 | 20 | 79 |
| 17 | Female | 35 | 21 | 35 |
| 18 | Male | 20 | 21 | 66 |
| 19 | Male | 52 | 23 | 29 |
| 20 | Female | 35 | 23 | 98 |
| 21 | Male | 35 | 24 | 35 |
| 22 | Male | 25 | 24 | 73 |
| 23 | Female | 46 | 25 | 5 |
| 24 | Male | 31 | 25 | 73 |
| 25 | Female | 54 | 28 | 14 |
| 26 | Male | 29 | 28 | 82 |
| 27 | Female | 45 | 28 | 32 |
| 28 | Male | 35 | 28 | 61 |
| 29 | Female | 40 | 29 | 31 |
| 30 | Female | 23 | 29 | 87 |
| 31 | Male | 60 | 30 | 4 |
| 32 | Female | 21 | 30 | 73 |
| 33 | Male | 53 | 33 | 4 |
| 34 | Male | 18 | 33 | 92 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 35 | Female | 49 | 33 | 14 |
| 36 | Female | 21 | 33 | 81 |
| 37 | Female | 42 | 34 | 17 |
| 38 | Female | 30 | 34 | 73 |
| 39 | Female | 36 | 37 | 26 |
| 40 | Female | 20 | 37 | 75 |
| 41 | Female | 65 | 38 | 35 |
| 42 | Male | 24 | 38 | 92 |
| 43 | Male | 48 | 39 | 36 |
| 44 | Female | 31 | 39 | 61 |
| 45 | Female | 49 | 39 | 28 |
| 46 | Female | 24 | 39 | 65 |
| 47 | Female | 50 | 40 | 55 |
| 48 | Female | 27 | 40 | 47 |
| 49 | Female | 29 | 40 | 42 |
| 50 | Female | 31 | 40 | 42 |
| 51 | Female | 49 | 42 | 52 |
| 52 | Male | 33 | 42 | 60 |
| 53 | Female | 31 | 43 | 54 |
| 54 | Male | 59 | 43 | 60 |
| 55 | Female | 50 | 43 | 45 |
| 56 | Male | 47 | 43 | 41 |
| 57 | Female | 51 | 44 | 50 |
| 58 | Male | 69 | 44 | 46 |
| 59 | Female | 27 | 46 | 51 |
| 60 | Male | 53 | 46 | 46 |
| 61 | Male | 70 | 46 | 56 |
| 62 | Male | 19 | 46 | 55 |
| 63 | Female | 67 | 47 | 52 |
| 64 | Female | 54 | 47 | 59 |
| 65 | Male | 63 | 48 | 51 |
| 66 | Male | 18 | 48 | 59 |
| 67 | Female | 43 | 48 | 50 |
| 68 | Female | 68 | 48 | 48 |
| 69 | Male | 19 | 48 | 59 |
| 70 | Female | 32 | 48 | 47 |
| 71 | Male | 70 | 49 | 55 |
| 72 | Female | 47 | 49 | 42 |
| 73 | Female | 60 | 50 | 49 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 74 | Female | 60 | 50 | 56 |
| 75 | Male | 59 | 54 | 47 |
| 76 | Male | 26 | 54 | 54 |
| 77 | Female | 45 | 54 | 53 |
| 78 | Male | 40 | 54 | 48 |
| 79 | Female | 23 | 54 | 52 |
| 80 | Female | 49 | 54 | 42 |
| 81 | Male | 57 | 54 | 51 |
| 82 | Male | 38 | 54 | 55 |
| 83 | Male | 67 | 54 | 41 |
| 84 | Female | 46 | 54 | 44 |
| 85 | Female | 21 | 54 | 57 |
| 86 | Male | 48 | 54 | 46 |
| 87 | Female | 55 | 57 | 58 |
| 88 | Female | 22 | 57 | 55 |
| 89 | Female | 34 | 58 | 60 |
| 90 | Female | 50 | 58 | 46 |
| 91 | Female | 68 | 59 | 55 |
| 92 | Male | 18 | 59 | 41 |
| 93 | Male | 48 | 60 | 49 |
| 94 | Female | 40 | 60 | 40 |
| 95 | Female | 32 | 60 | 42 |
| 96 | Male | 24 | 60 | 52 |
| 97 | Female | 47 | 60 | 47 |
| 98 | Female | 27 | 60 | 50 |
| 99 | Male | 48 | 61 | 42 |
| 100 | Male | 20 | 61 | 49 |
| 101 | Female | 23 | 62 | 41 |
| 102 | Female | 49 | 62 | 48 |
| 103 | Male | 67 | 62 | 59 |
| 104 | Male | 26 | 62 | 55 |
| 105 | Male | 49 | 62 | 56 |
| 106 | Female | 21 | 62 | 42 |
| 107 | Female | 66 | 63 | 50 |
| 108 | Male | 54 | 63 | 46 |
| 109 | Male | 68 | 63 | 43 |
| 110 | Male | 66 | 63 | 48 |
| 111 | Male | 65 | 63 | 52 |
| 112 | Female | 19 | 63 | 54 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 113 | Female | 38 | 64 | 42 |
| 114 | Male | 19 | 64 | 46 |
| 115 | Female | 18 | 65 | 48 |
| 116 | Female | 19 | 65 | 50 |
| 117 | Female | 63 | 65 | 43 |
| 118 | Female | 49 | 65 | 59 |
| 119 | Female | 51 | 67 | 43 |
| 120 | Female | 50 | 67 | 57 |
| 121 | Male | 27 | 67 | 56 |
| 122 | Female | 38 | 67 | 40 |
| 123 | Female | 40 | 69 | 58 |
| 124 | Male | 39 | 69 | 91 |
| 125 | Female | 23 | 70 | 29 |
| 126 | Female | 31 | 70 | 77 |
| 127 | Male | 43 | 71 | 35 |
| 128 | Male | 40 | 71 | 95 |
| 129 | Male | 59 | 71 | 11 |
| 130 | Male | 38 | 71 | 75 |
| 131 | Male | 47 | 71 | 9 |
| 132 | Male | 39 | 71 | 75 |
| 133 | Female | 25 | 72 | 34 |
| 134 | Female | 31 | 72 | 71 |
| 135 | Male | 20 | 73 | 5 |
| 136 | Female | 29 | 73 | 88 |
| 137 | Female | 44 | 73 | 7 |
| 138 | Male | 32 | 73 | 73 |
| 139 | Male | 19 | 74 | 10 |
| 140 | Female | 35 | 74 | 72 |
| 141 | Female | 57 | 75 | 5 |
| 142 | Male | 32 | 75 | 93 |
| 143 | Female | 28 | 76 | 40 |
| 144 | Female | 32 | 76 | 87 |
| 145 | Male | 25 | 77 | 12 |
| 146 | Male | 28 | 77 | 97 |
| 147 | Male | 48 | 77 | 36 |
| 148 | Female | 32 | 77 | 74 |
| 149 | Female | 34 | 78 | 22 |
| 150 | Male | 34 | 78 | 90 |
| 151 | Male | 43 | 78 | 17 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 152 | Male | 39 | 78 | 88 |
| 153 | Female | 44 | 78 | 20 |
| 154 | Female | 38 | 78 | 76 |
| 155 | Female | 47 | 78 | 16 |
| 156 | Female | 27 | 78 | 89 |
| 157 | Male | 37 | 78 | 1 |
| 158 | Female | 30 | 78 | 78 |
| 159 | Male | 34 | 78 | 1 |
| 160 | Female | 30 | 78 | 73 |
| 161 | Female | 56 | 79 | 35 |
| 162 | Female | 29 | 79 | 83 |
| 163 | Male | 19 | 81 | 5 |
| 164 | Female | 31 | 81 | 93 |
| 165 | Male | 50 | 85 | 26 |
| 166 | Female | 36 | 85 | 75 |
| 167 | Male | 42 | 86 | 20 |
| 168 | Female | 33 | 86 | 95 |
| 169 | Female | 36 | 87 | 27 |
| 170 | Male | 32 | 87 | 63 |
| 171 | Male | 40 | 87 | 13 |
| 172 | Male | 28 | 87 | 75 |
| 173 | Male | 36 | 87 | 10 |
| 174 | Male | 36 | 87 | 92 |
| 175 | Female | 52 | 88 | 13 |
| 176 | Female | 30 | 88 | 86 |
| 177 | Male | 58 | 88 | 15 |
| 178 | Male | 27 | 88 | 69 |
| 179 | Male | 59 | 93 | 14 |
| 180 | Male | 35 | 93 | 90 |
| 181 | Female | 37 | 97 | 32 |
| 182 | Female | 32 | 97 | 86 |
| 183 | Male | 46 | 98 | 15 |
| 184 | Female | 29 | 98 | 88 |
| 185 | Female | 41 | 99 | 39 |
| 186 | Male | 30 | 99 | 97 |
| 187 | Female | 54 | 101 | 24 |
| 188 | Male | 28 | 101 | 68 |
| 189 | Female | 41 | 103 | 17 |
| 190 | Female | 36 | 103 | 85 |

| CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 191 | Female | 34 | 103 | 23 |
| 192 | Female | 32 | 103 | 69 |
| 193 | Male | 33 | 113 | 8 |
| 194 | Female | 38 | 113 | 91 |
| 195 | Female | 47 | 120 | 16 |
| 196 | Female | 35 | 120 | 79 |
| 197 | Female | 45 | 126 | 28 |
| 198 | Male | 32 | 126 | 74 |
| 199 | Male | 32 | 137 | 18 |
| 200 | Male | 30 | 137 | 83 |

**Lampiran 2.** *Output* Dataset *"Mall_Customer.csv" Cluster "Careless"*

Tabel. *Cluster* 1

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 2 | Male | 21 | 15 | 81 | 0 |
| 4 | Female | 23 | 16 | 77 | 0 |
| 6 | Female | 22 | 17 | 76 | 0 |
| 8 | Female | 23 | 18 | 94 | 0 |
| 10 | Female | 30 | 19 | 72 | 0 |
| 12 | Female | 35 | 19 | 99 | 0 |
| 14 | Female | 24 | 20 | 77 | 0 |
| 16 | Male | 22 | 20 | 79 | 0 |
| 18 | Male | 20 | 21 | 66 | 0 |
| 20 | Female | 35 | 23 | 98 | 0 |
| 22 | Male | 25 | 24 | 73 | 0 |
| 24 | Male | 31 | 25 | 73 | 0 |
| 26 | Male | 29 | 28 | 82 | 0 |
| 28 | Male | 35 | 28 | 61 | 0 |
| 30 | Female | 23 | 29 | 87 | 0 |
| 32 | Female | 21 | 30 | 73 | 0 |
| 34 | Male | 18 | 33 | 92 | 0 |
| 36 | Female | 21 | 33 | 81 | 0 |
| 38 | Female | 30 | 34 | 73 | 0 |
| 40 | Female | 20 | 37 | 75 | 0 |
| 42 | Male | 24 | 38 | 92 | 0 |
| 46 | Female | 24 | 39 | 65 | 0 |

**Lampiran 3.** *Output* Dataset *"Mall_Customer.csv" Cluster "Target"*

Tabel. *Cluster* 2

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 44 | Female | 31 | 39 | 61 | 1 |
| 47 | Female | 50 | 40 | 55 | 1 |
| 48 | Female | 27 | 40 | 47 | 1 |
| 49 | Female | 29 | 40 | 42 | 1 |
| 50 | Female | 31 | 40 | 42 | 1 |
| 51 | Female | 49 | 42 | 52 | 1 |
| 52 | Male | 33 | 42 | 60 | 1 |
| 53 | Female | 31 | 43 | 54 | 1 |
| 54 | Male | 59 | 43 | 60 | 1 |
| 55 | Female | 50 | 43 | 45 | 1 |
| 56 | Male | 47 | 43 | 41 | 1 |
| 57 | Female | 51 | 44 | 50 | 1 |
| 58 | Male | 69 | 44 | 46 | 1 |
| 59 | Female | 27 | 46 | 51 | 1 |
| 60 | Male | 53 | 46 | 46 | 1 |
| 61 | Male | 70 | 46 | 56 | 1 |
| 62 | Male | 19 | 46 | 55 | 1 |
| 63 | Female | 67 | 47 | 52 | 1 |
| 64 | Female | 54 | 47 | 59 | 1 |
| 65 | Male | 63 | 48 | 51 | 1 |
| 66 | Male | 18 | 48 | 59 | 1 |
| 67 | Female | 43 | 48 | 50 | 1 |
| 68 | Female | 68 | 48 | 48 | 1 |
| 69 | Male | 19 | 48 | 59 | 1 |
| 70 | Female | 32 | 48 | 47 | 1 |
| 71 | Male | 70 | 49 | 55 | 1 |
| 72 | Female | 47 | 49 | 42 | 1 |
| 73 | Female | 60 | 50 | 49 | 1 |
| 74 | Female | 60 | 50 | 56 | 1 |
| 75 | Male | 59 | 54 | 47 | 1 |
| 76 | Male | 26 | 54 | 54 | 1 |
| 77 | Female | 45 | 54 | 53 | 1 |
| 78 | Male | 40 | 54 | 48 | 1 |
| 79 | Female | 23 | 54 | 52 | 1 |
| 80 | Female | 49 | 54 | 42 | 1 |

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 81 | Male | 57 | 54 | 51 | 1 |
| 82 | Male | 38 | 54 | 55 | 1 |
| 83 | Male | 67 | 54 | 41 | 1 |
| 84 | Female | 46 | 54 | 44 | 1 |
| 85 | Female | 21 | 54 | 57 | 1 |
| 86 | Male | 48 | 54 | 46 | 1 |
| 87 | Female | 55 | 57 | 58 | 1 |
| 88 | Female | 22 | 57 | 55 | 1 |
| 89 | Female | 34 | 58 | 60 | 1 |
| 90 | Female | 50 | 58 | 46 | 1 |
| 91 | Female | 68 | 59 | 55 | 1 |
| 92 | Male | 18 | 59 | 41 | 1 |
| 93 | Male | 48 | 60 | 49 | 1 |
| 94 | Female | 40 | 60 | 40 | 1 |
| 95 | Female | 32 | 60 | 42 | 1 |
| 96 | Male | 24 | 60 | 52 | 1 |
| 97 | Female | 47 | 60 | 47 | 1 |
| 98 | Female | 27 | 60 | 50 | 1 |
| 99 | Male | 48 | 61 | 42 | 1 |
| 100 | Male | 20 | 61 | 49 | 1 |
| 101 | Female | 23 | 62 | 41 | 1 |
| 102 | Female | 49 | 62 | 48 | 1 |
| 103 | Male | 67 | 62 | 59 | 1 |
| 104 | Male | 26 | 62 | 55 | 1 |
| 105 | Male | 49 | 62 | 56 | 1 |
| 106 | Female | 21 | 62 | 42 | 1 |
| 107 | Female | 66 | 63 | 50 | 1 |
| 108 | Male | 54 | 63 | 46 | 1 |
| 109 | Male | 68 | 63 | 43 | 1 |
| 110 | Male | 66 | 63 | 48 | 1 |
| 111 | Male | 65 | 63 | 52 | 1 |
| 112 | Female | 19 | 63 | 54 | 1 |
| 113 | Female | 38 | 64 | 42 | 1 |
| 114 | Male | 19 | 64 | 46 | 1 |
| 115 | Female | 18 | 65 | 48 | 1 |
| 116 | Female | 19 | 65 | 50 | 1 |
| 117 | Female | 63 | 65 | 43 | 1 |
| 118 | Female | 49 | 65 | 59 | 1 |
| 119 | Female | 51 | 67 | 43 | 1 |

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 120 | Female | 50 | 67 | 57 | 1 |
| 121 | Male | 27 | 67 | 56 | 1 |
| 122 | Female | 38 | 67 | 40 | 1 |
| 123 | Female | 40 | 69 | 58 | 1 |
| 127 | Male | 43 | 71 | 35 | 1 |
| 133 | Female | 25 | 72 | 34 | 1 |
| 143 | Female | 28 | 76 | 40 | 1 |

**Lampiran 4.** *Output* Dataset *"Mall_Customer.csv" Cluster "Planner"*

Tabel. *Cluster* 3

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 124 | Male | 39 | 69 | 91 | 2 |
| 126 | Female | 31 | 70 | 77 | 2 |
| 128 | Male | 40 | 71 | 95 | 2 |
| 130 | Male | 38 | 71 | 75 | 2 |
| 132 | Male | 39 | 71 | 75 | 2 |
| 134 | Female | 31 | 72 | 71 | 2 |
| 136 | Female | 29 | 73 | 88 | 2 |
| 138 | Male | 32 | 73 | 73 | 2 |
| 140 | Female | 35 | 74 | 72 | 2 |
| 142 | Male | 32 | 75 | 93 | 2 |
| 144 | Female | 32 | 76 | 87 | 2 |
| 146 | Male | 28 | 77 | 97 | 2 |
| 148 | Female | 32 | 77 | 74 | 2 |
| 150 | Male | 34 | 78 | 90 | 2 |
| 152 | Male | 39 | 78 | 88 | 2 |
| 154 | Female | 38 | 78 | 76 | 2 |
| 156 | Female | 27 | 78 | 89 | 2 |
| 158 | Female | 30 | 78 | 78 | 2 |
| 160 | Female | 30 | 78 | 73 | 2 |
| 162 | Female | 29 | 79 | 83 | 2 |
| 164 | Female | 31 | 81 | 93 | 2 |
| 166 | Female | 36 | 85 | 75 | 2 |
| 168 | Female | 33 | 86 | 95 | 2 |
| 170 | Male | 32 | 87 | 63 | 2 |
| 172 | Male | 28 | 87 | 75 | 2 |
| 174 | Male | 36 | 87 | 92 | 2 |
| 176 | Female | 30 | 88 | 86 | 2 |
| 178 | Male | 27 | 88 | 69 | 2 |

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 180 | Male | 35 | 93 | 90 | 2 |
| 182 | Female | 32 | 97 | 86 | 2 |
| 184 | Female | 29 | 98 | 88 | 2 |
| 186 | Male | 30 | 99 | 97 | 2 |
| 188 | Male | 28 | 101 | 68 | 2 |
| 190 | Female | 36 | 103 | 85 | 2 |
| 192 | Female | 32 | 103 | 69 | 2 |
| 194 | Female | 38 | 113 | 91 | 2 |
| 196 | Female | 35 | 120 | 79 | 2 |
| 198 | Male | 32 | 126 | 74 | 2 |
| 200 | Male | 30 | 137 | 83 | 2 |

**Lampiran 5.** *Output* Dataset *"Mall_Customer.csv" Cluster "Sensible"*

Tabel. *Cluster* 4

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 125 | Female | 23 | 70 | 29 | 3 |
| 129 | Male | 59 | 71 | 11 | 3 |
| 131 | Male | 47 | 71 | 9 | 3 |
| 135 | Male | 20 | 73 | 5 | 3 |
| 137 | Female | 44 | 73 | 7 | 3 |
| 139 | Male | 19 | 74 | 10 | 3 |
| 141 | Female | 57 | 75 | 5 | 3 |
| 145 | Male | 25 | 77 | 12 | 3 |
| 147 | Male | 48 | 77 | 36 | 3 |
| 149 | Female | 34 | 78 | 22 | 3 |
| 151 | Male | 43 | 78 | 17 | 3 |
| 153 | Female | 44 | 78 | 20 | 3 |
| 155 | Female | 47 | 78 | 16 | 3 |
| 157 | Male | 37 | 78 | 1 | 3 |
| 159 | Male | 34 | 78 | 1 | 3 |
| 161 | Female | 56 | 79 | 35 | 3 |
| 163 | Male | 19 | 81 | 5 | 3 |
| 165 | Male | 50 | 85 | 26 | 3 |
| 167 | Male | 42 | 86 | 20 | 3 |
| 169 | Female | 36 | 87 | 27 | 3 |
| 171 | Male | 40 | 87 | 13 | 3 |
| 173 | Male | 36 | 87 | 10 | 3 |
| 175 | Female | 52 | 88 | 13 | 3 |
| 177 | Male | 58 | 88 | 15 | 3 |

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 179 | Male | 59 | 93 | 14 | 3 |
| 181 | Female | 37 | 97 | 32 | 3 |
| 183 | Male | 46 | 98 | 15 | 3 |
| 185 | Female | 41 | 99 | 39 | 3 |
| 187 | Female | 54 | 101 | 24 | 3 |
| 189 | Female | 41 | 103 | 17 | 3 |
| 191 | Female | 34 | 103 | 23 | 3 |
| 193 | Male | 33 | 113 | 8 | 3 |
| 195 | Female | 47 | 120 | 16 | 3 |
| 197 | Female | 45 | 126 | 28 | 3 |
| 199 | Male | 32 | 137 | 18 | 3 |

**Lampiran 6.** *Output* Dataset *"Mall_Customer.csv" Cluster "Moderate"*

Tabel. *Cluster* 5

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 | 4 |
| 3 | Female | 20 | 16 | 6 | 4 |
| 5 | Female | 31 | 17 | 40 | 4 |
| 7 | Female | 35 | 18 | 6 | 4 |
| 9 | Male | 64 | 19 | 3 | 4 |
| 11 | Male | 67 | 19 | 14 | 4 |
| 13 | Female | 58 | 20 | 15 | 4 |
| 15 | Male | 37 | 20 | 13 | 4 |
| 17 | Female | 35 | 21 | 35 | 4 |
| 19 | Male | 52 | 23 | 29 | 4 |
| 21 | Male | 35 | 24 | 35 | 4 |
| 23 | Female | 46 | 25 | 5 | 4 |
| 25 | Female | 54 | 28 | 14 | 4 |
| 27 | Female | 45 | 28 | 32 | 4 |
| 29 | Female | 40 | 29 | 31 | 4 |
| 31 | Male | 60 | 30 | 4 | 4 |
| 33 | Male | 53 | 33 | 4 | 4 |

| Customer ID | Gender | Age | Annual Income | Spending Score | Cluster |
|---|---|---|---|---|---|
| 35 | Female | 49 | 33 | 14 | 4 |
| 37 | Female | 42 | 34 | 17 | 4 |
| 39 | Female | 36 | 37 | 26 | 4 |
| 41 | Female | 65 | 38 | 35 | 4 |
| 43 | Male | 48 | 39 | 36 | 4 |
| 45 | Female | 49 | 39 | 28 | 4 |