BAB II
KAJIAN PUSAKA

1.1 Penelitian Terkait

Eksplorasi yang diarahkan dengan mengacu pada penelitian masa lalu sebagai hipotesis atau penemuan melalui konsekuensi dari penyelidikan masa lalu yang berbeda sangat penting dan dapat dimanfaatkan sebagai informasi pendukung. Salah satu informasi pendukung yang menurut ilmuwan harus dibuat bagian yang berbeda adalah pemeriksaan terkait yang berlaku untuk masalah yang dibicarakan dalam ulasan ini. Untuk situasi ini, titik fokus penelitian masa lalu yang digunakan sebagai sumber perspektif diidentifikasi dengan masalah kerangka informasi waktu yang terbatas. Dengan cara ini, para ahli memimpin penyelidikan beberapa penelitian tentang jenis teori dan buku harian melalui web.

Beberapa penelitian telah dilakukan sebelumnya yang dilaksanakan langsung oleh: Fina.Nasari, Surya Darma 2015 **"Penerapan K-Means Clustering pada data penerimaan Mahasiswa baru (studi kasus: universitas potensi utama)"** menuliskan mengelompokkan data mahasiswa baru dengan teknik *clustering* pengelompokan yang penulis terapkan mengunakan alogritma *K-Means Clustering*, Alogirtma Clustering mampu mengelompokan data pada kelompok yang sama dan data yang beda pada kelompok yang berbeda. Sehingga akan terlihat kelompok data mahasiswa baru hasil *K-Means Clustering* yang diperoleh ada dua kelompok, pusat *cluster* dengan *cluster* 1 = 1; 1.75; 1.5 dan cluster 2 = 2.95; 1.65; 1.4, *Cluster* pertama jika asal sekolah adalah SMA atau Sekolah menengah Pertama maka rata-rata jurusan yang di ambil adalah system informasi.

Lalu penelitian selanjutnya telah dilakukan oleh : Johan Oscar yang berjudul**"Implementasi Algoritma K-Means Clustering untuk menentukan promosistrategimarketingPresidentUniversity"**
membahas mengenai pengelolahan data mahasiswa yang telah lulus. Pengelolahan mengunakan algoritma K-Means Clustering yaitu dengan mengelompokan data mahasiswa kedalam beberapa cluster berdasarkan karakteristik data yang digunakan dalam penelitian ini sudah ditentukan yaitu nama,kota asal, jurusan dan IPK mahasiswa. Informasi yang diperoleh dari penelitian ini dapat digunakan

sebagai referensi dalam menentukan strategi promosi pemasaran yang tepat bagi departemen pemasaran.

Hasil penelitian ketiga dilakukan oleh Wiwit Agus Triyanto 2015 **"Algoritma K-Medoids untuk penentuan strategi pemasaran produk"** pengelompokan datapenjualan,sehingga akan ditemukan informasi yang dapat dig unakan untuk penentuan strategi pemasaran produk yang tepat. Hasil dari penelitian ini menghasilkan 5 cluster pertama terdiri dari 909 record transaksi, cluster kedua terdiri dari 166 record transaksi , cluster ketiga terdiri dari 66 record transaksi, cluster keempat terdiri dari 132 record transaksi. Strategi pemasaran produk dapat dilakukan dengan melakukan promosi pada cluster kelima yang memiliki kombinasi jumlah barang dibeli yang paling tinggi.

Penelitian selanjutnya yang dilakukan oleh Maulida pada tahun 2018 mengenai "**implementasi clustering dengan penerapan data mining dalam melakukan pengelompokan kunjungan wisata ke objek wisata unggulan"** di Prov. DKI Jakarta dengan pengelompokan tiga *cluster*. Sehingga menghasilkan informasi berupa nilai centroid akhir yang digunakan pada C1=15.438.488, C2=4.464.577 dan C3=342.332 dengan hasil kunjungan tertinggi terdapat di Taman Impian Jaya Ancol[5] Penelitian yang dilakukan oleh Wardhani pada tahun 2016 mengenai pengelompokan penyakit pasien pada puskesmas Kajen Pekalogan dengan menerapkan *algoritma K-Means*. Data yang dikelola sebanyak 1000 data. Inisialisasi jumlah *cluster* diolah menjadi 2 buah sesuai dengan pendefinisian nilai k. Sehingga jumlah cluster kategori akut terbagi sebanyak 376 item dan cluster kategori tidak akut sebanyak 624 item. Oleh sebab itu dilakukan penerapan algoritma K-Means dalam menentukan jumlah cluster yang sesuai.

Penelitian selanjutnya O.J et al., (2010) mengunakan Shovon & Haque, (2012) yang berjudul **"Appilcation of k-means clustering algorithm for prediction of students' academic Performance" dan "Prediction of student Academic Performance by an Application of K-Means Clustering Algorithm"**dapat digunakan untuk memonitor kinerja mahasiswa di suatu universitas. Metode ini juga dapat digunakan untuk memonitor kinerja persemester dalam meningkatkan prestasi akademik. Penelitian yang dilakukan O. J et a., (2010)

mengunakan 79 data mahasiswa untuk diuji clustering pada universitas Nigeria, sedangkan penelitian shovon & Haque, (2012) menggunakan 60 data mahasiswa untuk uji coba penelitian Penelitian Arora & Badal, (2013) yang berjudul "**Evaluating Student's Performance Using K-Means Clustering"**, menggunakan Algoritma K-Means karena dinilai dapat dengan cepat dan efisien membantu memantau perkembangan kinerja mahasiswa di suatu instasi pendidikan. Jumlah data yang dianalisis adalah 118 data siswa untuk mendaptkan nilai rata-rata mahasiswa tiap semester. Metode ini dapat memainkan peran penting bagi analisis akademik untuk menentukan alasan penurunan kinerja mahasiswa selama semester tertentu sehingga dapat diambil tidakan untuk meningkatkan kinerja tersebut disemester berikutnya.

Hasil dari penelitian diatas bahwa clustering adalah teknik pengelompokan data pada setiap kelompok yang memiliki kesamaan dan dapat digunakan sebagai acuan segmentasi sehingga menghasilkan pengelompokan data yang valid dengan mengunakan algoritma *k-means clustering* setiap segementasi mempunyai cluster semakin banyak data yang ingin dicluster semakin baik proses pengolalaan data.

1.2 Landasan Teori

1.2.1 Pengertian data mining

Data mining merupakan proses untuk menemukan pola yang menarik dan pengatahuan dalam data yang besar. Pada data mining mengunakan teknik matematika, statistik, artifical intelegent serta machine learning untuk menidentifikasi dan mengekstrasi menjadi informasi atau pengatahuan yang bermanfaat yang diambil dari berbagi database yang besar (E.Irfiani and S. S Rani. 2018)

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu :

a) Description / Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.

Sebagai contoh : mengumpulkan suara tidak akan dapat menemukan data atau kenyataan bahwa individu-individu yang tidak cukup mahir akan ditegakkan dalam perlombaan politik resmi. Penggambaran contoh dan pola secara teratur memberikan klarifikasi potensial untuk contoh atau pola.

b) Estimation / Estimasi

Penilaian secara praktis setara dengan urutan, sebenarnya variabel sasaran penilaian lebih bersifat matematis daripada kelas kategori. Model ini dibangun dengan memanfaatkan catatan total yang menawarkan manfaat dari variabel objektif sebagai nilai sekarang. Kemudian, pada saat itu pada hari berikutnya, nilai yang dinilai dari variabel tujuan dibuat tergantung pada harapan. Misalnya: pengukur tekanan peredaran darah sistolik pasien klinik darurat akan dibuat tergantung pada usia pasien, orientasi seksual, file berat badan dan tingkat natrium darah. Keterkaitan antara regangan peredaran darah sistolik dan faktor nilai prescient dalam sistem pembelajaran akan menghasilkan penilaian model. Model gambar berikutnya dapat digunakan untuk kasus baru lainnya. Model lainnya adalah penilaian nilai inspirasi penting mahasiswa pascasarjana dengan melihat catatan prestasi mahasiswa saat mengikuti program sarjana.

c) Prediction / Prediksi

Ekspektasi secara praktis setara dengan karakterisasi dan penilaian nilai, sekali lagi, sebenarnya dalam ramalan hasil akan nanti. Ilustrasi ekspektasi dalam bisnis dan eksplorasi adalah:

a. Hasil prediksi harga beras dalam empat bulan yang akan datang.

b. Prediksi persentase peningkatan kecelakaan pada lalu lintas tahun depan jika batas bawah kecepatan dinaikan. Metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan untuk keadaan yang tepat untuk prediksi.

d) Classification / Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Contoh: Penggolongan pendapatan, dapat dipisahkan dalam 3 kategori yaitu pendapatan tinggi, pendapatan sedang dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitan adalah :

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.

b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.

c. Dokter mendiagnosa penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit apa.

e) Clustering / Pengklusteran

clustering adalah kumpulan catatan, persepsi, atau persepsi dan kelas objek tempat mereka terikat. *cluster* adalah bermacam-macam record yang memiliki satu sama lain dan memiliki ketidaksamaan dengan *record* dalam kelompok yang berbeda. Pengumpulan tidak sama dengan game plan karena tidak ada variabel target dalam gathering tersebut. Pengelompokan tidak berusaha untuk menggambarkan, mengevaluasi atau mengukur nilai dari variabel tujuan. Akan melakukan perhitungan *cluster* untuk mencoba membagi informasi secara umum menjadi cluster yang memiliki kondisi *(homogen)*, kehormatan dalam satu kumpulan akan bernilai membantu dengan penilaian pada kumpulan lain akan bernilai diabaikan.

Contoh pengklusteran dalam bisnis dan penelitian adalah:

a. Terdapat kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.

b. Tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku finansial dalam baik dan mencurigakan.

c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

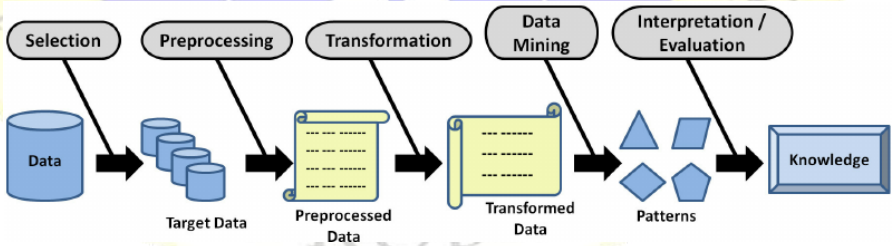
f) Association / Asosiasi

Tugas assosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut sebagai analisis keranjang belanja.

Contoh assosiasi dalam bisnis dan penelitian adalah:

a. Meneliti jumlah pelanggan dari perusahaan telekomuniasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran upgrade layanan yang diiberikan.
b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan.

Data mining adalah salah proses pada suatu tahap *Knowledge Discovery in Database* (KDD) yang terdiri dari penerapan analisis data dan penemuan algoritma yang menghasilkan enumerasi pola tertentu atau model terhadap data yang ada. Proses dalam KDD bersifat interaktif dan berulang, yang melibatkan banyak langkah dengan banyak keputusan yang dibuat oleh pengguna.



Gambar 2. 1 Tahapan Dalam KDD

1.2.2 Prapemrosesan Data

Pada Tahap prapemrosan data data adalah langkah awal untuk melakukan proses data mining pada tahap ini proses yang dilakukan yakni: *Data Cleaning, data Integration, Data Transformation, Data Mining, Pattern Evulation, Knowladge.*

a) *Data Cleaning*

Informasi yang dapat diterapkan juga lebih baik dibersihkan dengan alasan bahwa esensinya dapat mengurangi kualitas atau ketepatan hasil penambangan informasi adalah istilah yang sering digunakan untuk menggambarkan hal ini. Pembersihan informasi juga akan mempengaruhi penyajian kerangka kerja penambangan informasi karena informasi yang disimpan akan mengurangi jumlah dan kerumitan.

b) *Data Integration*

Data mining tidak hanya berasal dari kumpulan data tetapi juga berasal dari beberapa kumpulan data atau dokumen teks, pencampuran informasi dilakukan pada atribut yang mengenali zat khusus, misalnya, kredit nama, jenis barang, nomor klien, dll.

c) *Data Transformation*

Merupakan interaksi perubahan pada informasi yang telah dipilih, sehingga informasi tersebut sesuai untuk ukuran information mining. Sistem pengkodean di KDD adalah pendekatan inventif dan sangat bergantung pada jenis atau contoh data yang akan dicari dalam informasi penting.

d) *Data Mining*

Data mining adalah cara paling umum untuk mencari contoh atau data menarik sehubungan dengan informasi yang dipilih dengan menggunakan strategi atau teknik tertentu. Strategi, teknik, atau perhitungan dalam penambangan informasi umumnya berfluktuasi. Pilihan dari 17 teknik atau perhitungan yang tepat sangat bergantung pada tujuan dan ukuran KDD secara umum.

e) *Pattern Evolution*

Dalam tahap ini, efek samping dari strategi penambangan informasi adalah sebagai contoh umum dan model terkini untuk mengevaluasi apakah spekulasi saat ini pasti tercapai. Dalam hal hasil yang didapat tidak sesuai dengan teori, ada beberapa pilihan yang dapat diambil sebagai masukan untuk lebih mengembangkan langkah *data mining*,

mencoba berbagai metode data mining informasi yang lebih masuk akal, atau mengakui hasil sebagai hasil awal yang mungkin bisa membantu. Ada beberapa prosedur *data mining* yang menghasilkan hasil logis yang lebih besar seperti investigasi perkiraan. Representasi efek samping dari pemeriksaan akan sangat berguna untuk bekerja dengan pemahaman tentang konsekuensi *Data Mining*.

f) Knowledge

Tahap terakhir dari *data mining* adalah cara yang digunakan untuk membentuk pilihan atau kegiatan dari efek samping dari pemeriksaan yang diperoleh. Ada kalanya ini perlu mempengaruhi individu yang tidak memahami *data mining*. Dengan demikian, memperkenalkan konsekuensi *data mining* sebagai informasi yang dapat dilihat oleh semua orang adalah kemajuan utama yang diperlukan dalam tindakan pada data mining.

Secara teoritis, data mining tentu memiliki kelebihan dan kekurangan yakni sebagai berikut:

Kelebihan:

a. Memproses data skala besar.
b. Memungkinkan penerapan dalam masalah kompleks yang tidak terbatasi pada otak manusia.

Kekurangan:

a. Data mining belum tentu menjadi solusi untuk setiap masalah, mungkin dengan statistik sederhana solusi dapat dicapai.
b. Pengetahuan tidak diproses secara instan. (Kurniawan, Data Mining, 2019)

1.2.3 Pengertian Haji & Umrah

Umroh adalah salah satu kegiatan ibadah dalam agama islam. Hampir mirip dengan ibadah haji, ibadah ini dilaksanakan dengan cara melakukan beberapa ritual ibadah di kota suci makkah, khususnya di Masjidil Haram. Pada istilah syari'ah, umrah berarti melaksanakan tawaf di Ka'bah dan Sa'i antara shofa dan marwah, setelah memakai ihram yang diambil dari miqat. Sering disebut pula dengan haji kecil. Perbedaan umrah dengan haji adalah waktu dan tempat. Umrah dapat dilaksanakan sewaktu-waktu (setiap hari,setiap bulan,setiap tahun) dan hanya di mekkah, sedangkan haji hanya dapat dilaksanakan pada beberapa waktu antara tanggal 8 Dzullhijjah hingga 12 Dzullhijjah serta dilaksanakan sampai ke luar kota mekkah (Dena Tour Haji Umrah. 2017)

1.2.4 Pengertian Clustering

Cluster merupakan kumpulan objek data yang memilik kemiripan antara satu dengan yang lain dalam kelompok lain. Clustering atau lebih di kenal dengan analisis cluster merupakan proses pengelompokan satu set benda fisik ataupun abstrak kedalam satu kelas objek yang sama (E. Irfiani and S. S. Rani. 2018)

1.2.5 Algoritma K-Means

Algortima K-Means merupakan metode nonheirarchial yang pada awalnya mengambil sebagian dari banyak-nya komponen dari populasi untuk dijadikan pada cluster awal. Pada tahap ini tempat kelompok dipilih secara acak dari kumpulan informasi. Uji *K-Means* berikut setiap bagian dalam populasi informasi dan bagian ke salah satu tempat kelompok yang tidak diatur dalam bergantung pada jarak dasar antara bagian dengan setiap komunitas tandan akan dihitung ulang sampai semua bagian dapat dicirikan ke dalam setiap tandan tengah terakhir posisi tandan tengah lainnya dibentuk. Beberapa pemanfaatan elektif *K-Means* dengan beberapa perbaikan spekulasi komputasi terkait telah diusulkan. Manfaat menggunakan Algoritma K-Means adalah kecepatan pengumpulan objek yang lebih tinggi. Kelemahannya adalah menentukan jumlah kelompok sebelum diuji (S. Agustina, D. Yhudo, H. Santoso, N. Marnasusanto, A. Tritana, and F. 2012)

1.2.6 Pengertian Segmentasi

Segmentasi adalah kegiatan mengkelompokan pasar yang bersifat heterogen ke dalam tiap-tiap pasar yang bersifat *homogen* atau proses membagi pasar ke dalam segmen-segmen pelanggan potensial dengan kesamaan karakteristik yang menunjukkan kesamaan perilaku pembeli. Apapun jenis skema segmentasi yang kita gunakan, kuncinya adalah menyesuaikan program pemasaran untuk mengenali perbedaan pelanggan. Variabel utama segmentasi adalah: Segmentasi Geografis Segmentasi Demografis, Segmentasi Psikografis, dan Segmentasi Tingkah Laku (dosenpendidikan. 2021)

1.2.7 Promosi

Menurut Kotler, *promotion*, the fourth marketing mix tools, stand for various activities, the company undertakes to communicate its products merits and topersuade target customers to buy them. Definisi tersebut mempunyai pengertian bahwa promosi meliputi semua alat yang terdapat dalam bauran promosi yang peranan utamanya adalah mengadakan komunikasi yang bersifat membujuk (Rias Dias Ramadhanis, 2014)

1.2.8 Strategi Promosi

Variabel yang ada di dalam promotional mix ada lima, yaitu:

a. Periklanan (advertising)
b. Penjualan Personal (personal selling)
c. Promosi penjualan (sales promotion)
d. Hubungan masyarakat (public relation)
e. Pemasaran langsung (direct marketing)

(Rima Dias Ramadhanis, 2014)

1.2.9 Clustering K-Means

*Algoritma K-Mean*s merupakan algoritma pengelompokan *iterative* yang melakukan partisi set data ke dalam sejumlah *K cluster* yang sudah ditetapkan di awal. Algoritma *K-Means* sederhana untuk diimplementasikan dan dijalankan, *relative* cepat, mudah beradabtasi, umum penggunannya dalam peraktek. Secara historis, *K-Means* menjadi salah satu algoritma yang paling penting dalam bidang Data mining (Wu dan Kumar, 2009).

K-Means merupakan salah satu metode data clustering non hirarki berusaha mempartisi data yang ada kedalam bentuk satu atau lebih cluster atau kelompok. Metode ini mempartisi ke dalam cluster atau kelompok sehingga data yang dimiliki karakteristik yang sama (High intra clas similarity) dikelompokan pada kelompok yang lain. Proses clustering dimulai dengan mengidenfi mengidentifkasi data yang dicluster,

X_{ij} (i=1,…,n; j=1,…,m) dengan n adalah jumlah data yang (2.1)

Akan dicluster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap cluster ditetapkan secara bebas (sembarang), $\mathbf{C_{kj}}$ (k=1,…,k;j=1,…,m) kemudian dihitung jarak antara setiap data dengan setiap pusat cluster. Untuk melakukan penghitungan jarak data ke-i (x_i) pada pusat cluster ke-k (c_k), diberi nama (d_{ik}), dapat digunakan formula Eucliden, seperti pada persamaan (1), yaitu:

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(X_{ij} - C_{ij})^2} \quad (2.2)$$

Suatu data akan menjad anggota dari *cluster ke-k* apabila jarak data tersebut ke pusat cluster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lainnya. Hal ini dapat dihitung dengan menggunakan persamaan selanjutnya, kelompokkan data yang menjadi anggota pada setiap cluster.

$$\min \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m}(X_{ij} - C_{ij})^2} \quad (2.3)$$

Nilai pusat cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada cluster tersebut, dengan mengunakan rumus pada persamaan:

$$c_{ij} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (2.4)$$

Dimana $\boldsymbol{x}_{ij} \quad \epsilon \, cluster \, ke - k$ (2.5)

p = banyaknya anggota cluster ke k

1.2.10 Tujuan Clustering K-Means

Tujuan di balik pekerjaan pengumpulan informasi dapat dipisahkan menjadi dua, yaitu pengumpulan untuk pemahaman dan pengumpulan untuk digunakan. Jika tujuannya adalah untuk pemahaman, kumpulan yang dibentuk harus menangkap konstruksi normal informasi, biasanya sistem pengumpulan dalam tujuan ini hanyalah interaksi dasar untuk kemudian melanjutkan dengan pekerjaan pusat seperti rundown (rata-rata, standard deviasi), pelabelan kelas di setiap pertemuan untuk beberapa waktu di masa depan. sebagai penyusunan, penyiapan informasi, dll. Sementara itu, bila digunakan, tujuan pengumpulan utama biasanya untuk menemukan model pengumpulan yang paling tepat menangani informasi, memberikan pertimbangan setiap objek informasi dalam pengumpulan di mana suatu informasi ditemukan. Contoh tujuan di balik pengumpulan untuk pemahaman adalah sebagai berikut:

a. Biologi

Sebagaimana diketahui, bahwa hewan di alam tersusun oleh berbagai karakter berjenjang tertentu, menjadi alam, filum, kelas, permintaan, klan, varietas, dan spesies tertentu. Level paling signifikan adalah kerjaaan, level paling rendah adalah spesies. Satu jenis makhluk memiliki nama spesiesnya sendiri. Dua makhluk dari berbagai spesies dapat memiliki tempat dengan varietas yang sama. jumlah makhluk dengan berbagai genera dapat memiliki klan yang sama (satu kumpulan di tingkat ranah, untuk menjadi makhluk tertentu. Contoh strategi pengumpulan di berbagai bidang ilmu

pengetahuan adalah pengumpulan gen, gen yang kapasitasnya serupa.

b. Information retrieval

Situs di web mengungkap miliaran. Ketika ditanya, indeks web akan mengembalikan hasil halaman. Prosedur pengelompokan dapat digunakan untuk mengelompokkan hasil dari halaman tertentu oleh perayap web menjadi kumpulan yang lebih sederhana di mana setiap kumpulan berisi halaman dengan kualitas yang serupa atau komparatif. Misalnya, dengan kata kunci *query "movie"* dapat diberikan hasil halaman yang dibedakan dalam kategori seperti "genre", "star", "theaters", dan sebagainya setiap kategori dapat dipecah kembali menjadi subkategori yang membetuk hierarki sehingga membantu pengguna mengekspolarsi hasil query.

c. Klimatologi

Memahami iklim dunia membutuhkan pencarian contoh udara dan laut. Investigasi pengumpulan dapat diterapkan untuk menemukan contoh ketegangan gas di daerah kutub dan laut yang sebagian besar mempengaruhi iklim di darat.

d. Bisnis

Perusahaan biasanya mempunyai data informasi dalam sejumlah yang besar tentang seluruh pelanggan saat itu dan pelanggan yang berpotensi. Pengelompokan dapat diterapkan untuk memecah pelanggan ke dalam kelompok – kelompok kecil untuk analisis dan strategi pemasaran.

1.2.11 Langkah Clustering K-Means

Proses clustering dengan menggunakan algoritma K-Means memiliki langkah-langkah sebagai berikut:

a. Inisialisasi: tentukan K sebagai jumlah cluster yang diinginkan dan metrik ketidak miripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang atas perubahan fungsi objektif dan ambang batas perubahan fungsi objektif dan ambang batas perubahan fungsi objektif dan ambang batas perubahan centroid.
b. Pilih K data baru set data X sebagai centroid.
c. Alokasikan semua data ke centroid terdekat dengan metrik jarak yang sudah ditetapkan (memperbaharui ID setiap data).
d. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
e. Ulangi langkah tiga dan empat hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah dibawah ambang batas yang diingingkan; atau (b) tidak ada data yang berpindah cluster; atau (c) perubahan posisi centroid sudah dibawah ambang batas yang ditetapkan.

1.2.12 Jenis Data Dalam Set Data

Indeks informasi dapat dilihat sebagai bermacam-macam objek informasi. Satu lagi nama untuk objek informasi adalah catatan, titik, vektor, desain, peristiwa, kasus, tes, persepsi, atau elemen. Objek informasi digambarkan oleh berbagai kualitas yang menangkap atribut penting dari suatu item, misalnya, massa artikel yang sebenarnya atau waktu terjadinya. sebuah episode terjadi. Nama yang berbeda untuk menganggap adalah variabel, merek dagang, bidang, elemen atau pengukuran.

Sifat adalah sifat atau sifat atau atribut informasi yang dapat bergeser mulai dari satu artikel kemudian ke artikel berikutnya, dimulai dengan satu waktu lalu ke artikel berikutnya. Misalnya, warna kulit seseorang tidak bisa dengan warna kulit orang lain, berat badan sesorang juga bisa berubah dari waktu ke waktu. Warna

kulit bisa mempunyai nilai simbolik (hitam, putih, kuning, langsat, cokelat, sawo matang), sedangkan berat badan bisa berupa nilai angka atau numerik.

Atribut yang menjadi element setiap data mempunyai jenis yang beragam. Berat badan, pada contoh sebelumnya, mempunyai nilai numeric sehingga dapat dibandingkan suatu sama lain, sedangkan warna kulit tidak bisa dibandingkan karena menggunakan nilai yang sifatnya kualitatif. Umumnya, tipe atribut ada dua, yaitu kategoris (kualitatif) dan numeric (kuantitatif)

1.2.13 Transformasi Data

Pada setiap data yang memiliki jenis nominal seperti kota asal dan program studi harus dilakukan proses inisialisasi data terlebih dahulu ke dalam bentuk angka/numerikal. Untuk melakukan inisialisasi kota asal dapat dilakukan dengan

a. Pada daerah asal mahasiswa dilakukan pembagian wilayah-wilayah menjadi beberapa bagian wilayah.

Tabel 2. 1 Tabel Kota Asal

Kota Asal	Frekuensi	Inisial
JAWA TENGAH 1	1821	1
JAWA TENGAH 2	909	2
JAWA TENGAH 5	492	3
JAWA TENGAH 4	135	4
JAWA TENGAH 3	121	5
JAWA TIMUR	94	6
JAWA BARAT	46	7
SUMATERA SELATAN	42	8
KALIMATAN TENGAH	24	9
D.I YOGYAKARTA	23	10

SUMATERA UTARA	23	11
D.K.I JAKARTA	19	12
RIAU	11	13
NUSA TENGGARA BARAT	11	14
KALIMATAN BARAT	10	15
KALIMATAN SELATAN	8	16
SULAWESI SELATAN	7	17
KALIMATAN TIMUR	5	18
MALUKU	5	19
SUMATERA BARAT	4	20
NUSA TENGGARA TIMUR	4	21
BALI	3	22
PAPUA	3	23
SULAWESI UTARA	2	24
SULAWESI TENGAH	2	25

b. Kemudian wilayah-wilayah tersebut dlakukan pengurutan angka dari yang terbesar berdasarkan frekuensi mahasiswa yang berasal dari wilayah tersebut.

Tabel 2. 2 Program Studi

Program Studi	Kota Asal	Inisial
Teknik Informatika / S1	1110	1
Akuntansi / S1	501	2
Sistem Informasi / S1	435	3
DKV / S1	348	4
Manajemen / S1	295	5
Kesehatan Masyarakat / S1	269	6
Rekam Medis & Info Kes / D3	231	7
Teknik Informatika / D3	127	8
Sastra Inggris / S1	106	9
Broadcast / D3	104	10
Teknik Industri / S1	83	11
Sastra Jepang / S1	72	12
Manajemen Informatika / S1	72	13

c. Wilayah yang memiliki frekuensi terbesar diberi inisial dengan angka 1 dan wilayah yang memiliki frekuensi terbesar kedua diberi inisial dengan angka 2, begitu seterusnya hingga wilayah dengan frekuensi paling sedikit.Selain kota asal, program studi juga termasuk ke dalam jenis data nominal kedalam bentuk angka/numerikal.

Tabel 2. 3 hasil ininsial

No	Kota Asal	Program Studi	Nilai
1.	3	1	3,85
2.	7	2	3,64
3.	8	5	2,91
4.	1	8	3,20
5.	3	1	3,85
6.	6	1	2,99
7.	12	4	2,93
8.	6	14	3,62
9.	7	2	3,64
10.	14	13	3,02
11.	7	2	3,45
12.	14	4	3,10
13.	11	1	3,29
14.	11	1	3,63
15.	5	7	3,28
16.	8	1	2,75
17	5	4	3,11
18	3	2	3,05
19	3	18	3,14