BAB 2
LANDASAN TEORI

2.1 Teknik RFM

Teknik RFM adalah teknik segmentasi yang terdiri dari tiga ukuran: (Recency, Frequency dan Monetary), yang digabungkan menjadi kode sel RFM tiga digit, mencakup lima bagian kuintil (20% kelompok). Di antara tiga langkah RFM, Recency sering dianggap sebagai yang paling penting. Namun, menurut temuan sebelumnya, nilai RFM adalah nilai cenderung spesifik perusahaan dan didasarkan pada sifat produk [5]. Penelitian yang dilakukan oleh Fader [6] menemukan bahwa untuk Recency yang lebih rendah, pelanggan dengan frekuensi yang lebih tinggi cenderung memiliki pembelian masa depan yang memiliki potensial lebih rendah. Lumsden memiliki temuan serupa bahwa ada perbedaan yang signifikan antar kelompok Recency dan Frequency.

Proses untuk mengukur perilaku pelanggan melalui RFM modelnya adalah sebagai berikut. Pertama, data diurutkan berdasarkan masing-masing dimensi RFM dan kemudian data tersebut dibagi dalam daftar pelanggan dengan menjadi lima segmen. Metode tersebut diketahui memiliki ukuran yang sama persis. Kuintil RFM yang berbeda memiliki perbedaan tingkat respons. Untuk Recency, pelanggan diurutkan berdasarkan tanggal pembelian. Recency biasanya didefinisikan oleh jumlah periode sejak pembelian terakhir, yang mengukur interval antara transaksi terbaru waktu dan waktu analisis (hari atau bulan), yaitu, semakin rendah jumlah hari, semakin tinggi skor Recency. Pelanggan yang memiliki skor Recency yang tinggi menunjukkan kemungkinan tinggi untuk melakukan pembelian ulang. Segmen tertinggi 20% diberi kode lima, sedangkan 20% segmen berikutnya diberi kode empat dan seterusnya. Akhirnya, Recency untuk masing-masing pelanggan dalam data dilambangkan dengan angka dari lima sampai satu [7].

Untuk Frequency, database diurutkan berdasarkan pembelian frekuensi (jumlah pembelian) yang dilakukan dalam jangka waktu. Definisi Frequency sering disederhanakan untuk mempertimbangkan dua keadaan, termasuk tunggal dan berulang pembelian. Kuintil teratas diberi nilai lima dan yang lain diberi nilai empat, tiga, dua dan satu. Namun, skor frekuensi yang lebih tinggi menunjukkan loyalitas pelanggan yang lebih besar. Pelanggan yang memiliki skor frekuensi tinggi menyiratkan bahwa dia memiliki permintaan besar untuk produk dan lebih

banyak lagi cenderung membeli produk berulang kali.

Untuk Monetary, pelanggan dikodekan dengan jumlah total uang yang dihabiskan selama periode waktu tertentu. Angka dari Monetary ditentukan oleh nilai dolar yang pelanggan dihabiskan dalam periode waktu ini atau dengan jumlah uang rata-rata per pembelian atau semua pembelian hingga saat ini. Penelitian yang dilakukan oleh Marcus menyarankan bahwa lebih baik menggunakan pembelian rata-rata jumlah daripada 7 total akumulasi pembelian [8].

2.2 K-Means Clustering

K-means *clustering* adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi [9]. K-means adalah jenis algoritma yang digunakan untuk mengelompokkan data berdasarkan titik pusat data. Pengelompokkan data dilaksanakan dengan cara memaksimalkan kesamaan data pada satu bagian *cluster* dan meminimalkan kesamaan data antar *cluster*. Jarak di dalam *cluster* digunakan sebagai ukuran kesamaan. Proses pemaksimalan kesamaan data didapatkan dari jarak terpendek antara data terhadap titik pusat data [10].

Algoritma K-Means dapat terlihat pada kode semu atau pseudocode berikut

K-Means Algorithm

Input:
D= t1, t2, ... tn // Set of elements
K // Number of desired clusters
Output:
L // Set of clusters
K-Means Algorithm:
Assign initial values for m1, m2,. . . . mL
repeat
assign each item to the clusters which has the closest mean;
calculate new mean for each cluster;
until convergence criteria is met;

Pada kode semu (pseudocode) di atas, input menerima berbagai jenis elemen dan jumlah cluster yang diinginkan dan output menghasilkan cluster dengan data yang sudah dikelompokkan. Cara kerja algoritma K-Means dengan menetapkan

8 nilai awal mean *cluster* kemudian membagi data ke *cluster* yang mempunyai nilai mean terdekat dan dihitung kembali nilai mean *cluster* tersebut dikarenakan ada data terbaru yang sudah dimasukkan kemudian proses ini akan terus berulang sampai kriteria konvergensi terjadi yang berarti semua data sudah masuk pada *cluster* yang tepat dengan nilai mean yang dimiliki data itu sendiri [11].

Kemudian, proses *clustering* dimulai dengan mengidentifikasi data yang akan dibagi ke dalam *cluster*, X_{ij}(i = 1, ..., n; j = 1, ..., m) dengan n = jumlah data yang akan di *cluster* dan m = jumlah variabel. Pada awal iterasi, titik pusat data setiap *cluster* akan ditetapkan secara bebas (sembarang), C_{kj}(k = 1, ..., p; j = 1, ..., m). Selanjutnya dilakukan perhitungan jarak antara setiap data dengan setiap titik pusat data. Untuk melakukan perhitungan jarak data ke-i (X_i) pada pusat cluster ke-k (C_k), atau bisa disebut (d_{ik}), diperlukan rumus *Euclidean* seperti pada persamaan berikut [12]:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (X_{ij} - C_{kj})^2} \tag{2.1}$$

Suatu data akan menjadi anggota dari suatu *cluster* ke-k apabila jarak data ke pusat ke-k bernilai minimum apabila dibandingkan dengan jarak ke pusat cluster lainnya. Nilai ini dapat dihitung dengan menggunakan persamaan (2.2). Kemudian data-data tersebut dikelompokkan yang menjadi anggota pada setiap *cluster*.

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

2.3 Normalisasi MinMaxScaler

Teknik yang menyediakan transformasi linier pada rentang data asli disebut Normalisasi Min-Max . Teknik yang menjaga hubungan antara data asli disebut normalisasi Min-Max. Normalisasi Min-Max adalah salah satu teknik sederhana yang secara khusus dapat menyesuaikan data pada batas yang telah ditentukan dengan batas yang telah ditentukan sebelumnya. Teknik normalisasi Min-Max diperjelas dalam persamaan sebagai berikut [13]:

$$A' = \left(\frac{A \text{ - } min\ value\ of\ A}{max\ value\ of\ A \text{ - } min\ value\ of\ A}\right) * (D - C) + C \tag{2.3}$$

Dimana,

A' berisi data yang sudah dinormalisasikan

Batas yang ditentukan sebelumnya adalah [C, D]

A adalah rentang data asli

2.4 Standardisasi StandarScaler

Teknik pendekatan sebuah variabel untuk membuat kesesuaian, persamaan dan mempertahankan kesalahan pengukuran pada tingkat terendahnya. Oleh karena itu standardisasi adalah variabel acak ternormalisasi yang ditransformasikan. Teknik yang memberikan nilai atau rentang data yang dinormalisasi dari data tidak terstruktur asli menggunakan konsep seperti mean dan standar deviasi maka parameter tersebut disebut Normalisasi *Z-score* . Sehingga data tidak terstruktur dapat dinormalisasi menggunakan parameter *Z-Score*, rumus yang digunakan adalah [14]:

$$Z = \frac{X - \mu}{\sigma} \tag{2.4}$$

Dimana,

Z adalah nilai yang distandarisasi.

X adalah nilai baris dari kolom ke-i

σ adalah standar deviasi

μ adalah mean

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \mu)^2} \tag{2.5}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} V_i \tag{2.6}$$

2.5 Silhoutte Analysis

Silhoutte Analysis adalah salah satu metode evaluasi kinerja yang tidak membutuhkan satu set pelatihan data dalam mengevaluasi hasil pengelompokan. Ini membuatnya lebih tepat untuk tugas pengelompokan. Rumus *Silhoutte Analysis* $s(x_i)$ untuk titik x_i didefinisikan sebagai berikut [15]:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max[b(x_i), a(x_i)]} \tag{2.7}$$

Nilai x_i adalah sebuah elemen di dalam *cluster* k, $a(x_i)$ adalah jarak rata-rata x_i terhadap elemen lainnya di dalam *cluster* k (dengan syarat l (nilai semua elemen) tidak sama dengan nilai rata-rata x_i).

$$b(x_i) = min[d_l(x_i)], \text{di antara semua } cluster \text{ l tidak sama dengan k} \tag{2.8}$$

Nilai $d_l(x_i)$ adalah jarak rata-rata dari x_i ke semua titik dalam cluster l untuk l tidak sama dengan k (di antara perbedaan). Dari Persamaan (2.7) nilai lebar Silhouette dapat bervariasi antara -1 hingga 1. Nilai negatif tidak diinginkan karena terkait dengan kasus di mana $a(x_i)$ adalah lebih besar dari $b(x_i)$, dan rata-rata dalam perbedaan nilai yang lebih besar di antara ketidaksamaan nilai l dengan k. Nilai positif diperoleh di mana $a(x_i)$ lebih kecil dari $b(x_i)$, dan nilai *Silhouette Analysis* mencapai maksimum $s(x_i)$ = 1 untuk $a(x_i)$ = 0. Semakin besar nilai (positif) $s(x_i)$ dari an elemen, semakin tinggi kemungkinan untuk dikelompokkan dalam kelompok yang benar. Elemen dengan $s(x_i)$ negatif lebih cenderung mengelompok dalam kelompok yang salah [15]. Rata-rata dari nilai *Silhoutte Analysis* untuk sebuah *cluster* adalah rata-rata $s(x_i)$ untuk semua titik di *cluster*, dan rata-rata lebar *Silhouette* untuk seluruh hasil *clustering* adalah rata-rata $s(x_i)$ dari semua titik di setiap *cluster*.

Nilai *Silhouette* digunakan untuk mengevaluasi dan menetapkan bobot untuk setiap kernel. Kernel merupakan perluasan k-means untuk mengelompokkan objek yang secara linier dapat dipisahkan. Gagasan pengelompokan kernel bergantung pada proyeksi elemen ke dalam ruang fitur berdimensi lebih tinggi menggunakan fungsi non-linier untuk membuatnya secara linier dapat dipisahkan ke dalam ruang yang diproyeksikan. Bobot untuk kernel harus berupa nilai asli non-negatif antara 0 dan 1 dan harus berjumlah satu. Bobot non-negatif yang dinormalisasi akibatnya dihitung dengan membagi menggeser skor Silhouette ke jumlah skor Silhouette yang digeser. Selanjutnya, untuk setiap titik data, kita jumlahkan bobotnya (δ_j) ditugaskan ke kernel yang mengelompokkan titik data ke dalam kelompok yang sama. Untuk memastikan konsistensi label grup dengan kernel yang berbeda, *cluster* ditemukan oleh kernel pertama dianggap sebagai label grup

dasar. Jarak *Euclidean* antara pusat *cluster* yang ditemukan oleh setiap kernel dan pusat *cluster* yang ditemukan oleh yang pertama kernel (grup referensi) dihitung untuk menetapkan label *cluster* yang konsisten. Gugus label kemudian ditentukan berdasarkan jumlah minimum jarak *Euclidean* [16].