BAB II
TINJAUAN PUSTAKA

A. Tinjauan Pustaka

Pada penelitian yang dilakukan oleh (Chen, Sain, & Guo, 2012) berfokus untuk mengetahui pola penjualan, pelanggan mana yang paling berharga, pelanggan mana yang paling setia, pola kebiasaan pelanggan dalam melakukan pembelian pada sebuah toko online. untuk menjawab pertanyaan berikut maka metode penambangan data diterapkan pada penelitian ini. Dimana metode tersebut adalah metode klasterisasi K-Means dan metode pohon keputususan, lalu metode tersebut digabungkan dengan seperangkat metric bisnis tentang pelanggan seperti Recency, Frequency, and Monetary (RFM) dan Customer Value Life Model. Dengan menggunakan model RFM pelanggan disegmentasikan ke dalam tiap-tiap kelompok dengan metode K-Means dan metode pohon keputusan. Setelah proses segmentasi didapatkan dari proses klasterisasi maka metode pohon keputusan digunakan untuk memperbaiki segmentasi yang ada dengan cara membuat beberapa segmen bertingkat secara internal. Sehingga hasil analisis yang telah dilakukan dapat membantu bisnis lebih baik lagi dalam memahami dan perilaku pelanggan agar pemasaran lebih efektif.

Sementara itu penelitian yang dilakukan oleh (Bonnema & Waldt, 2008) meneliti mengapa praktisi komunikasi pada perguruan tinggi belum bisa mengidentifikasikan perspektif siswa dalam menentukan pilihannya untuk melanjutkan studinya di perguruan tinggi yang akan dipilih. Dalam hal pemasarannya, perguruan tinggi masih menggunakan satu pesan dalam satu media untuk semua target pasar. Sehingga hal ini dilihat kurang optimal dalam pemasaran maupun segementasi pelanggan. Untuk membantu permasalahan tersebut maka digunakan metode analisis K-Means untuk melakukan proses klasterisasi. Selanjutnya klasterisasi yang didapatkan dianalisis untuk informasi yang dibutuhkan dan preferensi sumber responden. Hasil yang didapatkan dari penelitian ini adalah informasi yang berharga pada preferensi untuk siswa yang

berencana melanjutkan studinya, selain itu juga bahwa pemasaran tidak hanya dilakukan pada satu media saja, tetapi menggunakan beragam komunikasi pemasaran yang ada.

Pada penelitian yang dilakukan oleh (Hiziroglu, Patwa, & Talwar, 2012) mencoba untuk menerapkan metode klasterisasi fuzzy dalam model pengelolaan portofolio pelanggan. Selain itu pada penelitian ini membandingkan antara metode Crisp classification dengan metode klasterisasi fuzzy. Penelitian ini diangkat karena dari semua model yang ada untuk mengelola portofolio pelanggan mempunyai keterbatasan dalam menangangi ketidakpastian, data pelanggan yang bersifat ambigu dan tidak lengkap. Sehingga pengguna metode fuzzy C-Means sebagai pendekatan untuk mengurangi ambuguitas data yang ada. Pendekatan dengan metode klasterisasi fuzzy dalam menarik sebuah kesimpulan memberikan hasil yang lebih baik. Hasil yang didapatkan dari penelitian ini adalah penggunaan klasterisasi fuzzy menghasilkan klaster yang lebih besar serta portofolio pelanggan yang lebih seimbang.

B. Landasan Teori

1. Metode *K-Means*

Dalam statistik dan mesin pembelajaran, pengelompokan *K-Means* merupakan metode analisis kelompok yang mengarah pada pemartisian N objek pengamatan ke dalam K kelompok *(cluster)* di manasetiap objek pengamatan dimiliki oleh sebuah kelompok dengan mean (rata-rata) terdekat, mirip dengan algoritma *Expectation-Maximazation* untuk *Gaussian Mixture* di mana keduanya mencoba untuk menemukan pusat dari kelompok dalam data sebanyak iterasi perbaikan yang dilakukan oleh kedua algoritma.

K-Means merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokan data ini adalah untuk

meminimalkan fungsi objektif yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antarkelompok.

Pengelompokan data dengan metode *K-Means* ini secara umum dilakukan dengan algoritma yang dijabarkan dibawah ini (Prasetyo, 2012) :

1. Tentukan K-pusat sentroid secara acak.
2. Tentukan kelompok untuk setiap data berdasarkan kedekatan terhadap sentroid masing-masing kelompok.
3. Hitung pusat kelompok yang baru dari data yang ada di masing-masing kelompok.
4. Ulangi langkah 2 sampai sentroid masing-masing kelompok tidak berubah.

Pada langkah 3 dari algoritma diatas, lokasi centroid (titik pusat) setiap kelompok yang diambil dari rata-rata (mean) semua nilai data pada setiap fiturnya harus dihitung kembali. Jika M menyatakan jumlah data sebuah kelompok, *i* menyatakan fitur ke-*i* dalam sebuah kelompok, dan p menyatakan dimensi data, untuk menghitung centroid fitur ke-*i* digunakan formula

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \qquad [2.1]$$

Formula tersebut dilakukan sebanyak *p* dimensi sehingga *i* mulai dari 1 sampai *p*.

Ada beberapa cara yang dapat digunakan untuk mengukur jarak data ke pusat kelompok, di antaranya Euclidean (Bezdek, 1981), Manhattan/*City Block* (Miyamoto dan Agusta, 1995) dan Minkowsky (Miyamoto dan Agusta, 1995). Masing-masing cara mempunyai kelebihan dan kekurangan.

Pengukuran jarak pada ruang jarak (*distance space*) Euclidean menggunakan formula

$$D(x_2,x_1) = \|x_2 - x_1\| = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2}$$

[2.2]

D adalah jarak antara data x_2 dan x_1, dan | . | adalah nilai mutlak. Pengukuran jarak pada jarak Manhattan menggunakan formula

$$D(x_2,x_1) = \|x_2 - x_1\| = \sum_{j=1}^{p} |x_{2j} - x_{1j}|$$

[2.3]

Pengukuran jarak pada ruang jarak Minkowsky menggunakan formula

$$D(x_2,x_1) = \|x_2 - x_1\| = \sqrt[\lambda]{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^{\lambda}}$$

[2.4]

λ adalah parameter jarak Minkowsky. Secara umum, λ merupakan parameter penentu dalam karakteristik jarak. Jika λ=1, ruang jarak pada Minkowsky sama dengan Manhattan. Jika λ=2, ruang jaraknya akan sama dengan Euclidean; jika λ=∞, ruang jaraknya akan sama dengan ruang jarak Chebyshev. Namun demikian, cara yang paling banyak digunakan adalah Euclidean dan Manhattan. Euclidean menjadi pilihan jika kita ingin memberi jarak terpendek antara dua titik (jarak lurus), seperti yang ditunjukkan pada formula 3.2, sedangkan Manhattan memberikan jarak terjauh pada dua data. Manhattan juga sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus, seperti keberadaan outlier, dengan lebih baik (Agusta, 2005).

Pada langkah 4 pada algoritma diatas, pengalokasian kembali data ke dalam masing-masing kelompok dalam metode *K-Means* didasarkan pada perbandingan jarak antara data dengan centroid setiap kelompok yang ada. data dialokasikan ulang secara tegas ke kelompok yang mempunyai centroid dengan jarak terdekat dari data tersebut. Pengalokasian ini dapat dirumuskan sebagai berikut (MacQueen, 1967) :

$$a_{il} = \begin{cases} 1 & d=\min\{D(x_i,C_1)\} \\ 0 & \end{cases}$$

[2.5]

a_{il} adalah nilai keanggotanan titik x_i ke pusat kelompok C_i, d adalah jarak terpendek dari data x_i ke K kelompok setelah dibandingkan, dan C_i adalah centroid (pusat kelompok) ke-*1*.

Fungsi objektif yang digunakan untuk *K-Means* ditentukan bersasarkan jarak dan nilai keanggotaan data dalam kelompok. Fungsi objektif yang digunakan adalah sebagai berikut (MacQueen, 1967) :

$$J = \sum_{i=1}^{N}\sum_{j=1}^{K} a_{ic} D(x_i, c_1)^2$$

[2.6]

N adalah jumlah data, *K* adalah jumlah kelompok, a_{il} adalah nilai keanggotaan titik data x_i ke pusat kelompok C_i, C_1 adalah pusat kelompok ke-*1*, dan $D(x_i,C_i)$ adalah jarak titik x_i ke kelompok C_1 yang diikuti. *a* mempunyai nilai 0 atau 1. Apabila suatu data merupakan anggota suatu kelompok, nilai a_{i1}=1. Jika tidak, nilai a_{i1}=0.

2. Metode *Fuzzy C-Means*

Pengelompokan dengan metode *Fuzzy C-Means* (FCM) didasarkan pada teori logika fuzzy. Teori ini pertama kali diperkenalkan oleh Lotfi Zadeh (1965) dengan nama himpunan fuzzy *(fuzzy set)*. Dalam teori fuzzy, keanggotaan sebuah data tidak diberi nilai secara tegas dengan nilai 1 (menjadi anggota) dan 0 (tidak menjadi anggota), melainkan dengan suatu nilai derajat keanggotaan yang jangkauan nilainya 0 sampai 1. Nilai keanggotaan suatu data dalam sebuah himpunan menjadi 0 ketika data sama sekali bukan anggota, dan 1 ketika data menjadi anggota secara penuh dalam suatu himpunan. Umumnya nilai keanggotaannya antara 0 dan 1. Semakin tinggi nilai keanggotaannya, semakin tinggi derajat keanggotaanya. Dikaitkan dengan *K-Means*, sebenarnya FCM merupakan versi fuzzy dari *K-Means* dengan beberapa modifikasi yang membedakannya dengan *K-Means*.

Contoh sederhana adalah umur orag. Umumnya umur orang ada dua, yaitu muda dan tua (ada juga yang lain, seperti remaja dan paruh baya, tetapi disini dicontohkan muda dan tua saja). Orang yang berumur 10 atau atau 25 disebut muda; umur 45 atau 65 disebut tua. Bagaimana jika berumur 35 tahun? Apakah muda atau tua? Jika himpunan secara tegas menyatakan batas usia muda dan tua adalah 35 tahun, orang yang berusia 35 tahun disebut muda dan yang berumur 36 tahun disebut tua. Beda antara muda dan tua sangat tegas. Untuk menentukan status umur orang apakah muda atau tua, teori fuzzy menggunakan derajat keanggotaan. Misalnya, umur 35 tahun disebut 50% muda dan 50% tua; umur 25 tahun disebut 80% muda dan 20% tua. Dengan cara fuzzy, penentuan status sebuah data pada setiap himpunan berdasarkan nilai derajat keanggotaan pada setiap himpunan.

Asumsikan ada sejumlah data dalam set data *(X)* yang berisi *m* data : x_1, $x_2, \ldots, x_m$, dinotasikan $X=\{ x_1, x_2, \ldots, x_m\}$, di mana setiap data mempunyai fitur *n* dimensi : x_{i1}, x_{i2}, ….,x_{in}, dinotasikan $x_i=\{ x_{i1}, x_{i2}, \ldots, x_{in}\}$. Ada sejumlah kelompok *C* dengan centroid : c_1, c_2,….,c_k, di mana *k* adalah jumlah kelompok. Setiap data mempunyai derajat keanggotaan pada setiap kelompok, dinyatakan dengan u_{ij}, dengan nilai di antara 0 dan 1. *i* menyatakan data x_i, dan *j* menyatakan kelompok c_j. jumlah nilai derajat keanggotaan setiap data x_i selalu sama dengan 1. Formulasinya :

$$\sum_{j=1}^{k} u_{ij} = 1$$

[2.7]

Setiap kelompok c_j berisi paling sedikit satu data dengan nilai keanggotaan tidak nol, tetapi tidak berisi derajat satu ada semua data. Formulasinya :

$$0 < \sum_{j=1}^{k} u_{ij} < m$$

[2.8]

Seperti halnya teori himpunan fuzzy yang menyatakan bahwa suatu data bisa menjadi anggota di beberapa himpunan yang dinyatakan dengan nilai derajat keanggotaan pada setiap himpunan, dalam FCM, setiap data juga menjadi anggota pada setiap kelompok dengan derajat keanggotaan u_{ij}.

Nilai keanggotaan data x_i pada kelompok v_j diformulasikan dalam

$$u_{ij} = \frac{D(x_i, c_j)^{\frac{2}{w-1}}}{\sum_{I=1}^{k} D(x_i, c_j)^{\frac{2}{w-1}}}$$

[2.9]

Parameter c_j adalah centroid kelompok-*j*, dan D() adalah jarak antara data dengan centroid. *w* adalah parameter bobot pangkat *(weighting exponent)* yang diperkenalkan dalam FCM. Tidak ada nilai ketetapan, biasanya nilai *w>1*, dan umumnya diberi nilai 2.

Untuk menghitung centroid pada kelompok c_i pada fitur *j*, kita menggunakan formula berikut :

$$c_{ij} = \frac{\sum_{I=1}^{M} (u_{il})^w x_{ij}}{\sum_{I=1}^{M} (u_{il})^w}$$

[2.10]

Parameter *M* adalah jumlah data, *w* adalah bobot pangkat, dan u_{il} adalah nilai derajat keanggotaan data x_l ke kelompok c_i. Sementara, fungsi objektif yang digunakan adalah

$$J = \sum_{i=1}^{M} \sum_{j=1}^{K} (u_{ij})^w D(x_i, c_j)^2$$

[2.11]

Secara prinsip, algoritma FCM memiliki banyak kesamaan dengan *K-Means*.

3. Pengukuran Jarak Data Biner Dengan Similaritas Jaccard

Misalkan dua obyek i dan j masing-masing diamati pada p variabel *random* diskret bertipe biner, maka tabel kontingensi dapat disajikan sebagaimana tabel 1. Pada tabel 1., nilai a dan nilai d, menunjukkan frekuensi data yang sama (*matches*), yaitu baik obyek i maupun obyek j, mempunyai kategori 0 (nol) sebanyak a, dan mempunyai kategori 1 (satu) sebanyak d. Sebaliknya, nilai *b* dan nilai *c*, menunjukkan frekuensi data yang tidak sama (*mismatches*). Secara sederhana, jika frekuensi *a* dan frekuensi *d* dijumlahkan hasilnya mendekati jumlah seluruh variabel (*p*), maka obyek *I* dan obyek *j*, dikatakan semakin mirip. Apabila $a + d = p$, maka obyek *i* dan obyek *j*, dikatakan identik.

Tabel 1. Tabel kontingensi data biner pada dua obyek

Hasil		Objek *i*		Jumlah
		1	0	
Objek *j*	1	*a*	*b*	*a* + *b*
	0	*c*	*d*	*c* + *d*
Jumlah		*a* + *c*	*b* + *d*	*p* = *a* + *b* + *c* + *d*

Untuk pengukuran similaritas antara objek *i* dan objek *j* dapat menggunakan rumus similaritas Jaccard sebagai berikut :

$$S_{ij} = \frac{a}{a + b + c}$$

[2.12]

4. Simpangan baku (*standard deviation*)

Penghitungan terhadap simpangan di bawah nilai rata-rata akan diperoleh hasil negatip sedangkan di atas nilai rata-rata akan diperoleh hasil positif. Jumlah simpangan adalah nol. Di dalam penghitungan rata-rata simpangan, simpangan baik negatip maupun positip ke duanya dianggap positip (diperhatikan harga mutlaknya) karena yang dipentingkan dalam hal ini adalah jarak antara nilai rata-

rata dengan suatu nilai tertentu. Oleh karena itu |X - X | diartikan sebagai jarak antara X dengan X . Simpangan baku biasa pula dikatakan standard deviation atau deviasi standar atau simpangan standar, adalah ukuran variabilitas yang terpenting. Simpangan baku untuk statistik diberi simbol "s" atau "SD", sedangkan untuk populasi diberi simbol δ (baca : sigma). Dalam pengertiannya simpangan baku biasa diartikan sebagai akar pangkat dua dari jumlah kuadrat simpangan dibagi banyaknya frekuensi atau banyaknya subyek. Sedangkan dalam rumus statistiknya biasa ditulis (Wismanto, 2007) :

$$s = \sqrt{\frac{\sum x^2}{N-1}} \quad \text{atau} \quad s = \sqrt{\frac{\sum (X - \bar{X})^2}{N-1}} \qquad [2.13]$$