BAB II

TINJAUAN PUSTAKA

Pada bab II akan dibahas mengenai hasil dari tinjauan pustaka dan landasan teori yang telah teruji dan dipakai sampai saat ini. Diharapkan pada bab ini pembahasan tentang penelitian terdahulu dan teori yang digunakan dapat dipahami dengan baik.

2.1 STATISTIK DESKRIPTIF

Pada awalnya statistik adalah kumpulan data, dari data kuantitatif (angka-angka) maupun data kualitatif (nilai yang didasarkan pada kualitas yang terkandung didalamnya), dimana keduanya ini memiliki makna penting dan fungsi yang besar bagi negara (Sholikhah, 1970). Namun setelah berkembangnya zaman pengertian statistik ini ada perubahan yang mana hanya dibatasi oleh data kuantitatif(Vinarti & Baskara Joni, 2018).

Sehingga pengertian statistik menjadi metode dan aturan mengenai analisa, pengumpulan, pengolahan, dan penafsiran dari data yang berupa kuantitatif yaitu mengenai angka-angka saja yang memberikan hasil pengamatan, analisa dari penjelasan data.

Secara asal usul dari kata statistik yakni berasal dari kata status (bahasa latin) yang artinya *state* (bahasa Inggris) atau kata staat (bahasa Belanda), dan dalam bahasa Indonesia diterjemahkan menjadi negara. Dahulu kala statistik hanya dipergunakan untuk kepentingan negara.

Fungsi statistik dalam berbagai aspek negara, yaitu dari segi kehidupan maupun penghidupan. Sehingga muncullah istilah statistik, dimana fungsinya disesuaikan dengan cakupan datanya. Statistik memiliki pengertian luas dan sempit. Pengertian sempit yakni statistik di pergunakan untuk menunjuk semua hasil kenyataan yang berwujud data kuantitatif tentang beberapa kejadian nyata, misalnya statistik kelahiran dan kematian, statistik tingkat pertumbuhan penduduk, dan lain sebagainya.

Pengertian luasnya statistik yakni sebagai teknik metodologi, dengan cara-cara ilmiah mengenai analisa, pengolahan, pengumpulan, penafsiran data penelitian yang berwujud data kuantitatif. Sehingga dapat menghasilkan penjelasan data yang benar dan dapat digunakan untuk mengambil sebuah keputusan yang baik.

Statistik deskriptif mempunyai 3 macam ciri pokok yang pertama bekerja dengan angka yaitu angka ini dapat menjadi jumlah (frekuensi), dan angka dapat juga menjadi sebagai nilai. Kedua bersifat objektif yaitu yang mana sebagai alat penilaian hasil nyata yang dilihat dari objektif (data), bukan dilihat dari sisi subjektif. Ketiga bersifat universal merupakan dapat digunakan untuk segala bidang penelitian yang mana berhubungan dengan data.

Statistik dibagi menjadi dua golongan yaitu statistik inferensial dan statistik deskriptif. Statistik Inferensial yang sering dikenal dengan statistik induktif adalah statistik yang memiliki aturan atau fungsinya sebagai alat dalam

menarik kesimpulan yang bersifat umum, penyusunan atau ramalan, penaksiran, dan sebagainya dari pengumpulan data yang telah disusun dan diolah.

Statistik deskriptif yang mana juga disebut statistik deduktif adalah metode yang berkaitan dengan cara pengumpulan, menyusun, mengolah, dan menyajikan suatu data sehingga dapat memberikan hasil yang jelas berupa analisa, pengelompokkan data, dan sebagainya. Statistik deskriptif ini memiliki aturan pengolahan data yang di hitung dengan Mean dan Standar Deviasi (SD). Berikut rumus Mean, seperti pada persamaan (1), yaitu ;

$$\text{Mean} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} \quad (1)$$

Dimana :

X = data ke n

Mean = nilai rata-rata data

N = banyaknya data

Berikut rumus Standar Deviasi, seperti pada persamaan (2), yaitu ;

$$\text{SD} = \sqrt{\sum \frac{(x_{1-Mean})^2 + (x_{2-Mean})^2 + \cdots + (x_{n-Mean})^2}{N-1}} \quad (2)$$

Dimana :

X = data ke n

SD = Standar Deviasi

Mean = nilai rata-rata data

N = banyaknya data

Didalam penelitian ini menggunakan statistik deskritif sebagai pembanding dari hasil pengelompokkan data.

2.2 METODE K-MEANS

Klasterisasi adalah metode penganalisaan data, dimana bertujuan untuk mengelompokkan data dengan mengklasifikasi karakteristik data yang nilainya berdekatan untuk menjasi suatu klaster atau kelompok- kelompok tertentu yang telah di tentukan(Widiarina, 2015). Klasterisasi terdapat 2 cara pendekatan, yakni pendekatan pertama adalah pendekatan partisi (*partition based clustering*) yaitu mengelompokkan data kedalam klaster-klaster yang telah ditentukan(AKBAR, 2015). Pendekatan yang kedua yakni pendekatan hirarki (*hirerarchical clustering*) yaitu mengelompokkan data dengan membetuk suatu hirarki. Dalam penelitian ini menggunakan pendekatan partisi (*partition based clustering*) yang mana metode K-means yang digunakan untuk klasterisasi.

Algoritma K-means merupakan algoritma pengelompokan iterative yang melakukan partisi set data ke dalam sejumlah K cluster yang sudah ditetapkan di awal(Lynda, Widya, & Esti, 2014). Algoritma K-means sederhana untuk diimplementasi dan dijalankan, relative cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-means menjadi salah satu algoritma yang paling penting dalam bidang data mining(Ulya, 2011).

K-means ditemukan oleh beberapa orang yaitu Lloyd (1957, 1982), Forgey (1965), Friedman dan Rubin (1967), dan McQueen (1967). K-means merupakan salah satu algoritma cluster non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok. Metode ini mengelompokkan ke dalam cluster dengan memilih acuan data random dari data yang tersedia. Dari cluster yang ditentukan secara *random* tersebut menjadi penentu keterdekatan nilai dengan data lain yang ada satu lingkup.

Menurut Kotu & Deshpande (2015: 233), klasterisasi K-means merupakan metode pengkelompokkan berbasis *prototype* yang kumpulan data di bagi menjadi cluster k. Tujuan cluster k tersebut untuk menemukan titik *prototype* data dimasukkan pada ketentuan setiap cluster yang di tetapkan(Ningrat, Maruddani, & Wuryandari, 2016). *Prototype* ini adalah sebagai pusat cluster, atau pusat massa dimana dapat menjadi rata-rata (mean) dari semua objek data dalam cluster.

Proses clustering dimulai dengan mengidentifikasi data yang akan di cluster , X_{ij}(i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan di cluster dan m adalah jumlah variabel. Pada awal iterasi, cluster yang ditentukan ditetapkan secara acak atau *random,* C_{ij}(k=1,...,k); j=1,...,l). Kemudian dihitung jarak antara setiap data dengan cluster yang telah ditentukan secara acak diawal. Berikut rumus Euclidiean Distance (ED), seperti pada persamaan (3), yaitu :

$$(X,Y)\ ,\ (A,B) = \sqrt{(X - A)^2 + (Y - B)^2} \qquad (3)$$

Suatu data akan menjadi anggota dari cluster ke-k apabila data tersebut ke pusat cluster k-k bernilai paling kecil jika dibandingkan dengan jarak ke cluster

lainnya. Kelompokan data sesuai dengan clusternya. Nilai cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data yang menjadi anggota pada cluster tersebut, dengan menggunakan rumus pada persamaan (4) :

$$C_{ij} = \frac{\sum_i^p X_{ij}}{P} \quad (4)$$

Dimana X_{ij} cluster ke –k

P = banyaknya anggota cluster ke k

Tujuan clustering data dapat ditentukan dengan tujuan clustering tersebut ditujukan untuk pemahan, dimana cluster yang terbentuk harus menangkap struktur alami data, biasanya proses clustering dalam tujuan ini hanya sebagai proses awal untuk kemudian dilanjutkan dengan peringkasan, pelabelan kelas pada setiap cluster (Pratama & Harjoko, 2017).

Sementara jika tujuan utama pengelompokan biasanya adalah mencari prototype kelompok yang paling resprentatif terhadap data, memberikan abstraksi dari setiap objek data dalam cluster dimana sebuah data terletak didalamnya. Contoh-contoh tujuan pengelompokan untuk pemahan adalah sebagai berikut :

2.2.1 Aplikasi di bidang biologi

K-means dapat digunkan untuk mengelompokkan gen berdasarkan polanya. Hal ini diperlukan untuk menemukan gen yang memiliki fungsi serupa. Contoh teknik pengelompokkan lain yakni seperti yang sudah banyak diketahui, bahwa hewan-hewan di alam ini di kelompokkan menurut karakter-karakter yaitu

kerajaan, filum, kelas, ordo, suku, genus, dan spesies. Level tertinggi adalah kerajaan, sedangkan level terendah adalah spesies.

2.2.2 Aplikasi di bidang bisnis

K-means dapat digunakan untuk melakukan segementasi pasar. Segementasi pasar adalah pengelompokkan pelanggan sesuai karakteristik pelanggan (misal: kebutuhan, gaya hidup). K-means juga dapat digunakan dalam sistem pemberi rekomendasi untuk mengelompokkan objek-objek yang saling terkait.

2.2.3 Aplikasi di bidang temu kembali informasi

K-means dapat digunakan untuk mengelompokkan dokumen sehingga memudahkan temu kembali doumen berdasarkan topiknya.

Clustertering K-means dapat digunakan mengolah data yang besar untuk melakukan pengelompokkan bertujuan memudahkan analisis data atau pemrosesan data lebih lanjut. Untuk tujuan ini, *centroid* dari cluster memegang peran lebih berarti.

Contohnya K-means dapat digunakan untuk kompresi data multimedia (citra, audio, video) yang mana ukuran dari data multimedia tersebut besar. Perhitungannya setiap objek dalam data (detik pada video, pixel pada gambar, dan frekuensi pada audio) direpresentasikan dengan *centroid* dari cluster yang memuat objek tersebut. Teknik ini disebut kuantisasi vektor.

Adapun proses klasterisasi menggunakan metode algoritma K-means ini memiliki langkah-langkah sebagai berikut :

a. Inisialisasi : tentukan berapa K sebagai jumlah cluster yang diinginkan.
b. Pilih K buah titik centroid secara acak atau *random.*
c. Lakukan perhitungan data yang tersedia dengan K yang telah dipilih, sehingga menghasilkan nilai keterdekatan masing-masing data terhadap K.
d. Kelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari setiap cluster merupakan titik centroid yang telah dipilih sebelumnya.
e. Perbaharui nilai titik centroid
f. Kemudian hitung kembali centroid berdasarkan data yang mengikuti cluster masing-masing
g. Ulangi langkah c, d, e hingga kondisi konvergen tercapai yaitu perubahan fungsi objektif sudah dibawah ambang batas yang diinginkan, atau tidak ada yang berpindah cluster atau posisi centroid sudah di ambang batas yang ditetapkan.

Keuntungan menggunakan metode *Clustering* K-means ini diantaranya yakni;

a. Sangat fleksibel, mudah adaptasi untuk di lakukan
b. Waktu yang dibutuhkan dalam melakukan pembelajaran dan perhitungan relatif lebih cepat
c. Mudah dilakukan saat implementasi dan dijalankan
d. Menggunakan prinsip yang sederhana dan dapat dijelaskan dalam non statistik

e. Sangat umum penggunaannya

Algoritma K-means ini memiliki banyak aplikasi *real time,* akan tetapi prosesnya tidak dapat dijamin sesuai dengan yang di inginkan yaitu centroid awal secara acak. Kompleksitas komputasi dari K-means cukup tinggi karena kebutuhan untuk menetapkan titik data yang cukup banyak (Wahyu, 2017).