BAB II

LANDASAN TEORI

2.1 *Text mining*

Text mining adalah sebuah proses mengambil data tidak terstruktur seperti mengambil data dari sebuah paragraf. Pemilihan target dari *text mining* dan kategori yang dijadikan acuan dapat dilakukan secara otomatis pada proses *text mining* dengan cara atau metode yang dikuasai (Lam & Lai, 2016).

Text Mining merupakan proses mengekstrak informasi penting dan sebuah pola untuk mendapatkan pengetahuan dari sebuah sumber data tekstual (Talib, Hanif, Ayesha, & Fatima, 2016).

Text Mining adalah proses mendapatkan informasi baru yang belum diketahui menggunakan komputer dengan mengekstrak data secara otomatis dari berbagai jenis sumber tertulis (Gupta & Lehal, 2009).

Text mining yang paling sesuai dengan penelitian ini adalah pengertian menurut Gepta & Lehal, 2009 yang mendefinisikan *text mining* merupakan proses mendapatkan data dengan menggunakan alat komputer untuk mengekstra data dari sumber tertulis.

2.2 *Big Data*

Data adalah informasi yang berharga dan terus naik yang memiliki informasi untuk menjelaskan kejadian yang terjadi, data semakin lama semakin berkembang begitupun data visual.

Big data didefinisikan sebagai Vs yaitu *Volume* terkait dengan jumlah data tersebut dan penggunaan *disk space*, *Velocity* merupakan aliran data yang selalu bertambah dan informasi yang ada di dalamnya, *Variety* yang berarti data berasal dari banyak sumber yang terhubung dan berbeda-beda menjadi data yang tergabung (Tonidandel, King, & Cortina, 2016).

2.3 Visualisasi Data

Tujuan utama dari analisa visual adalah untuk mengembangkan ilmu pengetahuan, metode, teknologi, dan praktek yang menjelaskan gabungan dari kemampuan manusia dan data processing elektronik. Visualisasi adalah hasil dari manusia dan komputer yang bekerja sama dengan kemampuannya untuk menghasilkan hasil yang efektif (Andrienko & Andrienko, 2012).

Data visual merupakan proses merepresentasikan data dalam bentuk tabel, grafik dari yang data yang simpel menjadi informasi yang kompleks dan interaktif (Hill, Kennedy, & Gerrard, 2016).

2.4 *Cross-Industry Standard Process for Data mining (CRISP-DM)*

Gambar 2.1 *CRISP-DM Cycle*

Sumber : (Umair, 2014)

Cross-industry standard process for data mining ditemukan oleh Daimler Chrysler pada tahun 1999. *CRISP-DM* merupakan sebuah *framework* dan petunjuk untuk *data miner* (Umair, 2014). *CRISP-DM* ini terdiri dari enam tahap yang terstruktur dan jelas, yaitu:

1. ***Business Understanding***

   Tahap pertama *CRISP-DM* berfokus pada faktor penting termasuk kriteria sukses, tujuan dari bisnis dan *data mining* dan kebutuhan yang diperlukan dan kebutuhan teknologi.

2. Data Understanding

Tahap kedua dari *CRISP-DM* yang berfokus pada pengkoleksian data, mengecek kualitas ,mengenal data, dan membuat hipotesa untuk mendapatkan informasi tersembunyi.

3. Data preparation

Tahap ketiga yang berfokus pada seleksi dan persiapan data hingga *final set*. Pada tahap ini ada banyak perubahan *record*, *table* dan *attribut*e serta pembersihan dan transformasi data.

4. Modelling

Tahap keempat yaitu pemilihan dan penerapan teknik *Modelling* yang digunakan, parameter yang digunakan menyesuaikan dengan permasalahan *data mining*.

5. Evaluation

Tahap kelima berfokus pada evaluasi dan menentukan bagaimana cara menggunakan hasil data. Hasil yang didapat tergantung algoritma dan *model* yang digunakan dan di *review* apakah sudah sesuai tujuan objektif atau belum.

6. Deployment

Tahap terakhir, tahap ini berfokus untuk mengorganisir, melaporkan dan melakukan visual data informasi yang sudah diolah.

2.5 Tahap *Data mining*

Tahap-tahap *data mining* ada 7 yaitu (Asriningtias & Mardhiyah, 2014) :

1. **Pembersihan Data (*Data Cleaning*)**

   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak relevan. Pada umumnya data yang diperoleh, baik dari database memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik.Datadata yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga mempengaruhi performasi dari teknik *data mining* karena data yang dibersihkan berkurang jumlah dan kompleksitasnya.

2. **Integrasi Data (*Data Integration*)**

   Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. **Seleksi Data (*Data Selection*)**

   Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang diambil dari database. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang

membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi Data (*Data Transformation*)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan *cluster*ing hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa *Interval*. Proses ini sering disebut transformasi data.

5. Proses *Mining*

Proses *mining* Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi Pola (*Pattern Evaluation*)

Evaluasi Pola Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasilnya berupa pola-pola yang khas maupun *model* prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

Presentasi pengetahuan merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.Tahap terakhir adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat.

Presentasi dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan.Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil *data mining* (Han, 2006).

2.6 *Naive Bayes Classifier*

Naive bayes classifier adalah salah satu ilmu probabilistic yang mengklasifikasikan data berdasarkan asumsi umum bahwa semua data adalah independent tiap data, memiliki variable khusus. Naive bayes classifier mengansumsikan bahwa setiap data terdistribusi dan dapat dijadikan *model* berdasarkan class, naïve bayse sering digunakan untuk klasifikasi text (Xu, 2016).

Naïve Bayes Classifier (NBC) merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan (*independent*) yang tinggi. Keuntungan penggunan *NBC* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variable independent, maka hanya varians dari suatu variable dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari *matriks kovarians*. Secara garis besar *model NBC* adalah sebagai berikut:

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

Rumus 2.1 *Naive Bayes Classifier*

Persamaan tersebut dapat juga digambarkan sebagai:

$$Posterior = \frac{Prior * Likelihood}{Evidence}.$$

Rumus 2.2 *Posterior Naive Bayes Classifier*

2.7 *K-means Clustering*

K-means clustering merupakan salah satu metode data *cluster*ing non-hirarki yang mengelompokan data dalam bentuk satu atau lebih *cluster*/kelompok. Data yang memiliki karakteristik yang sama dikelompokan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan *cluster*/kelompok yang lain sehingga data yang berada dalam satu *cluster*/kelompok memiliki tingkat variasi yang kecil (Agusta, 2007).

Menurut (Santosa, 2007), langkah-langkah melakukan *cluster*ing dengan metode *K-mean*s adalah sebagai berikut:

1. Pilih jumlah *cluster k.*
2. Inisialisasi *k* pusat *cluster* ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara *random*. Pusat-pusat *cluster* diberi nilai awal dengan angka-angka *random*.
3. Alokasikan semua data/ objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak paling antara satu data dengan satu *cluster* tertentu menentukan suatu data masuk dalam *cluster* mana. Untuk

menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad \ldots (1)$$

dimana:
$D\ (i,j)$ = Jarak data ke i ke pusat cluster j
X_{ki} = Data ke i pada atribut data ke k
X_{kj} = Titik pusat ke j pada atribut ke k

Rumus 2.3 Jarak *Cluster*

4. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/ objek dalam *cluster* tertentu.

   Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata (*mean*) bukan satu-satunya ukuran yang bisa dipakai.
5. Tugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *cluster*ing selesai. Atau, kembali ke langkah nomor tiga.
6. Ulang sampai pusat *cluster* tidak berubah lagi.

2.8 Tools

1. *Python*

Gambar 2.2 *Python*

Sumber : (Python, 2018)

Python adalah sebuah bahasa pemrograman yang digunakan untuk perancangan sebuah program. *Python* juga dikenal sebagai bahasa pemrograman yang menggabungkan kapabilitas, kemampuan, dan sintaks dari kode yang jelas, dan dilengkapi dengan *library* atau modul standar yang besar serta mudah digunakan.

Python mendukung berbagai teknik pemrograman seperti pemrograman berorientasi objek, pemrograman *imperative*, dan pemrograman fungsional. Salah satu fitur yang tersedia pada *python* adalah bahasa pemrograman yang dinamis yang dilengkapi dengan manajemen memori secara otomatis.

Python umumnya digunakan sebagai *script* meski pada praktiknya penggunaan bahasa ini lebih luas yaitu mencakup konteks pemanfaatan yang umumnya tidak dilakukan dengan menggunakan *script*. *Python* dapat

digunakan untuk berbagai keperluan pengembangan perangkat lunak dan dapat berjalan di berbagai *platform* sistem operasi (Rossum, 2018).

2. *PHP*

Gambar 2.3 *PHP*

Sumber : (Wikipedia, 2018)

PHP adalah singkatan dari *Hypertext Prepocessor* yaitu sebuah bahasa pemrograman yang digunakan untuk membangun, merancang dan memelihara sebuah aplikasi berbasis web dan biasanya digunakan bersamaan dengan *HTML, CSS, Jquery,* dan *Javascript*. *PHP* diciptakan oleh Rasmus Lerdorf pertama kali pada tahun 1994. Pada awalnya *PHP* adalah singkatan dari *Personal Home Page*. *PHP* berubah nama menjadi *FI* ("*Forms Interpreter*"). Sejak versi 3.0, nama bahasa ini secara resmi menjadi *Hypertext Prepocessor* dengan singkatannya yaitu *PHP*. *PHP* versi terbaru adalah versi ke-5 (PHP, 2018).

3. *Tweepy*

Gambar 2.4 *Tweepy*

Sumber : (Tweepy, 2018)

Tweepy merupakan sebuah *library python* yang digunakan untuk mengakses *Twitter API* , *library ini* cukup mudah digunakan. Sebuah situs atau aplikasi terutama *social media* sering digunakan oleh masyarakat terutama *Twitter* dimana kita bisa melakukan beberapa fitur di *Twitter* seperti *tweet*, *reply*, dan *retweet*. Data yang diambil dari sebuah *Social media* perlu menggunakan suatu teknologi bernama *API* sebagai koneksi antar program dengan *social media*. *Tweepy* adalah *library python* yang bisa mengakses API itu secara langsung di *console* maupun *script* (Tweepy, 2018).

4. ***Power BI***

Gambar 2.5 *Power BI*

Sumber : (PowerBI, 2018)

Power BI adalah *tools* bisnis analitik yang digunakan untuk menganalisa data dan membagikan informasi pengetahuan ke orang banyak. Dashboard dalam *Power BI* memberikan pandangan 360 derajat kepada *user* bisnis dengan data yang penting diperbaharui secara cepat dan dapat diakses oleh banyak perangkat (PowerBI, 2018).

5. R

Gambar 2.6 *R Script*

Sumber : (Wikipedia, 2018)

R Script adalah sebuah bahasa pemrograman dan perangkat lunak. *R Script* biasanya digunakan untuk analisis statistik dan grafik. *R* dibuat oleh Ross Ihaka dan Robert Gentleman di Universitas Auckland, Selandia Baru, dan kini dikembangkan oleh *R Development Core Team*. *R* dinamakan berasal dari nama dua pembuat nya yaitu Robert Gentleman dan Ross Ihaka.

Bahasa pemrograman *R* kini menjadi bahasa yang sering digunakan oleh statistikawan untuk pengembangan perangkat lunak statistik, serta digunakan secara luas untuk pengembangan perangkat lunak statistik dan analisis data. *R* menggunakan tampilan antarmuka berupa baris perintah tetapi antarmuka pengguna berupa grafik juga tersedia.

R menyediakan berbagai teknik statistik (*model linear* dan *nonlinear*, uji statistik klasik, analisis deret waktu, klasifikasi, klusterisasi, dan sebagainya) serta grafik. *R* dirancang sebagai bahasa komputer yang mengizinkan penggunanya untuk menambah fungsi tambahan dengan

mendefinisikan fungsi baru. Kekuatan besar dari *R* yang lain adalah fasilitas grafik nya, yang menghasilkan grafik dengan kualitas publikasi yang dapat memuat simbol matematika. R memiliki format dokumentasi seperti *LaTeX* yang digunakan sebagai penyedia dokumentasi yang lengkap baik secara *daring* (dalam berbagai format) maupun secara cetakan (Wikipedia, 2018).

2.9 Grafik

1) Line Chart

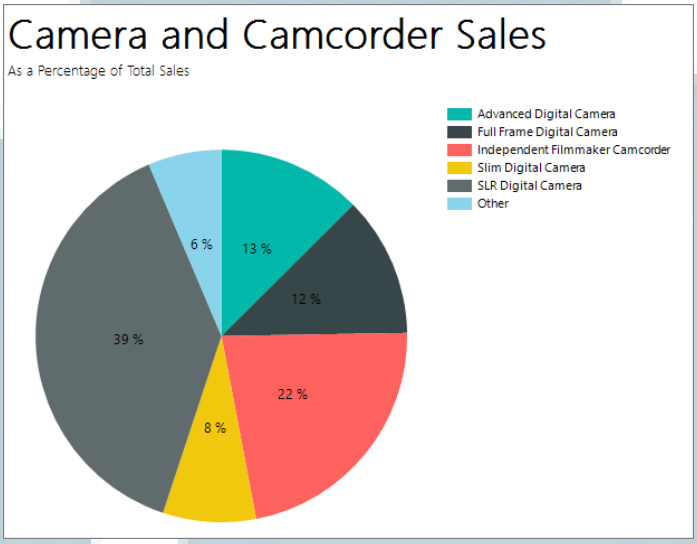
Grafik Garis merupakan grafik yang biasanya digunakan untuk menggambarkan perkembangan dan perubahan sesuai durasi waktu yang ditampilkan.

Gambar 2.7 *Line Chart*

Sumber : (Microsoft, 2018)

2) *Pie Chart*

Pie chart merepresentasikan data dalam bentuk pie atau lingkaran yang menampilkan jumlah dan presentase data sesuai ukuran dari keseluruhan data.

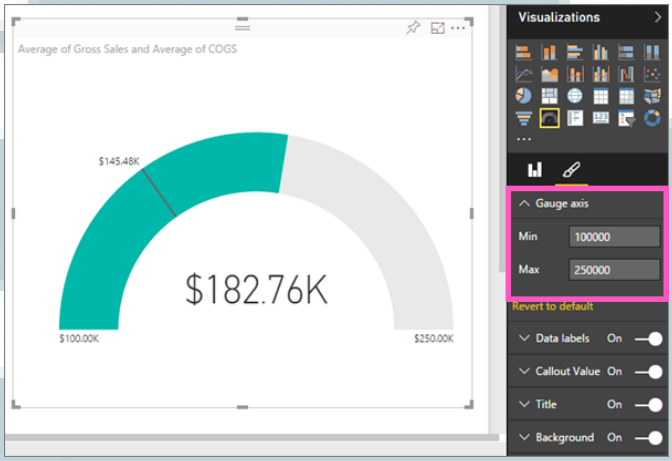


Gambar 2.8 *Pie Chart*

Sumber : (Microsoft, 2018)

3) Gauge Chart

Gauge chart menggambarkan sumbuh minimal dan sumbuh maksimal dari suatu data dengan isi rata-rata dari data tersebut.

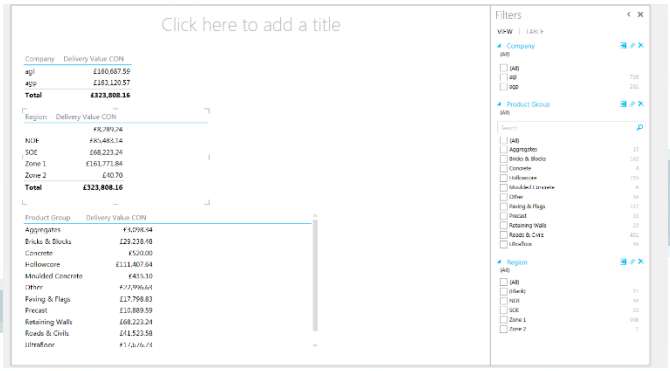


Gambar 2.9 *Gauge Chart*

Sumber : (Microsoft, 2018)

4) Tabel

Tabel juga dapat digunakan untuk menampilkan informasi data dalam bentuk kolom dan baris.

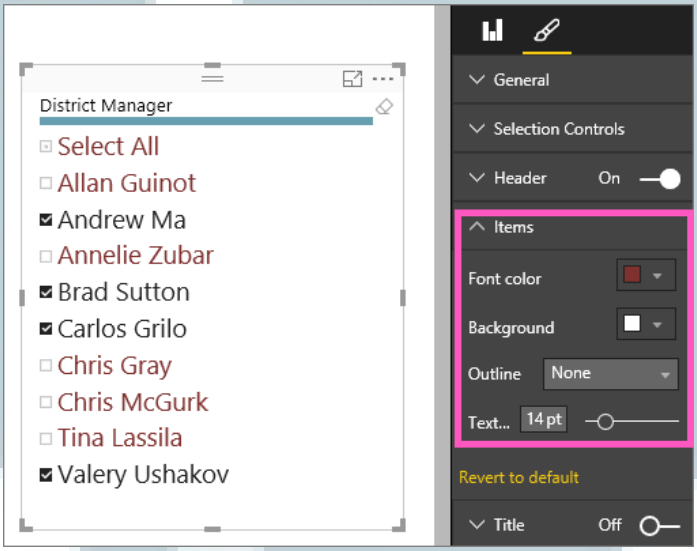


Gambar 2.10 *Power BI* Tabel

Sumber : (Microsoft, 2018)

5) *Slicer*

Slicer merupakan *filter* visualisasi data merupakan salah satu feature *Power BI*, *slicer* ini menampilkan data *unique* (hanya data yang berbeda di kolom tersebut) dan dapat digunakan sebagai *filter* untuk menampilkan data di grafik lain.

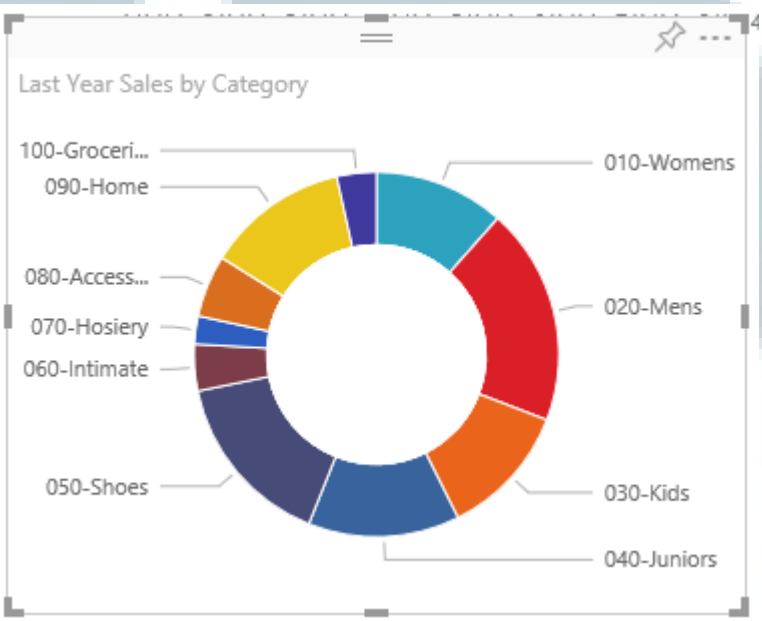


Gambar 2.11 *Power BI Slicer*

Sumber : (Microsoft, 2018)

6) *Donut Chart*

Donut chart sama seperti Grafik Pie merepresentasikan data dalam bentuk pie atau lingkaran yang menampilkan jumlah dan presentase data sesuai ukuran dari keseluruhan data hanya saja di tengah lingkaran terdapat lingkaran kosong seperti donut.



Gambar 2.12 *Donut Chart*

Sumber : (Microsoft, 2018)

7) Clustered Column Chart

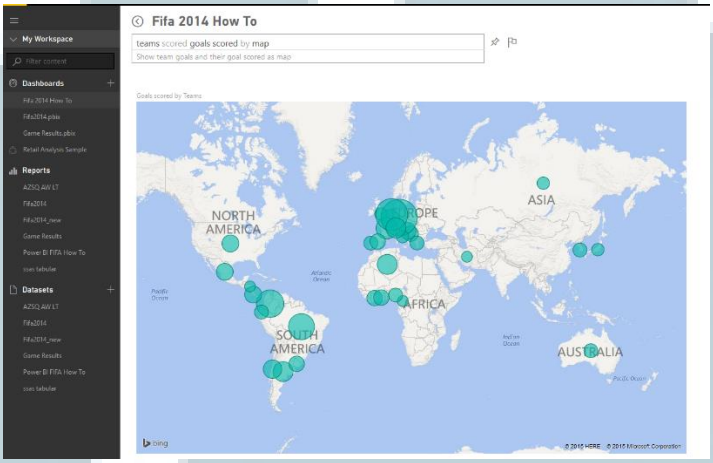
Grafik yang digunakan untuk membandingkan data berdasarkan kategorinya dengan menggunakan bar vertikal.

Gambar 2.13 *Clustered Column Chart*

Sumber : (Microsoft, 2018)

8) Map Chart

Map chart merepresentasikan data angka/jumlah dalam bentuk map bisa berupa warna, batang, maupun lingkaran sesuai dengan jumlah data yang direpresantasikan sesuai lokasi data(Location, Longtitude, Latitude).

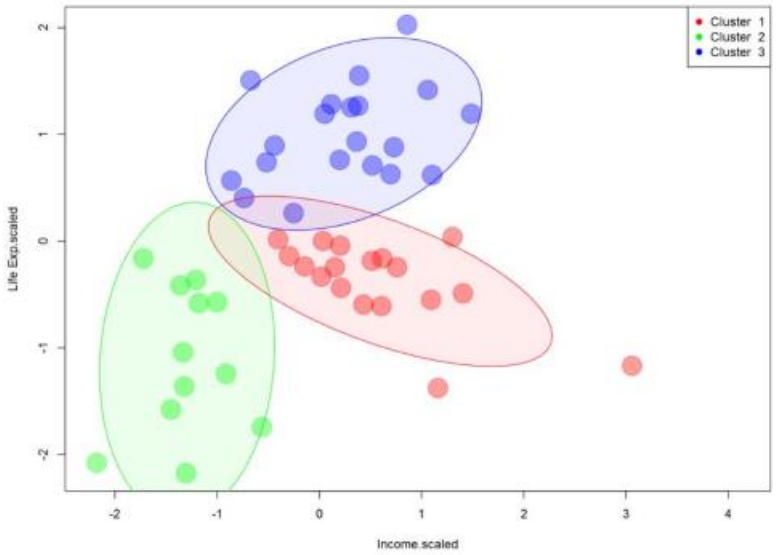


Gambar 2.14 *Map Chart*

Sumber : (Microsoft, 2018)

9) *K-mean Cluster*

Grafik yang menggambarkan pengelompokan data *cluster* biasanya berbentuk titik dan lingkaran dengan suatu algoritma/metode, penelitian ini menggunakan algoritma *k-mean*s untuk menampilkan data *clustering*.



Gambar 2.15 *Cluster Chart*

Sumber : (Microsoft, 2018)