

BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Berdasarkan judul yang ada terdapat beberapa Referensi yang di mana membahas hal yang sama. Referensi tersebut akan di jadikan sebagai acuan dan jika terdapat masalah yang akan dijadikan bahan pertimbangan. Adapun referensi-referensi tersebut adalah:

Algoritma K-Means bisa digunakan untuk mengelompokkan mahasiswa berdasarkan IPK dan beberapa atribut mata kuliah. Algoritma K-Means merupakan metode mengelompokkan data berdasarkan pusat cluster(centroid awal) terdekat dari data sumber. (Asroni & Adrian, 2015)

Dalam melakukan prediksi kelulusan tepat waktu mahasiswa terdapat berbagai macam permasalahan, diantaranya yaitu metode prediksi masih menggunakan prinsip kekeluargaan sehingga dirasakan kurangnya tingkat profesionalisme dalam melakukan prediksi kelulusan tepat waktu. (Saefulloh, 2014)

Dalam pemodelan K-Means, data yang memiliki Karakteristik sama akan dikelompokkan dalam kelompok yang sama dan data yang memiliki karakteristik berbeda akan masuk kelompok yang lainnya .Penerapan model K-Means dapat digunakan untuk mengelompokkan data seperti predikat kelulusan mahasiswa yang di dapat berdasarkan jumlah beban studi yang telah ditempuh, IPK yang diperoleh mahasiswa, dan kelulusan skripsi. (Rusdiana, 2016)

Metode yang digunakan adalah K-Means Clustering. Dengan menggunakan metode ini data-data yang telah didapatkan dapat dikelompokkan ke dalam beberapa cluster, berdasarkan kemiripan dari data-data tersebut. (Mulyati, 2015). Algoritma K-Means merupakan algoritma dari teknik clustering yang berbasis partisi. (Rohmawati W, 2015)

Clustering merupakan proses pembentukan suatu kelompok dari sebuah data yang tidak diketahui kelompok-kelompoknya berdasarkan kesamaan. Sehingga dalam suatu

cluster terdapat kelompok yang berisi data dan karakteristik yang sama. (Wicaksono, 2016)

Algoritma K-Means dapat diimplementasikan untuk membantu untuk mengelompokkan data siswa berdasarkan atribut yang dibutuhkan. (Eka, 2015). Dalam penelitian ini bertujuan untuk membagi siswa menjadi 3 *cluster* atau kelompok yang menggunakan *K-Means* memiliki kategorinya, Tinggi, Sedang, dan Rendah. (Naik, 2017)

Teknik mengelompokkan atau biasa disebut cluster merupakan hal yang sangat penting dari proses analisis suatu data. (Raval, 2016). Penelitian dengan menggunakan cluster ini berfungsi untuk mengurangi rasio terjadinya drop out terhadap siswa dan meningkatkan kinerja siswa menjadi lebih baik. (Singh, 2018)

2.2 Landasan Teori

2.2.1 Drop Out

Drop out adalah proses mahasiswa dikeluarkan dari suatu universitas belum pada waktunya atau biasa disebut sebelum lulus dari universitas. Mahasiswa yang *drop out* biasanya memiliki banyak faktor di antara lain adalah mahasiswa yang memiliki IP semester tidak sesuai dengan peraturan universitas, keduanya yaitu faktor jumlah Sks yang sedikit, mahasiswa yang termasuk dalam dua golongan ini hal yang pertama yang dilakukan oleh pihak universitas adalah melakukan tindakan memperingatkan tetapi jika mahasiswa tersebut masih memiliki faktor tersebut pihak kampus akan memberi teguran kembali tetapi jika dia semester 3 masih termasuk golongan tersebut maka pihak kampus memberi tindakan yang tegas dengan melakukan *Drop Out*.

Setiap universitas memiliki peraturan terhadap semua mahasiswanya untuk melakukan proses *Drop Out*. Mahasiswa program sarjana bisa dihentikan disuatu waktu jika:

1. Mahasiswa yang masa studi tahun pertama, kedua, ketiga memiliki IP semester rendah $IP < 2,00$ (satu koma nol), akan diberikan tindakan tegas terhadap mahasiswanya.

2. Mahasiswa yang masa studi tahun pertama, kedua, ketiga memiliki jumlah Sks rendah < 18.

2.2.2 Data Mining

Data Mining merupakan ekstraksi informasi yang tersembunyi dari database yang besar. Data Mining adalah teknologi baru yang kuat dengan potensi besar untuk membantu perusahaan agar fokus pada informasi yang paling penting dalam gudang data mereka (Ramamohan et al, 2012).

Data Mining mendukung proses dalam bisnis cerdas(*business intelligence*) karena data mining merupakan salah satu bidang yang terkait juga dengan gudang data(*data warehouse*) dalam bisnis cerdas. Selain bisnis cerdas, data mining juga memiliki kesamaan dengan bidang .ilmu kecerdasan buatan(*Artificial Intelegence*), pengenalan pola(*pattern recognition*), dan pembelajaran mesin(*machine learning*) seperti algoritma pencarian, teknik pemodelan, dan teori pembelajaran (Prasetyo, 2012).

Data Mining memiliki kemampuan untuk mencari suatu informasi bisnis dari suatu data yang jumlahnya sangat banyak. Teknologi Data Mining biasanya digunakan untuk:

1. Melakukan prediksi dalam bisnis, dimana Data Mining memudahkan untuk melakukan proses pencarian informasi dengan cara melakukan prediksi terhadap data-data dengan jumlah yang banyak.
2. Sebelum melakukan proses Data Mining dalam suatu data terdapat pola-pola yang tidak diketahui, Data Mining ini akan sangat berfungsi untuk mengidentifikasi pola-pola yang tidak jelas.
3. Data Mining berfungsi juga untuk membuat sebuah keputusan yang berguna dalam proses memprediksi suatu data.

Data Mining memiliki proses antara lain: deskripsi, prediksi, estimasi, klasifikasi, clustering dan asosiasi. Secara garis besar proses Data Mining dapat dijelaskan sebagai berikut :

1. Deskripsi

Deskripsi dalam Data Mining bertujuan untuk mengidentifikasi pola yang muncul secara berulang dan mengubah pola sesuai dengan kriteria yang dibutuhkan dalam suatu kasus pada data. Data Mining juga melakukan validasi dan menjelaskan proses yang dilakukan.

2. Prediksi

Proses prediksi memiliki kesamaan seperti proses klasifikasi. Contoh kasus dari prediksi adalah menentukan lulus mahasiswa secara tepat waktu yang bisa dihitung berdasarkan dengan IP per semester , contoh lainnya yaitu memonitoring mahasiswa rawan drop out yang bisa dihitung berdasarkan IP per semester dan jumlah Sks selama berkuliah di suatu universitas.

3. Estimasi

Proses estimasi hampir sama dengan proses prediksi. Proses estimasi dibangun menggunakan record lengkap yang menyediakan nilai dari suatu variabel target.

4. Klasifikasi

Proses klasifikasi merupakan proses menemukan suatu model yang biasanya dideskripsikan dan membagi data ke dalam kelas-kelas. Proses klasifikasi melibatkan proses pemeriksaan karakteristik dan memasukkan data ke dalam suatu kelas.

5. Clustering

Clustering merupakan proses pengelompokkan data tanpa berdasarkan kelas tertentu ke dalam kelas objek yang sama atau biasa disebut dengan cluster. Tujuan Clustering adalah mengelompokkan data yang sama dalam kelompok-kelompok tertentu.

6. Asosiasi

Proses asosiasi dalam Data Mining bertujuan untuk menemukan suatu atribut yang bisa muncul dalam suatu waktu tugas lain asosiasi berfungsi untuk mengungkap aturan yang berguna untuk mengukur hubungan antara dua atau lebih atribut.

2.2.3 Konsep Pengelompokan (*Clustering*) dan Metode K-Means

Menurut Baskoro *clustering* atau klusterisasi adalah salah satu alat bantu pada data mining yang bertujuan mengelompokkan objek-objek ke dalam *cluster-cluster*. *Cluster* adalah sekelompok atau kumpulan objek-objek data yang similar satu sama lain dalam cluster yang sama dan disimilar terhadap objek-objek yang berbeda cluster (Nango,2012).

Andayani (2007) menyatakan clustering memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spasial, dan analisis web. *Clustering* diterapkan dalam mesin pencari Internet. Web mesin pencari akan mencari ratusan dokumen yang cocok dengan kata kunci yang dimasukkan. Dokumen-dokumen tersebut dikelompokkan dalam *cluster-cluster* sesuai dengan kata-kata yang digunakan.

Algoritma *clustering* biasanya disebut algoritma mengelompokkan sejumlah data (N) menjadi sebuah kelompok-kelompok (cluster). Dalam suatu kelompok cluster harus memiliki suatu kemiripan sedangkan jika data tidak memiliki suatu kemiripan data tersebut tidak masuk dalam suatu kelompok (cluster).

Dalam *clustering* memiliki syarat untuk melakukan proses yang harus dipenuhi antara lain skalabilitas, kemampuan analisa beragam bentuk data, menemukan cluster dengan bentuk yang tidak terduga, kemampuan untuk dapat menangani noise, sensitifitas terhadap perubahan input, mampu melakukan clustering untuk data yang berdimensi tinggi, interpretasi dan kegunaan, ke enam

ini merupakan syarat untuk melakukan clustering adapun penjelasan dari syarat diatas adalah:

1. Skalabilitas.

Dalam metode clustering harus mampu menangani data dalam jumlah yang sangat banyak. Pada era saat ini data yang sangat banyak sudah digunakan diberbagai kepentingan.

2. Kemampuan analisa beragam bentuk data.

Pada proses clustering data yang berbentuk nominal, ordinal atau gabungannya harus bisa diimplementasikan juga seperti data yang bersifat numerik.

3. Kemampuan untuk melakukan noise.

Dalam melakukan penelitian kondisi data tidak selalu dalam keadaan yang baik. Ada kalanya terjadi kerusakan pada data yang menghambat penelitian. Dengan adanya algoritma clustering dituntut untuk menangani suatu data yang memiliki kerusakan.

4. Sensifitas terhadap perubahan input.

Dalam suatu penelitian biasa terajdinya penambahan data atau pengurangan yang bisa menyebabkan terjadinya perubahan cluster yang telah ada bahkan isa menyebabkan perubahan yang sangat mencolok apabila menggunakan clustering yang memiliki tingkat sensifitas yang rendah.

5. Mampu melakukan clustering untuk data dimensi tinggi.

Dalam suatu data biasanya berisi banyak dimensi ataupun atribut. Clustering mampu menangani data yang memiliki dimensi yang jumlahnya tidak sedikit.

6. Interpretasi dan Kegunaan

Hasil dari clustering harus berguna untuk suatu organisasi dan data itu dapat diinterpretasikan.

K-means merupakan metode analisis kelompok yang mengarah ke pembagian N objek pengamatan ke dalam kelompok (cluster) dan setiap objek pengamatan dimiliki oleh suatu kelompok dengan rata-rata (mean) terdekat (prasetyo, 2012)

Metode K-Means merupakan metode clustering yang mempartisi data yang dengan jumlah yang banyak ke dalam bentuk satu atau lebih cluster. Metode juga ini berfungsi untuk membagi kelompok ke dalam suatu cluster yang sama dan memiliki karakteristik yang sama sedangkan kelompok yang tidak memiliki karakteristik berbeda akan dicluster berbeda. Secara garis besar algoritma K-Means clustering sebagai berikut:

1. Tentukan jumlah cluster.

Pada tahap menentukan suatu cluster dilakukan dengan beberapa pertimbangan yang akan dilewati seperti pertimbangan teoritis dan pertimbangan konseptual yang dimana setelah melewati kedua pertimbangan tersebut mungkin lebih mudah menentukan suatu cluster.

2. Tempatkan data ke dalam cluster

Menentukan centroid awal pada metode K-Means dilakukan secara acak, kemudian untuk menghitung centroid cluster selanjutnya, digunakan rumus sebagai berikut seperti pada gambar 2.2 :

$$v = \frac{\sum_{i=1}^n x_i}{n} ; i = 1,2,3,\dots,n$$

Gambar 2.1 Rumus Menghitung Centroid

keterangan:

v : centroid pada cluster

x_i : objek ke- i

n : Banyaknya objek

3. Hitung jarak setiap hasil dari setiap cluster.

Untuk menghitung jarak antara objek dengan centroid menggunakan *Euclidian Distance*, adapun rumus yang digunakan sbisa dilihat pada gambar 2.3 :

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad ; i = 1, 2, 3, \dots, n$$

Gambar 2.2 Rumus Menghitung Jarak

Keterangan:

x_i : objek x ke- i .

y_i : daya y ke- i

n : banyaknya objek

4. Tempatkan data dengan centroid yang terdekat

Untuk menempatkan data atau objek kedalam setiap cluster secara umum dapat dilakukan dengan dua cara yaitu *hard k-means* dimana objek biasa disebut dengan anggota cluster dengan mengukur sifatnya cara lainnya menggunakan cara *fuzzy C-means*

Pada saat melakukan penerapan metode K-Means dalam suatu penelitian, data yang diolah biasanya data yang bisa diperhitungkan seperti data yang berbentuk numerik (angka). Sedangkan data yang tidak memiliki sifat numerik bisa juga diterapkan dalam metode K-Means tetapi terlebih dahulu untuk melakukan pengkodean terhadap data yang bukan numerik setelah melakukan pengkodean data pun mudah untuk dilakukan perhitungan jarak/kesamaan karakteristik yang dimiliki dari setiap objek. Metode K-Means sangat efektif dalam penelitian yang

bersifat prediksi algoritma K-Means juga tidak terpengaruh dengan urutan data yang digunakan

2.2.4 Weka

Weka merupakan aplikasi pendukung untuk melakukan data mining yang berbasis java. Weka pertama kali dikembangkan di Universitas Waikato yang terletak di Selandia Baru sebelum menjadi bagian dari Pentaho. Weka terdiri dari algoritma machine learning yang berfungsi untuk melakukan generalisasi atau formulasi dari sekumpulan data. Adapun tampilan awal Weka bisa dilihat pada gambar 2.4.



Gambar 2.3 Tampilan awal Weka

Gambar diatas merupakan tampilan awal dari Weka terdiri dari 5 tombol yang berfungsi untuk menjalankan aplikasi, adapun penjelasan dari 5 tombol antara lain:

1. Explorer

Explorer berfungsi untuk mencari data yang akan digunakan dan data tersebut akan diolah menjadi yang diinginkan dengan aplikasi Weka ini.

2. Experimenter

Experimenter digunakan untuk melakukan percobaan pada data sumber dengan pengujian statisic.

3. KnowledgeFlow

KnowledgeFlow untuk pengetahuan pendukung.

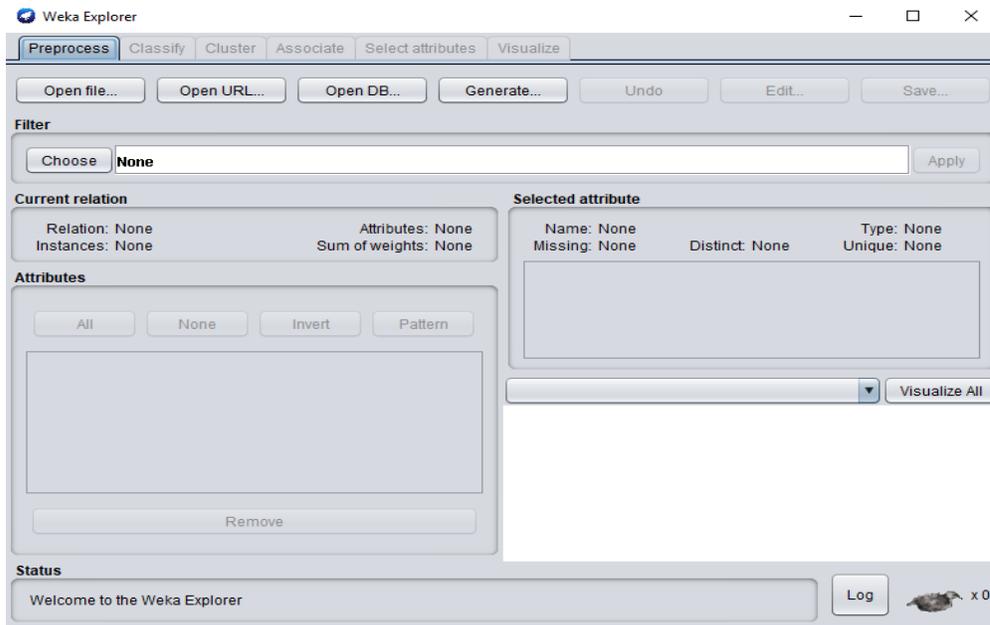
4. Workbench

Workbench berfungsi untuk melakukan visualization tools dan algortima untuk melakukan data analisis dan permodelan predikitif.

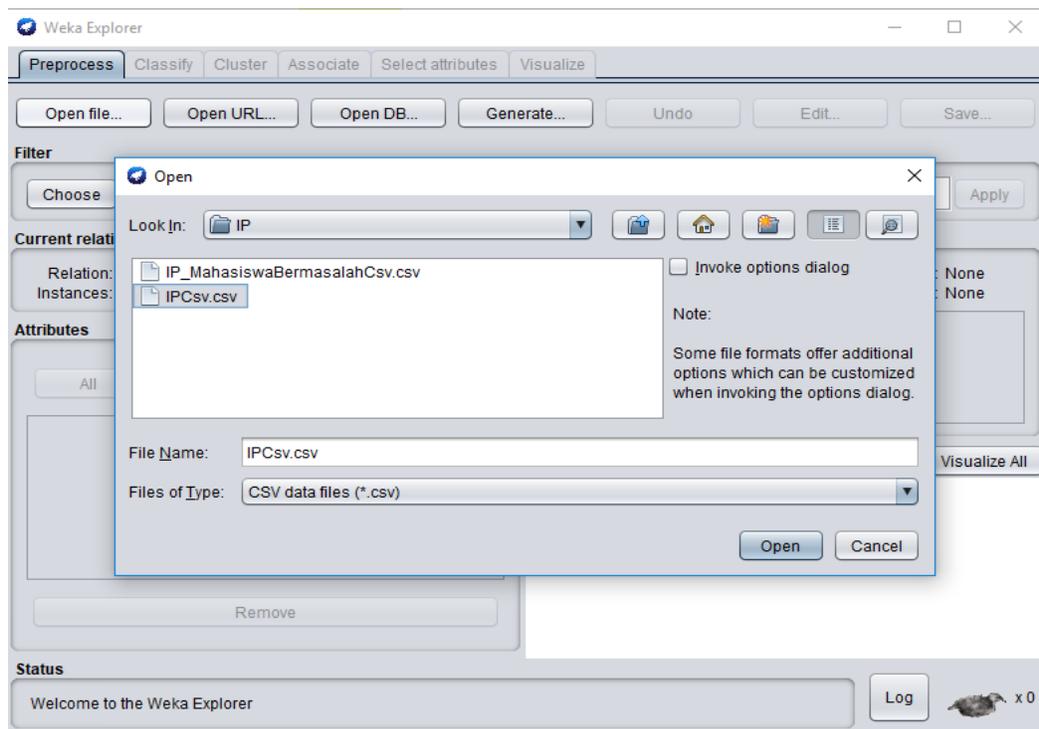
5. Simple CLI

Simple CLI berfungsi untuk mengeksekusi perintah pada weka untuk sistem operasi yang tidak tersedia secara langsung.

Sebelum melakukan cluster pada data akan muncul tampilan seperti pada gambar 2.5. yang pada bagian atas window terjadi beberapa jumlah bar seperti *preprocess, classify, cluster, associate, select attributes, visualize* namun yang hanya bisa digunakan adalah preopcess dikarenakan sebelum melakukan algoritma terlebih dahulu melakukan set file yang akan dieksekusi seperti pada gambar 2.6.

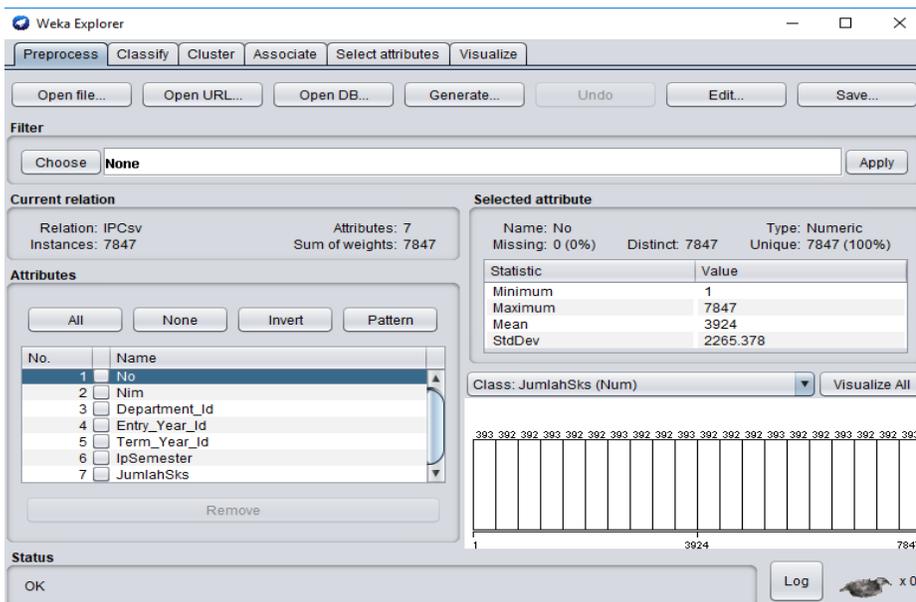


Gambar 2.4 Tampilan kedua Weka

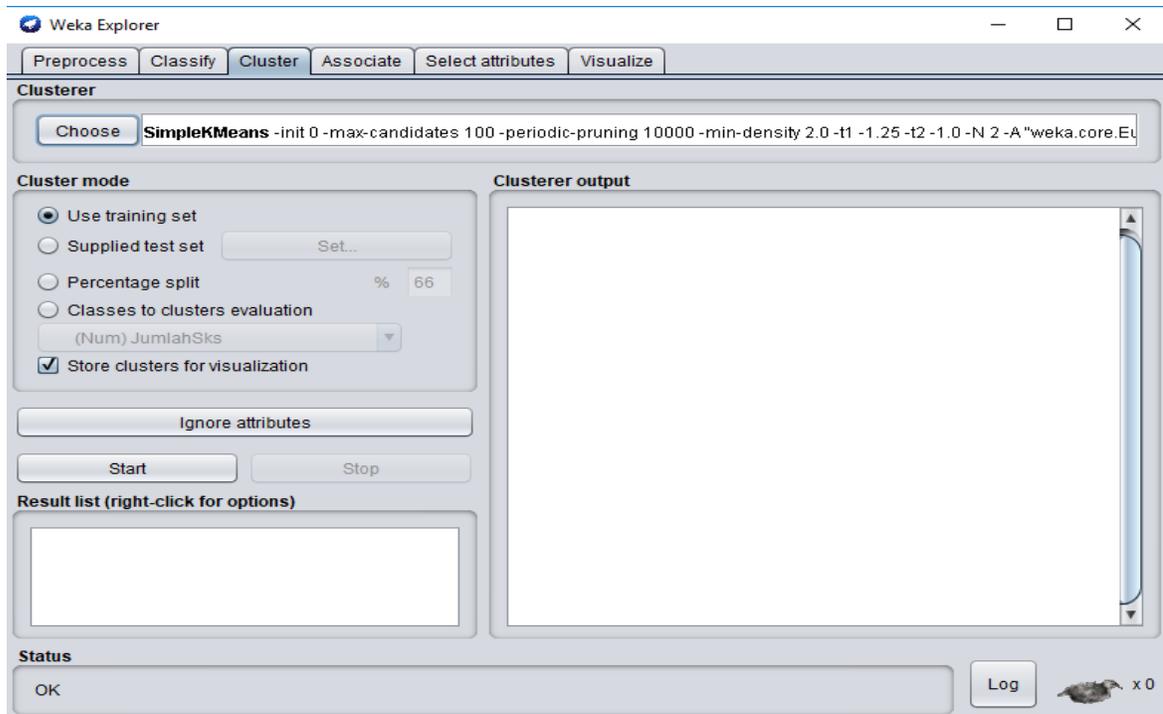


Gambar 2.5 Tampilan memasukkan data pada Weka

Pada gambar 2.6 diatas merupakan data yang akan diolah berformat .csv (Command Separated Values). Setelah membuka data selanjutnya akan muncul tampilan seperti pada gambar 2.7, ada perbedaan ketika data belum dipilih dan data sudah dipilih saat data .cvs telah dipilih bar yang berada pada window atas akan seketika aktif seperti *classify*, *cluster*, *associate*, *select attributes*, dan *visualize*. Setelah itu pilih bar cluster untuk melakukan metode K-Means seperti pada gambar 2.8 tampilan ini merupakan tampilan dari bar cluster untuk melakukan prediksi terhadap data yang berjumlah banyak. Contoh kasus yang menggunakan weka untuk *cluster* dengan menggunakan K-Means yaitu menghitung lamanya mahasiswa melakukan studi di salah satu universitas, memprediksi mahasiswa yang berprestasi berdasarkan nilai selama melakukan studi dan contoh terakhir adalah memprediksi mahasiswa yang bisa didrop out di salah satu universitas.



Gambar 2.6 Tampilan setelah data masuk



Gambar 2.7 Tampilan cluster pada Weka

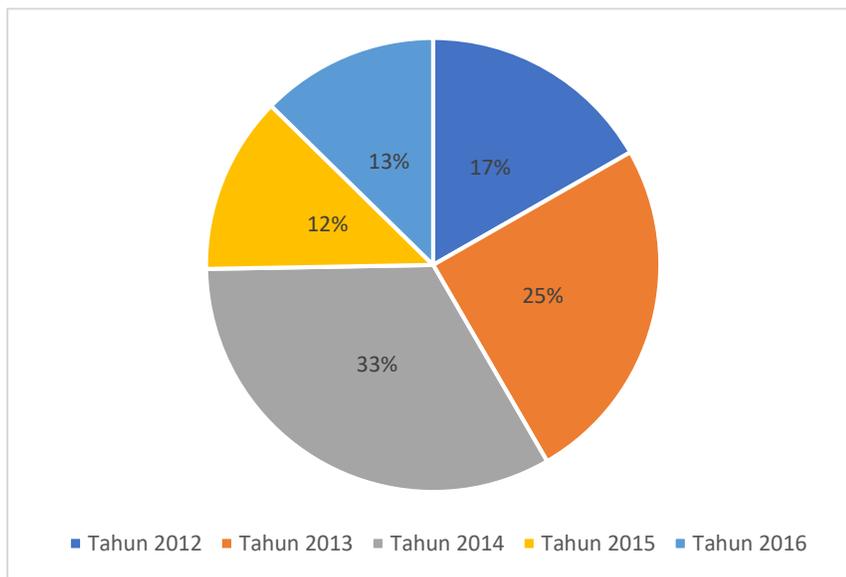
2.2.5 Diagram

Diagram merupakan suatu gambar yang bertujuan untuk memudahkan meneangkan data yang akan disajikan. Diagram biasa dikatakan sebagai lambang-lambang tertentu untuk menjelaskan sarana, prosedur terhadap kegiatan-kegiatan yang biasa dilakukan dalam suatu sistem. Diagram memiliki beberapa jenis seperti Diagram Lingkaran / Pie Chart, Diagram Batang, Diagram Garis / Line Chart, Diagram Kotak Garis, Diagram Batang Daun adapun penjelasan jenis-jenis diagram lain adalah :

1. Diagram Lingkaran / Pie Chart

Diagram Lingkaran atau biasa disebut Pie Chart adalah jenis penyajian data yang menggunakan gambar lingkaran, bagian-bagian yang ada di daerah lingkaran akan menunjukkan persentase data. Contoh Diagram

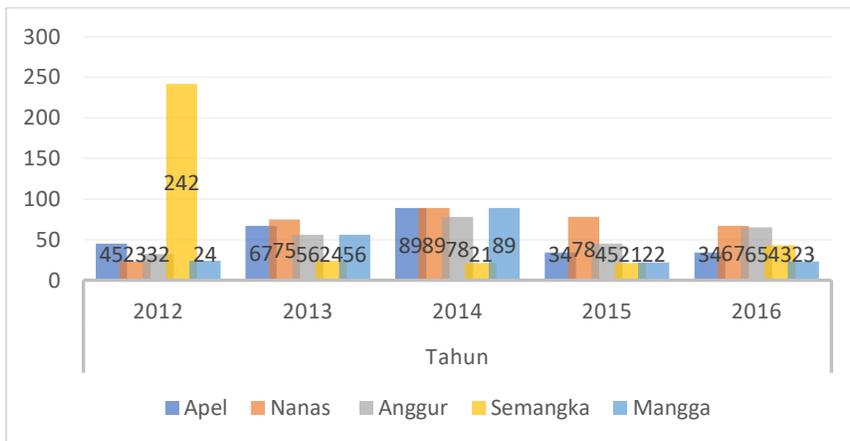
Lingkaran / Pie Chart bisa dilihat pada gambar 2.9.



Gambar 2.8 Diagram Lingkaran

2. Diagram Batang

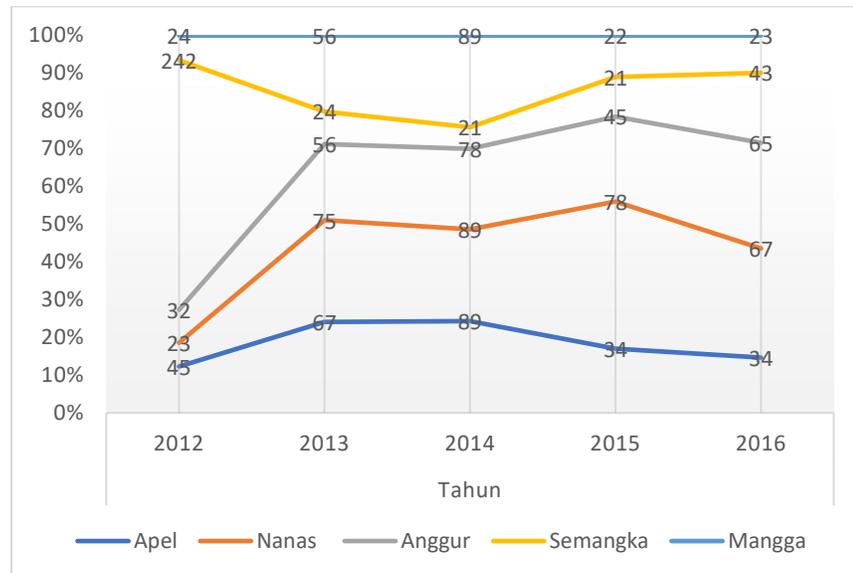
Diagram Batang merupakan diagram yang pada umumnya biasa digunakan untuk menggambarkan perkembangan nilai-nilai suatu objek yang diteliti. Contoh Diagram Batang bisa dilihat pada gambar 2.10.



Gambar 2.9 Diagram Batang

3. Diagram Garis / Line Chart

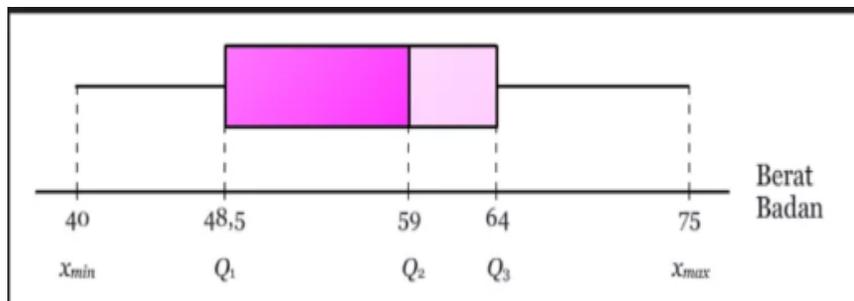
Diagram garis atau biasa disebut Line chart merupakan diagram yang pada umumnya digunakan untuk melakukan penyajian data statistik berdasarkan pengamatan waktu ke waktu secara berurutan. Contoh Diagram Garis / Line Chart bisa dilihat pada gambar 2.11.



Gambar 2.10 Diagram Garis

4. Diagram Kotak Garis

Diagram kotak garis memiliki statistik 5 serangkai, yang hal tersebut terdiri dari data terkecil, data terbesar, Q1, Q2, Q3. Contoh Diagram Kotak Garis bisa dilihat pada gambar 2.12.



Gambar 2.11 Diagram Kotak Garis

5. Diagram Batang Daun

Diagram batang daun digunakan untuk melakukan contoh penyebaran suatu data. Diagram daun ini akan diurutkan berdasarkan data yang berukuran besar hingga data yang berukuran besar. Contoh Diagram Daun bisa dilihat pada gambar 2.13.

Batang	Daun
4	4 7
5	6 9
6	1 3 4 5 6 8
7	0 1 1 1 1 2 3 3 3 4 5 5 6 6 6 9
8	3 4 4 5 5 9
9	1 4 5 7

Gambar 2.12 Diagram Daun

Jika dilihat pada gambar diatas maka digram diatas akan terbagi 2 menjadi bagian batang dan daun, pada bagian batang akan memuat angka puluhan sedangkan pada bagian daun akan memuat angka.