BAB II

LANDASAN TEORI

1.1 CRISP-DM

CRISP-DM merupakan singkatan dari *Cross Industry Standard Process for Data Mining*. CRISP-DM merupakan standarisasi data mining yang disusun oleh tiga penggagas data mining market. Yaitu Daimler Chrysler (Daimler-Benz), SPSS (ISL), NCR. Pada metodologi ini dilakukan pembagian siklus untuk proses *data mining* menjadi 6 tahap, dimana ketergantungan antara setiap tahap digambarkan dengan panah. Berikut merupakan gambaran dari metodologi CRISP-DM. (Larose, 2011)

Berikut adalah penjelasan dari setiap tahap (Larose, 2011):

1. *Business Understanding*

   Pada tahap ini berfokus pada pemahaman mengenai tujuan dari proyek dan kebutuhan secara persepektif bisnis, kemudian mengubah hal tersebut menjadi sebuah permasalahan *data mining* dan rencana awal untuk mencapai tujuan tersebut Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah *data mining*, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

2. *Data Understanding*

   Pada tahap ini dilakukan pengumpulan terhadap data, lalu kemudian mempelajari data tersebut dengan tujuan untuk mengenal data, melakukan identifikasi dan mengetahui kualitas dari data, serta mendeteksi subset yang menarik dari data yang dapat dijadikan hipotesa bagi informasi yang tersembunyi

3. *Data Preparation*

   Pada tahap ini dilakukan persiapan mengenai data yang akan digunakan pada tahap berikutnya. Kegiatan yang dilakukan antara lain: memilih kasus dan parameter yang akan dianalisis (*Select Data*), melakukan transformasi terhadap parameter tertentu (*Transformation),* dan melakukan pembersihan data agar data siap untuk tahap *modeling(Cleaning)*

4. *Modeling*

   Pada tahap ini dilakukan penentuan terhadap teknik *data mining,* alat bantu *data mining,* dan algoritma *data mining* yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma *data mining* tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik *data mining* tertentu, dapat kembali ke tahap persiapan data.

5. *Evaluation*

   Melakukan interpretasi terhadao ohasil dari *data mining* yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model

yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

6. *Deployment*

   Melakukan penyusunan laporan terhadap hasil yang didapat dari evalusasi pada tahap sebelumnya atau dari proses *data mining* yang dilakukan secara keseluruhan.

1.2 SPSS

SPSS adalah applikasi untuk melakukan analisis statistic. SPSS adalah singkatan dari *Statistical Package for the Social Sciences*. SPSS merupakan hal yang paling banyak dipikirkan oleh para mahasiswa dalam menyelesaikan tugas akhirnya. Karena memang aplikasi ini merupakan aplikasi yang satu ini merupakan applikasi yang popular dalam kalangan para mahasiswa yang sedang melakukan penelitian atau menempuh tugas akhir. Oleh karena itu, disini statistician memberikan kesempatan pada para pembaca untuk mempelajari SPSS (Aripin, 2008).

Kegunaan SPSS dalam penelitian adalah untuk olah dan analissis statistic. Banyak sekali analisis yang dapat dikerjakan dengan aplikasi tersebut. Antara lain Uji Deskriptive, regresi linear, regresi logistic, analisis factor, uji normalisasi, uji F dan uji T, Independent T Test, ANOVA, MANOVA, dll Bahkan dapat juga digunakan untuk pembuatan grafik, seperti Histogram, Normal PP, Detrend PP, Boxplot, dll. Untuk uji instrument atau uji validitas dan uji realibilitas, SPSS juga dapat melakukanya dengan fitur yang lengkap (Aripin, 2008).

1.3 Applikasi R

R (juga idkenal sebagai GNU S) adalah Bahasa pemrograman dan perangkat lunak untuk analisis statistika dan grafik. R dibuat oleh Ross Ihaka dan Robert Gentleman di Universitas Auckland, Selandia Baru, dan kini dikembangkan oleh R Development Core Team, di mana Chambers merupakana anggotanya. R dinamakan sebagian setelah nama dua pembuatnya (Robert Gentleman dan Ross Ihaka), dan sebagian sebagian dari permainan nama dari S (Conti, 2010).

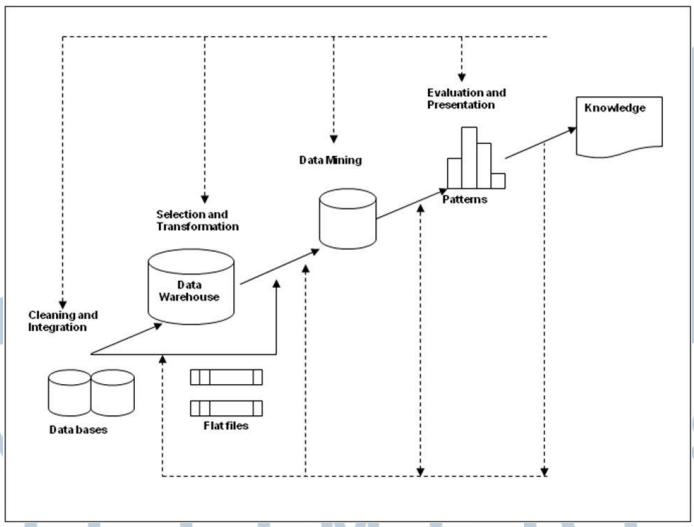
Bahasa R kini menjadi standar *de facto* di antara statistikawan untuk pengembangan perangkat lunak statistika, serta digunakan secara luas untuk pengembangan perangkat lunak statistika dan analisis data. R merupakan bagian dari proyek GNU. Kode sumbernya tersedia secara bebas di bawah Lisensi Publik Umum GNU, dan versi biner prekompilasinya tersedia untuk berbagai system operasi. R menggunakan antarmuka baris perintah, meski beberapa antarmuka pengguna grafik juga tersedia (Conti, 2010).

R menyediakan berbagai teknik statistika (pemodelan linier dan nonlinier, uji statistic klasik, analisis deret waktu, klasifikasi, klasterisasi, dan sebagainya) serta grafik. R, sebagaimana S, dirancang sebagai Bahasa computer sebenarnya, dan mengizinkan penggunanya untuk menambah fungsi tambahan dengan mendefinisikan fungsi baru. Kekuatan besar dari R yang lain adalah fasilitas tambahan dengan mendefinisikan fungsi baru. Kekuatan besar dari R yang lain adalah fasilitas grafiknya,

yang menghasilkan grafik dengan kualitas publikasi yang dapat membuat symbol matematika. R memiliki format dokumentasi seperti LaTeX, yang digunakan untuk menyediakan dokumentasi yang lengkap, baik secara daring (dalam berbagai format) maupun secara cetakan (Conti, 2010).

2.4 *Data Mining*

Nugroho (2014) mendefinisikan *data mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang menarik dari suatu data. Dalam *data mining* pengelompokan bias dilakukan, tujuanya untuk mengetahui pola universal data-data yang ada (Prasetyo,2014), Langkah-langkah melakukan *data mining.*



Gambar 2.1. Data Mining

Sumber: (Nugroho, 2014)

2.5 *Clustering*

Clustering adalah mengelompokkan item data ke dalam sejumlah kecil grup sedemeikian sehingga masing – masing grup mempunyai sesaut persamaan yang esensial (Andayani, 2007)

Ada beberapa pendekatan yang digunakan dalam mengembangkan metode *clustering*. Dua pendekatan utama adalah *clustering* dengan pendekatan partisi dan *clustering* dengan pendekatan hirarki. *Clustering* dengan pendekatan partisi atau sering disebut dengan *partition based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam *cluster-cluster* yang ada. *Clustering*dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering*mengelompokkan daya dengan membuat suatu hirarki berupa kurva yang menggambarkan penglompokkan *cluster* dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauha (Andayani, 2007).

Menurut Andayani (2007), Algoritma *clustering* dibagi ke dalam beberapa kelompok besar antara lain:

1. *Partitioning algorithms*: algoritma dalam kelompok ini membentuk bermacam partisi dan kemudian mengevaluasinya dengan berdasarkan beberapa kriteria.
2. *Hierarchy algorithms*: pembentukan dekomposisi hirarki dari sekumpulan data menggunakan beberapa kriteria.
3. *Density-based:* pembentukan *cluster*berdasarkan pada koneksi dan fungsi densitas.

4. *Grid-based:* pembentukan *cluster* berdasarkan pada struktur multiple-level granularity.
5. *Model-based:* sebuah model dianggap sebagai hipotesa untuk masing-masing *cluster* dan model yang baik dipilih diantara model hipotesa tersebut.

2.6 K-Means *Clustering*

Metode K-Means pertama kali diperkenalkan oleh MacQueen JB pada tahun 1976. Metode ini adalah salah satu metode *non hierarchi* yang umum digunakan. Metode ini termasuk dalam teknik penyekatan (*partition*) yang membagi atau memisahkan objek ke k daerah bagian yang terpisah. Pada K-Means, setiap objek harus masuk dalam kelompok tertentu, tetapi dalam satu tahapan proses tertentu, objek yang sudah masuk dalam satu kelompok, pada satu tahapan berikutnya objek akan berpindah ke kelompok lain (Santosa, 2009).

Hasil *cluster* dengan metode K-Means sangat bergantung pada nilali pusat kelompok awal yang diberikan. Pemberian nilai awal yang berbeda bias menghasilkan kelompok yang berbeda. Ada beberapa cara memberi nilai awal misalnya dengan mengambil sampel awal dari objek, lalu mencari nilai pusatnya, memberi nilai awal secara *random*, menentukan nilai awalnya atau menggunakan hasil dari kelompok hierarki dengan jumlah kelompok yang sesuai (Santosa, 2009).

K-Means adalah suatu metode penganalisaan data atau metode *Data Mining* yang melakukan proses pemodelan tanpa supervise (*Unsupervised)*

dang merupakan salah satu metode yang melakukan pengelompokan data dengan system partisi. Metode K-Means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya. (Agusta, 2011).

Menurut Nuningsih (2010), algortima K-Means memerlukan 3 komponen yaitu:

1. Jumlah *Cluster* K

   K-Means merupakan bagian dari metode non-hirarki sehingga dalam metode ini jumlah k harus ditentukan terlebih dahulu. Jumlah *cluster* k dapat ditentukan melalui pendekatan metode hirarki. Namun perlu diperhatikan bahwa tidak terdapat aturan khusus dalam menentukan jumlah *cluster* k, terkadang jumlah *cluster* yang diinginkan tergantung pada subyektif seseorang.

2. *Cluster* Awal

   Cluster awal yang dipilih berkaitan dengan penentuan pusat *cluster* awal (*centroid* awal). Dalam hal ini, terdapat beberapa pendapat dalam memlilih *cluster* awal untuk metode K-Means sebagai berikut:

a. Berdasarkan Hartigan (1975), pemilihan *cluster* awal dapat ditentukan berdasarkan interval dari jumlah setiap observasi.
b. Berdasarkan Rencher (2002), pemilihan *cluster* awal dapat ditentukan melalui pendekatan salah satu metode hirarki.
c. Berdasarkan Teknomo (2007), pemilihan *cluster* awal dapat dilakukan secara acak dari semua observasi.

Oleh karena adanya pemilihan *cluster* awal yang berbeda ini maka kemungkinan besar solusi *cluster* yang dihasilkan akan berbeda pula.

3. Ukuran Jarak

Metode K-Means dimulai dengan pembentukan prototype cluster di awal kemudian secara iterative prototype cluster ini diperbaiki hingga konvergen (tidak terjadi perubahan yang signifikan pada prototype cluster). Perubahan ini diukur dengan ukuran jarak *Euclidean*. Ukuran jarak ini digunakan untuk menempatkan observasi ke dalam *cluster* berdasarkan *centroid* terdekat.

Menurut Sarwono (2011), Algoritma K-Means adalah sebagai berikut:

a. Menentukan k sebagai jumlah *cluster* yang ingin dibentuk.
b. Membangkitkan nilai *random* untuk pusat cluster awal (*centroid)* sebanyak k

c. Menghitung jarak setiap data *input* terhadap masing-masing *centroid* menggunakan rumus jarak *Euclidiean (Euclidean Distance)* hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid.

Berikut adalah persamaan *Euclidean Distance:*

$$d(x_i, \mu_j) = \sqrt{(x_i - \mu_i)^2} \quad \ldots\ldots\ldots (1)$$

Rumus 2.1. Rumus Jarak Ecludiean Distance

Dimana:

Xi :data kriteria

μ j :*centroid* pada cluster ke-j

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil)
5. Memperbaharui nilai *centroid*. Nilai *Centroid* baru diperoleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

$$\mu j\,(t+1) = \frac{1}{Nsj} \sum_{j \in Sj} x_j \quad \ldots\ldots\ldots (2)$$

Dimana:

μ j(t + 1) : centroid baru pada iterasi ke (t+1)

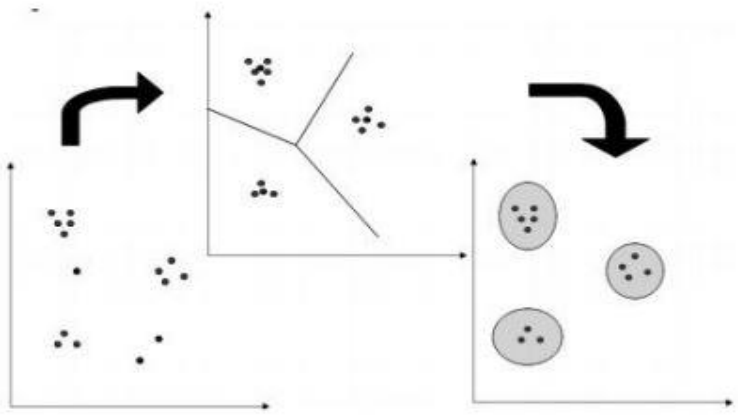
Nsj : banyak data pada cluster Sj

6. Melakukan perulangan dari langkah 2 hingga 5 hingga anggota tiap *cluster* tidak ada yang berubah.

7. Jika langkah 6 telah terpenuhi, maka nilai pusat cluster (JIj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data. Ilustrasi dari perubahan cluster/kelompok data ditungjukkan pada Gambar

Gambar 2.2. Diagram *Flowchart* Algoritma K-Means

Sumber : (Andayani,2007)



Gambar 2.3. Ilustrasi Algoritma K-Means

Sumber : (Andayani, 2007)

Berikut ini adalah ilustrasi penggunaan metode K-Means untuk menentukan *cluster* dari 4 buah objek dengan 2 atribut, seperti ditunjukkan dalam Tabel 2.1.

2.7 Penelitian Sebelumnya

Table 2.12. Penelitian Sebelumnya

1	Nama:	Narwati
	Judul:	Pengelompokan Mahasiswa Menggunakan Algoritma K-Means
	Tahun:	2016
	Metodologi:	Analisa Kebutuhan, Perancangan Desain, implementasi coding
	Nama Jurnal	*Jurnal Dinamika Informatika*, *2*(2). Vol 2 No 2 (2010)

	Hasil dan Simpulan:	Program applikasi menghasilkan pola dari prestasi mahasiswa yang klusternya tetap, turun dan naik. Pola Mahasiswa tersebut dapat terlihat dari asal program studi.
2	Nama:	Oyelade O. J, Oladipupo O.O, Obagbuwa, I.C
	Judul:	*Application of K-Means Clustering Algorihm for prediction of Student's Academic Performance*
	Tahun:	2010
	Metodologi:	Pengumpulan data langsung melalui badan kemahasiswaan
	Nama Jurnal	(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010
	Hasil dan Simpulan:	Algoritma pengelompokan ini berfungsi sebagai sesuatu yang baik, patokan untuk memantau perkembangan kinerja siswa..
3	Nama:	Zhongxiang Fan, Yan Sun, Hong Luo
	Judul:	*Clustering of College Students Based on Improved K-Means Algorithm*
	Tahun:	2014
	Metodologi:	Mengandalkan data *survey* dan observasi langsung.
	Nama Jurnal	978-1-5090-3438-3/16 $31.00 © 2016 IEEE DOI 10.1109/ICS.2016.138
	Hasil dan Simpulan:	K means adalah satu algoritma paling popular dalam algoritma pengelompokan. Namun, hasilnya tergantung pada pusat klaster awal. Sementara itu, outlier memiliki

		dampak besar pada hasil. Menghindari masalah ini, Algorima K-Means yang ditingkatkan berdasarkan kepadatan jaringan sangat direkomendasikan.

Penelitian sebelumnya perlu diperhatikan dan juga dipelajari agar menjadi landasan bagi penelitian berikutnya. Dengan mempelajari penelitian sebelumnya dapat memperkaya teori yang digunakan dalam mengkaji penelitian yang dilakukan. Dari lima jurnal yang ada diatas tidak ada penelitian yang memiliki judul yang sama, namun penelitian sebelumnya mengangkat penelitian penelitian yang mempunyai referensi yang sama sehingga dapat menjadi bahan kajian pada penelitian kali ini.

Dari penelitian-penelitian sebelumnya dapat disimpulkan para peneliti sebelumnya menggunakan metode pengumpulan data yaitu dengan observasi dokumen secara langsung. Dengan metode yang dikembangkan dapat terlihat dengan jelas hasil dari pada jurnal yang para peneliti buat bahwa implementasi K-Means Algoritma sangat memberikan dampak yang baik Mahasiswa atau Siswa. Beberapa jurnal diatas dapat disimpulkan bahwa dengan adanya K-Means Algoritma sangat berperan penting bagi kemajuan performa mahasiswa.

Berdasarkan penelitian yang sudah dilakukan sebelumnya maka metode yang akan digunakan pada implementasi kali ini adalah dengan observasi secara langsung sebagai teknik pengumpulan data yang kemudian akan di analisis. Sedangkan teknik baru pada penelitian kali ini yang akan digunakan adalah implementasi dengan syarat atau K-Means Algoritma.