BAB II

KAJIAN PUSTAKA

2.1 *Knowledge Discovery in Database* (KDD)

Konsep *data mining* dikenal sebagai *tools* penting dalam manajemen informasi karena jumlah data yang semakin lama semakan besar jumlahnya. *Knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam set data berukuran besar (Handoko, 2016). Istilah *data mining* dan *knowledge dicovery in database* (KDD) juga sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi yang tersembunyi dalam suatu basis dengan jumlah data yang besar. Kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut.

1. *Data Selection*

   Pemilihan himpunan data, menciptakan himpunan data target, atau fokuskan pada suatu sampel data, dimana penemuan *(discovery)* akan dilakukan. Pemilihan atau seleksi data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/ Cleaning*

   Pemrosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan *noise* dilakukan. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain menghapus duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

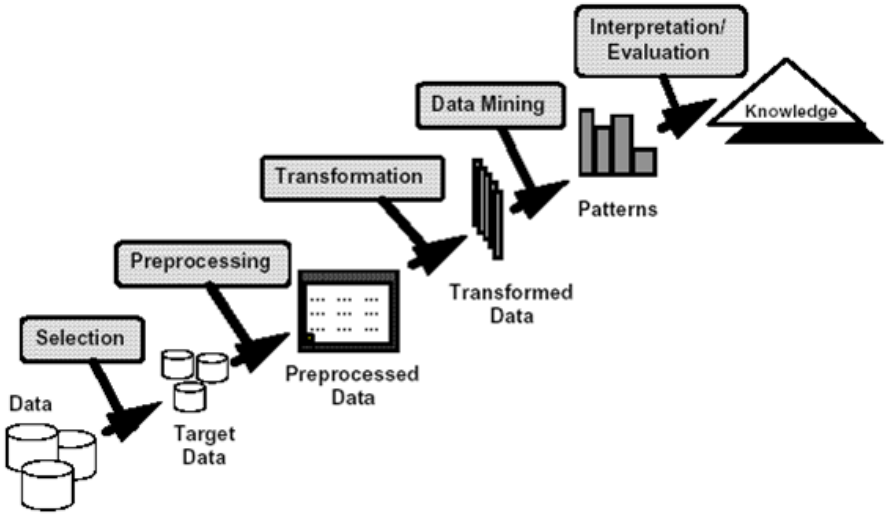
   Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

   Proses *Data mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/ Evaluation*

Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.



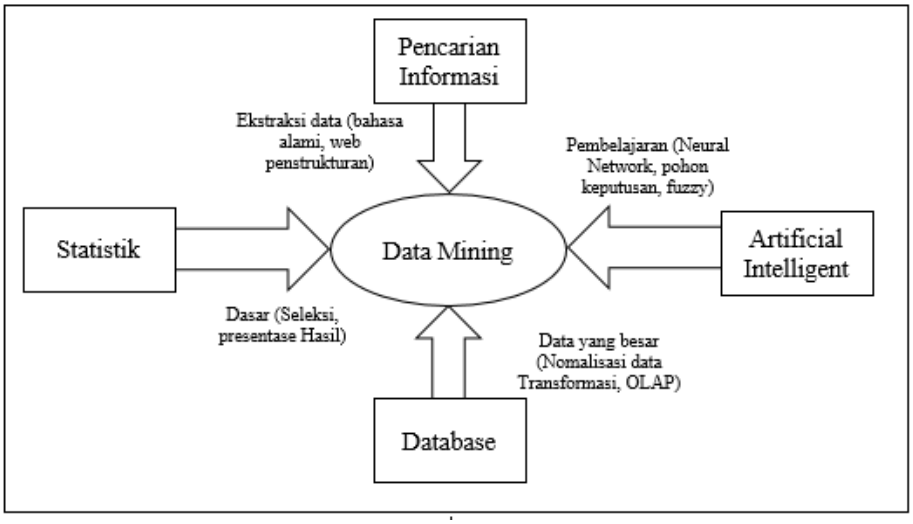
Gambar 2.1 Tahapan Knowledge Discovery in Database (KDD)

2.2 *Data Mining*

Pada tahun 1990 istilah *data mining* mulai dikenal, ketika pemanfaatan data menjadi suatu yang penting dalam berbagai bidang, Mulai dari bidang akademik, bisnis hingga medis. Munculnya *data mining* didasarkan pada jumlah data yang tersimpan dalam basis data yang semakin meningkat jumlahnya. *Data mining* juga disebut *knowledge discovery in database* atau *pattern recognition.* Istilah KDD atau disebut juga penemuan pengetahuan data karena tujuan utama data mining adalah memanfaatkan data dalam basis data dengan mengolahnya sehingga menghasilkan informasi baru yang berguna. Sedangkan istilah *pattern recognition* atau disebut

pengenalan pola mempunyai tujuan pengetahuan yang akan digali dari dalam bongkahan data yang sedang dihadapi (Fauziah Nur, Prof. M. Zarlis, 2015).

Data mining merupakan suatu rangkaian proses untuk dimana *data mining* ini menggali nilai tambah dari sekumpulan data yang berupa pengetahuan yang selama ini tidak diketahui secara manual (Handoko, 2016). *Data mining* bukanlah suatu bidang yang sama sekali baru. Namun terdapat kesulitan dalam mendefinisikan *data mining* yaitu kenyataan bahwa *data mining* ini mewarisikan banyak aspek dan teknik dari bidang-bidang ilmu. *Data mining* memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan *(artificial intelligent)*, *machine learning*, statistik, database, dan juga *information retrieval* yang dapat dilihat pada gambar 2.2 (Luthfi, 2009 : 6):



Gambar 2.2 Bidang Ilmu *Data Mining*

Data mining bisa juga bisa memberikan dampak negatif maupun positif bergantung pada pengunaanya *data mining*nya. Jika tidak memperhatikan etika dalam pengunaan data khususnya berhubungan dengan data-data pribadi pelanggan, maka *data mining* dapat berdampak negatif. Misalkan, klasterisasi

pelanggan berdasarkan suku bangsa, agama, ras, golongan, usia, maupun gender bisa berujung pada masalah diskriminasi dan bisa merugikan suatu kelompok tertentu. Tetapi, ketika *data mining* digunakan untuk masalah medis yang harus membedakan gender atau usia tertentu, maka pengunaan *data mining* dalam masalah medis ini justru berefek positif. Misalnya ada suatu penyakit yang peluangnya lebih besar diderita oleh kaum wanita atau kelompok usia tertentu, maka sudah seharusnya pihak medis melakukan penanganan secara berbeda (Dr. Suyanto, ST., 2017 : 7).

2.2.1 Manfaat *Data Mining*

Menurut Retno Tri Vulandari, S.SI., (2017 : 3) *data mining* terdapat dua sudut pandang dalam pemanfaatannya, yaitu sudut pandang komersial dan sudut pandang keilmuan. Dari sudut pandang komersial, pemanfaatan *data mining* dapat digunakan untuk menangani meledaknya volume data, dengan menggunakan teknik komputasi dapat digunakan untuk menghasilkan informasi-informasi yang dibutuhkan yang merupakan asset yang dapat meningkatkan daya saing suatu institusi. Sedangkan pada sudut pandang keilmuan *data mining* dapat digunakan untuk meng*capture,* menganalisis dan menyimpan data yang bersifat real time dan sangat besar.

Contoh dari sudut pandang komersial yaitu sebagai berikut:

1. Bagaimana mengetahui hilangnya pelanggan karena pesaing
2. Bagaimana mengetahui item produk atau konsumen yang memiliki kesamaan karakteristik

3. Bagaimana mengidentifikasi produk-produk yang terjual bersamaan dengan produk lain
4. Bagaimana memprediksi tingkat penjualan
5. Bagaimana menilai tingkat resiko dalam menentukan jumlah produksi suatu item
6. Bagaimana memprediksi prilaku bisnis dimasa yang akan datang.

Sedangkan contoh sudut pandang keilmuan *data mining* yakni:

1. *remote sensor* yang ditempatkan pada suatu satelit
2. telescope yang digunakan untuk memidai langit
3. simulasi saintifik yang membangkitkan data dalam ukuran terrabytes

2.2.2 Penerapan *Data Mining*

Dalam data mining terdapat dua penerapan dalam data mining yaitu sebagai berikut (Retno Tri Vulandari, S.SI., 2017 : 5) :

1. Analisis Pasar dan Manajemen

Sumber data yang digunakan seperti transaksi kartu kredit, kartu anggota club tertentu, kupon diskon, keluhan pembeli, ditambah dengan studi tentang gaya hidup public. Beberapa solusi yang dapat diselesaikan dengan data mining antara lain:

1) Menembak target pasar

Data mining dapat melakukan pengelompokan *(clustering)* dari model-model pembeli dan melakukan klasifikasi terhadap setiap pembeli sesuai dengan karakteristik yang diinginkan seperti tingkat penghasilan yang sama, kebiasaan pembeli dan karakteristik lainnya.

2) Melihat pola beli pemakai dari waktu ke waktu

Data mining dapat digunakan untuk melihat pola beli dari waktu ke waktu. Sebagai contoh, ketika seorang menikah bisa saja kemudian memutuskan untuk pindah dari *single account* ke *joint account.*

3) *Cross Market Analysis*

Kita dapat memanfaatkan *data mining* untuk melihat hubungan antara penjualan satu produk dengan produk lainnya.

4) Profil *Costumer*

Data mining dapat melihat profil costumer sehingga dapat mengetahui kelompok costumer tertentu suka membeli produk apa saja.

5) Identifikasi Kebutuhan costumer

Dapat mengidentifikasi produk apa saja yang terbaik untuk tiap kelompok costumer dan faktor apa saja yang dapat menarik kostumer.

6) Menilai loyalitas costumer

7) Informasi summary

Dapat digunakan untuk membuat laporan summary yang bersifat multi dimensi dan dilengkapi dengan informasi statistic lainnya.

2. Analisis Perusahaan dan Manajemen Resiko

1) Perencanaan keuangan dan evaluasi aset

Data mining dapat membantu melakukan analisis dan prediksi cash flow serta dapat melakukan *contingent claim analysis* untuk mengevaluasi aset. Selain itu dapat menggunakan untuk analisis trend.\

2) Perencanaan sumber daya

Dengan melihat ringkasan informasi serta pola pembelanjaan dan pemasukan dari masing-masing *resource.* Maka dapat memanfaatkan untuk *resource planning.*

3) Persaingan

Data mining dapat membantu untuk memonitoring pesaing-pesaing dengan melihat *marjet direction* mereka. Data mining juga dapat melakukan pengelompokan *costumer* dan dapat membertikan variasi harga untuk masing-masing grup.

4) Telekomunikasi

Data mining melihat jutaan transaksi yang masuk, dan melihat transaksi mana sajakah yang masih harus ditangani secara manual. Tujuannya adalah menambah layanan otomatis.

2.2.3 Kategori *Data Mining*

Kegunaan *data mining* adalah untuk menspesifikkan pola yang harus ditemukan dalam tugas data mining. Terdapat dua kategori utama dalam *data mining* yang dapat dijelaskan sebagai berikut (Retno, 2017:8), yaitu:

a. Prediktif

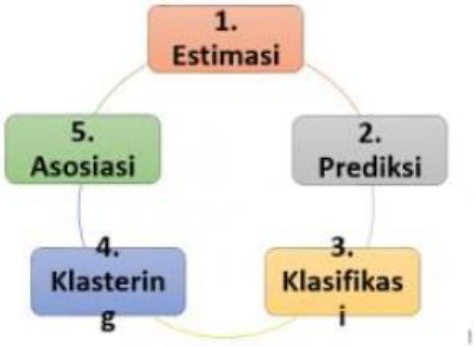
Tujuan dari prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory.*

b. Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, trend, cluster, teritori, dan anomali) yang meringkas hubungan pokok dalam data. Tugas *data mining* deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik post-processing untuk validasi dan penjelasan hasil.

2.3 Metode *Data Mining*

Data mining dibagi menjadi beberapa kelompok yang dibagikan berdasarkan tugas-tugasnya yang dapat dilakukan untuk menemukan, mengali dan menanmbang pengetahuan, yaitu sebagai berikut (Luthfi, 2009 :10):



Gambar 2.3 Metode *Data Mining*

a. Description

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari data untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menentukan keterangan atau fakta bahwa siapa yang tidak cukup professional

akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelesan untuk suatu pola atau kecenderungan.

b. Estimation

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik dari pada kearah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel predikasi. Sebagai contoh akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

c. Prediction

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam predikasi nilai dari hasik akan ada dimasa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

d. Classfication

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e. Cluestering

Clustering merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. *Cluster* adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (*homogeny*), yang mana kemiripan dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

f. Association

Tugas asosiasi dalam *Data Mining* adalah menemukan *attribut* yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.3.1 Teknik *Clustering*

Clustering merupakan bagian dari ilmu *data mining* yang bersifat tanpa arahan *(unsupervised). Clustering* merupakan pekerjaan yang memisahkan data atau vektor ke dalam sejumlah kelompok atau cluster menurut karakteristiknya masing-masing. Obyeknya merupakan untuk kasus pendistribusian (orang-orang,

objek, peristiwa dan lainnya) ke dalam kelompok, sedemikian hingga derajat tingkat keterhubungan antar anggota *cluster* yang sama adalah kuat dan lemah antar anggota dari *cluster* yang berbeda (Jannah & Yulianto, 2016). *Clustering* merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. *Cluster* adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogeny), yang mana kemiripan dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal (Luthfi, 2009 : 11).

Pada dasarnya *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik *(similarity)* antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan *(unsupervised),* maksudnya metode ini diterapkan tanpa adanya latihan *(training)* dan tanpa ada guru *(teacher)* serta tidak memerlukan target output. Metode *clustering* yang paling banyak digunakan ialah metode *K-Means clustering*. Kelemahan utama dari metode ini adalah hasil yang sensitif terhadap pemilihan pusat *cluster* awal dan perhitungan solusi lokal untuk mencapai kondisi optimal. Analisis *Cluster* merupakan teknik multivariat yang

mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis *Cluster* mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam *cluster* yang sama. *Cluster-cluster* yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi. Berbeda dengan teknik multivariat lainnya, analisis ini tidak mengestimasi set variabel secara empiris sebaliknya menggunakan set variabel yang ditentukan oleh peneliti itu sendiri (Handoko, 2016). *Clustering* banyak digunakan dalam berbagai bidang dengan beragam aplikasi vital yang sangat penting, diantaranya adalah riset pasar, dimana klasterisasi digunakan untuk segmentasi dan profilling pelanggan yang membantu dalam merancang strategi-strategi produk, harga, tempat, dan promosi. Klasterisasi juga dapat digunakan untuk mengimplementasikan *costumer relationship management* (CRM) yang efektif seperti sistem perekomendasian produk dalam sistem jual beli *online* yang biasanya menggunakan pendekatan *collaborative filtering,* dimana klasterisasi adalah bagian dasar dari *collaborative filtering*, *business intelligence*, sistem keamanan, mesin pencarian di internet *(search engine)*, dan sebagainya (Dr. Suyanto, ST., 2017 : 260)

2.3.2 Algoritma *K-Means*

Metode *K-means* merupakan algoritma klasterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya. Ide dasar algoritma *k-means* sangatlah sederhana, yaitu meminimalkan *Sum of Squared Error* (SSE) antara objek-objek data dengan

sejumlah *k centroid*. Algoritma *k-means* bekerja empat langkah. Pertama, dari himpunan data yang akan diklasterisasi, dipilih sejumlah k objek secara acak sebagai *centroid* awal. Kedua, setiap objek yang bukan *centroid* dimasukkan ke klaster terdekat berdasarkan ukuran jarak tertentu. Ketiga, setiap *centroid* diperbaruhi berdasarkan rata-rata dari objek yang ada di dalam setiap klaster. Keempat, langkah kedua dan ketiga tersebut diulang-ulang (diiterasi) sampai semua *centroid* stabil atau konvergen, dalam arti semua *centroid* yang dihasilkan dalam iterasi saat ini sama dengan semua *centroid* yang dihasilkan pada iterasi sebelumnya (Dr. Suyanto, ST., 2017 : 262).

Algoritma *K-Mean*s merupakan algoritma pengelompokan iteratif yang melakukan partisi set data ke dalam sejumlah K *cluster* yang sudah ditetapkan di awal. Algoritam *K-Means* sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. *K-means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efesien. Namun, *K-Means* memiliki kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode *K-Means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan (Handoko, 2016).

K-Means dapat diterapkan pada data yang direpresentasikan dalam r-dimensi ruang tempat. *K-means* mengelompokkan set data r-dimensi, X= x_i\{ |i=1,...,N}. Algoritma K-Means mengelompokkan semua titik data dalam X sehingga setiap titik x_i hanya jatuh dalam satu K partisi. Tujuan pengelompokan ini adalah untuk meminimalkan fungsi objek yang diset dalam proses pengelompokan, yang pada

umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antarkelompok. Parameter yang harus dimasukkan ketika menggunakan algoritma *K-Means* adalah nilai K. Nilai K yang digunkan pada umumnya didasarkan pada informasi yang diketahui sebelumnya mengenai sebenarnya berapa banyak *cluster* yang muncul dalam X, berapa banyak yang digunakan untuk penerapnnya, atau jenis *cluster* dicari dengan melakukan percobaan dengan beberapa nilai K. Set representatif *cluster* dinyatakan $C=\{ c_j | j=1,\ldots,K\}$. Sejumlah K representatif *cluster* tersebut sebagai *cluster* centroid (titik pusat *cluster*). Untuk set data dalam X dikelompokan berdasarkan konsep kedekatan atau kemiripan, namun kuantitas yang digunkan untuk mengukurnya adalah ketidakmiripan. Metrik yang umum digunakan untuk ketidakmiripan tersebut adalah Euclidean (Fauziah Nur, Prof. M. Zarlis, 2015). Beberapa teknik klastering yang paling sederhana dan umum adalah klastering *K-means*. Secara detail teknik ini menggunakan ukuran ketidak miripan untuk mengelompokkan obyek. Ketidak miripan dapat diterjemahkan dalam konsep jarak. Dua obyek dikatakan mirip jika jarak dua objek tersebut dekat. Semakin tinggi nilai jarak, semakin tinggi nilai ketidak miripannya. Metrik ketidak miripan tersebuat adalah *Euclidean. K-Means Clustering* merupakan metode yang termasuk ke dalam golongan algoritma *Partitioning Clustering*. Berikut adalah langkah-langkah dari metode *K-Means* sebagi berikut (Handoko, 2016):

1. Tentukan nilai k sebagai jumlah *cluster* yang ingin dibentuk.
2. Bangkitkan *k centroid* (titik pusat *cluster*) awal secara acak.

3. Hitung jarak setiap data ke masing-masing centroid menggunakan rumus korelasi antar dua objek *(Euclidea Distance).*
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya.
5. Tentukan posisi centroid baru (k C) dengan cara menghitung nilai rata-rata dari data yang ada pada centroid yang sama.

$$C_k=\left(\frac{1}{n_k}\right)\sum d_i$$

Rumus 2.1 Nilai Rata-rata pada *Centroid*

Dimana nk adalah jumlah dokumen dalam *cluster* k dan di adalah dokumen dalam *cluster* k.

6. Kembali ke langkah 3 jika posisi centroid baru dengan centroid lama, tidak sama.

2.4 Sofware Pendukung

RapidMiner sebelumnya dikenal sebagai YALE (*Yet Another Learning Environment),* mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer dari Unit Kecerdasan Buatan Universitas Teknik Dortmund. Mulai tahun 2006, perkembangannya didorong oleh *Rapid-I*, sebuah perusahaan yang didirikan oleh Ingo Mierswa dan Ralf Klinkenberg pada tahun yang sama. Pada tahun 2007, nama perangkat lunak itu berubah dari YALE menjadi *RapidMiner*. Pada tahun 2013, perusahaan melakukan *rebranding* dari *Rapid-I* menjadi *RapidMiner*. *RapidMiner* adalah platform perangkat lunak ilmu data yang

dikembangkan oleh perusahaan bernama sama dengan yang menyediakan lingkungan terintegrasi untuk persiapan data, pembelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis prediktif. Hal ini digunakan untuk bisnis dan komersial, juga untuk penelitian, pendidikan, pelatihan, *rapid prototyping*, dan pengembangan aplikasi serta mendukung semua langkah dalam proses pembelajaran mesin termasuk persiapan data, hasil visualisasi, validasi model, dan optimasi. *RapidMiner* dikembangkan pada model inti terbuka.

Gambar 2.4 *RapidMiner*

RapidMiner adalah salah satu *software* untuk pengolahan data mining. Pekerjaan yang dilakukan oleh *RapidMiner text mining* adalah berkisar dengan analisis teks, mengekstrak pola-pola dari data set yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan database. Tujuan dari analisis teks ini adalah untuk mendapatkan informasi bermutu tertinggi dari teks yang diolah. *RapidMiner* menyediakan prosedur *data mining* dan *machine learning*, di dalamnya termasuk: ETL *(extraction, transformation, loading)*, data *preprocessing, visualisasi, modelling* dan evaluasi. Proses data mining tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI. Penyajiannya dituliskan dalam bahasa pemrograman *Java*. *RapidMiner* merupakan *software* / perangkat lunak untuk pengolahan data. Dengan menggunakan prinsip dan algoritma data mining, *RapidMiner* dapat mengekstrak

pola-pola dari data set yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan database. *RapidMiner* memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan node-node pada operator kemudian kita hanya tinggal menghubungkannya ke node hasil untuk melihat hasilnya. Hasil yang diperlihatkan *RapidMiner* pun dapat ditampilkan secara visual dengan grafik.

2.5 Penelitian Terdahulu

Berikut merupakan penelitian terdahulu berupa beberapa jurnal terkait dengan penelitian, yakni:

1. **Handoko** (2016) yang berjudul **Penerapan *Data Mining* Dalam Meningkatkan Mutu Pembelajaran Pada Instansi Perguruan Tinggi Menggunakan Metode *K-Means Clustering*,** ISSN: 2476 – 8812, membahas tentang bagaimana Penelitian ini menerapkan Data Mining dengan menggunakan metode *clustering* untuk Meningkatkan mutu pembelajaran pada Instansi Perguruan Tinggi di Program Studi TKJ Akademi Komunitas Solok Selatan. Algoritma yang digunakan yaitu K Algoritma yang digunakan yaitu *K-Means Clustering* berupa proses pengelompokan sejumlah data atau objek ke dalam *cluster (group)* sehingga setiap dalam cluster tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam cluster yang lainnya. Pengujian dilakukan dengan aplikasi ilakukan dengan aplikasi *RapidMiner 5.3 sehingga* menghasilkan *cluster- RapidMiner*

5.3 sehingga menghasilkan *cluster-cluster* dalam meningkatkan mutu pembelajaran cluster dalam meningkatkan mutu pembelajaran cluster dalam meningkatkan mutu pembelajaran. Sampel yang digunakan diambil dari tabel data mahasiswa yang telah ditrasformasi. Di mana variabel yang pengujian pertama ditentukan sebanyak 4 variabel, yaitu: IP mahasiswa, jarak tempuh mahasiswa, jumlah kehadiran dan penghasilan orang tua. Di mana akan mempresentasikan data mahasiswa dengan mutu pembelajaran sangat baik, baik, cukup baik, dan kurang baik.

2. **Nur Jannah & Tony Yulianto** (2016) yang berjudul **Mengelompokkan Siswa Berprestasi Akademik Dengan Menggunakan Metode *K-Means* Kelas VII MTS Hidayatul Mabtadi'in Pancoran Kadur,** ISSN: 2459-9948 membahas tentang pengelompokan dari 30 siswa dengan kriteria pembobot pintar, sedang, dan tidak pintar menggunakan metode *clustering* dengan *algoritma k-means*. juga dapat digunakan untuk memantau perkembangan kemampuan setelah mengikuti belajar siswa. Dari hasil dan pembahasan dapat dapat diperoleh hasil dari validasinya data valid ada 33.3333% sedangkan data yang tidak valid ada 66.6667%.
3. **Fauziah Nur, Prof. M. Zarlis (2015)** yang berjudul **Penerapan Algoritma *K-means* Pada Siswa Baru Sekolah Menengah Kejuruan Untuk *Clustering* Jurusan,** ISSN:2540-7600 membahas tentang mengelompokkan data siswa baru sekolah menengah kejuruan tahun ajaran 2014/2015. Pengelompokan tersebut berdasarkan kriteria – kriteria data siswa. Pada penelitian ini, penulis menerapkan algoritma K-Means Clustering untuk

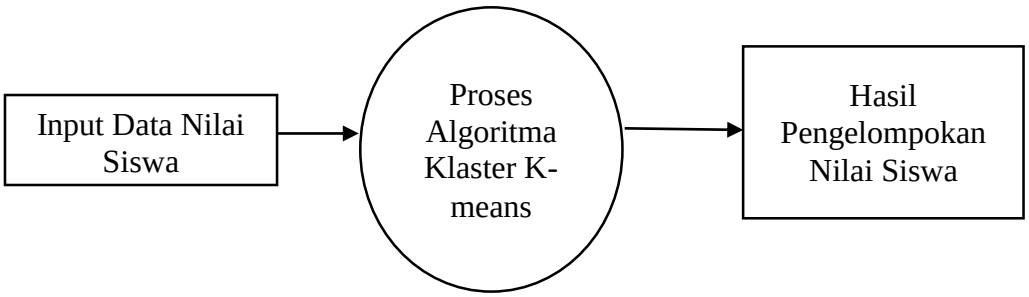
pengelompokan data siswa baru sekolah menengah kejuruan. Dalam hal ini, pada umumnya untuk memamasuki jurusan hanya disesuaikan dengan nilai siswa saja namun dalam penelitian ini pengelompokan disesuaikan kriteria – kriteria siswa seperti penghasilan orang tua, tanggungan anak orang tua dan nilai tes siswa. Penulis menggunakan beberapa kriteria tersebut agar pengelompokan yang dihasilkan menjadi lebih optimal. Tujuan dari pengelompokan ini adalah terbentuknya kelompok jurusan pada siswa yang menggunakan algoritma K-Means clustering. Hasil dari pengelompokan tersebut diperoleh tiga kelompok yaitu kelompok tidak lulus, kelompok rekayasa perangkat lunak dan kelompok teknik komputer jaringan. Terdapat pusat cluster dengan Cluster-1=1.4;2.2;2.2, Cluster-2= 2.28;1.64;4 dan Cluster-3=5;2;6. Pusat cluster tersebut didapat dari beberapa iterasi sehingga menghasilkan pusat cluster yang optimal.

4. **Sumathi, Kannan, & Nagarajan, (2016)** yang berjudul **Constrained K-means Clustering with Background Knowledge**, ISSN: 0975–8887 yang membahas tentang Data Analysis can be categorized into two forms. One is used for extracting models describing important classes; another is to predict future trends. Data classification can be used to generate models which are further used to predict the unknown classes. The accuracy of the models can be examined by checking the percentage of correctly classified instance. Lot of classification algorithms is available nowadays. One of the most commonly used algorithms is decision tree because of its simplicity of implementation and easier to understand when compared to other

classification algorithms. J48 is the one of the effective classification method. In this paper, J48 algorithm is applied for analyzing student dataset which includes academic year, department, academic grade and job position

5. **Purnamaningsih, Saptono, & Aziz, (2016)** yang berjudul **Pemanfaatan Merode K-Means Clustering dalam Penentuan Penjurusan Siswa SMA,** ISSN: 2301-7201 yang membahas tentang n setiap cluster diklasifikasikan berdasarkan kriteria mana yang lebih diprioritaskan. Cluster dengan nilai terbesar pada centroid akhir merupakan cluster yang diterima IPA/IPS, sedangkan cluster dengan nilai terkecil pada centroid akhir merupakan cluster yang ditolak IPA/IPS. Hasil penelitian pengujian terbaik pada praprocessing clustering K-Means IPA dengan hasil akurasi 0.905882, tingkat kesesuaian hasil prediksi dengan data sebenarnya (recall) 1, ketepatan hasil pengujian dalam memprediksi clustering (sensitivity) 0.876923, kesesuaian prediksi negatif terhadap aktual negatif (specificity) 0.714285. Sedangkan pengujian terbaik juga pada praprocessing clustering K-Means IPS didapatkan akurasi 0.905882, recall 0.714285, sensitivity 1, dan specificity 1. Hasil perbandingan clustering terbaik pada praprocessing clustering K- Means IPA dengan praprocessing clustering K-Means IPS menunjukkan bahwa tidak ada siswa yang diterima di dua jurusan IPA/IPS atau siswa ditolak di keduanya.

2.6 Kerangka Pemikiran



Gambar 2.5 Kerangka Pemikiran

Penelitian ini diawali dengan input data-data dari hasil tes siswa tersebut dan kemudian data-data hasil tes siswa tersebut diproses dengan *algoritma k-means clustering*, dan setelah data-data tersebut telah diproses akan memperoleh hasil pengelompokan siswa.

2.7 Hipotesis

Hipotesis dalam penelitian ini adalah dengan adanya pengelompokan siswa dengan hasil tes siswa dengan *data mining* menggunakan metodel *K-means Clustering* dapat membantu guru dalam pemantauan nilai siswa-siswa dalam pengambilan keputusan di Sekolah.