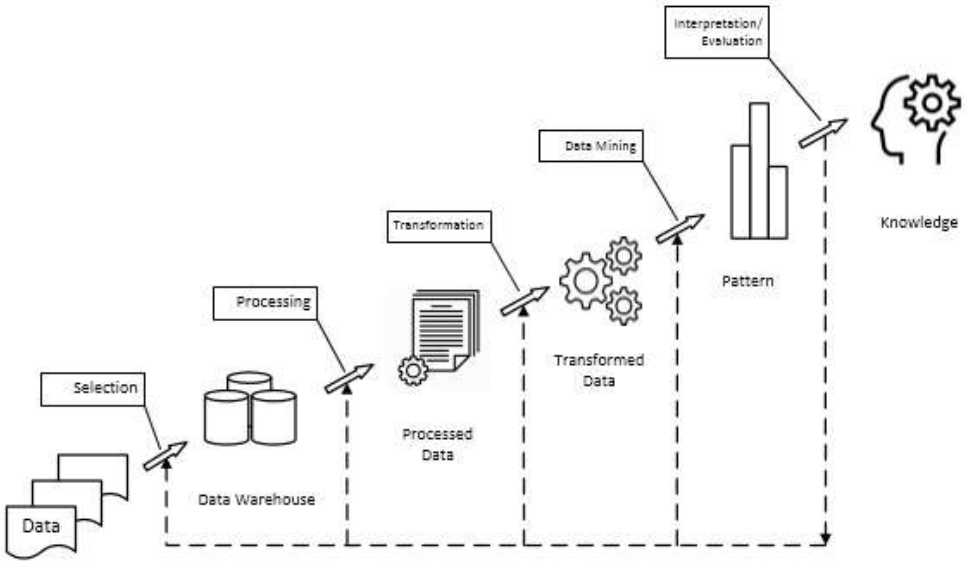
BAB II

KAJIAN PUSTAKA

2.1. *Knowledge Discovery in Database* (KDD)

Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan dengan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Secara garis besar proses KDD sebagai berikut (Nofriansyah, 2014)



Gambar 2. 1 Proses *Knowledg Discovery In Database*
Sumber: Gambar Peneliti (2020)

1. *Data Selection,* sebelum melakukan tahap penggalian data memerlukan tahap pemrosesan pemilihan data terlebih dahulu dari kumpulan-kumpulan data operasional. Kemudian data dari hasil seleksi/pemilihan akan digunakan untuk proses *data mining*, data tersebut disimpan ke sebuah berkas tersendiri yang berbeda dari berkas basis data operasional agar mempermudah pada penggunaan ke tahapan berikutnya.
2. *Pre-pocessing*/*cleansing*, proses di tahap ini berupa memeriksa data yang bersifat non-konsisten, membuang data ganda dan atribut-atribut data yang non-relevan serta memperbaiki kesalahan yang terdapat di data tersebut. Kemudian data yang telah diperoleh melalui database atau berkas di suatu perusahaan, terdapat keterangan atau isian yang kurang lengkap seperti data tersebut tidak valid, data tersebut hilang, atau data tersebut mengalami kesalahan pengetikan, keberadaan data-data tersebut dapat mengakibatkan kurangnya akurasi atau kulitas dari hasil *data mining* berikutnya. *Cleansing* data akan berpengaruh pada performasi melalui sistem *data mining* sebab data akan mengalami pengurangan kompleksitas dan jumlahnya berdasarkan data yang ditangani.
3. *Transformation,* dalam metode *data mining* sebelum bisa diaplikasikan ada beberapa metode yang membutuhkan format data yang khusus. Seperti beberapa metode *clustering* dan metode analisis, metode tersebut hanya bisa menerima data-data input kategorikal. Oleh karena itu, data berupa angka numerik yang berlanjut memerlukan pembagian menjadi beberapa interval. Di tahap ini akan melakukan proses pemilihan data yang dibutuhkan oleh

metode *data mining* yang digunakan, proses pemilihan dan transformasi data ini akan diperoleh kualitas melalui hasil *data mining* sebab terdapat beberapa kerakteristik di metode-metode *data mining* tertentu yang bergantung dari tahapan ini.

4. *Data mining,* tahapan ini berupa proses menemukan pola atau mencari informasi menarik yang terdapat pada data-data pilihan menggunakan metode maupun teknik tertentu. Algoritma maupun metode yang terdapat dalam *data mining* memiliki berbagai jenis dan sangat bervariasi, penggunaan algoritma atau metode yang tepat bergantung pada tujuan serta dengan keseluruhan berganutng pada proses KDD. Untuk data yang bisa digunakan menjadi sebuah model yang baik data harus tercukupi sebagai data riset, apabila semakin banyaknya data yang digunakan maka semakin sedikit kesalahan atau *error* sehingga semakin bagus model yang dijadikan sebagai acuan.
5. *Evaluation/Interpretation,* pada tahapan akhir ini proses *data mining* yang menghasilkan pola informasi memerlukan penampilan dalam bentuk yang mudah dimengerti oleh pihak yang membutuhkan atau pihak yang berkepentingan. Tahapan ini juga termasuk cakupan pemeriksaan pola maupun informasi yang ditemukan apakah menimbulkan tentangan dengan hipotesis dan fakta yang telah ada pada sebelumnya.

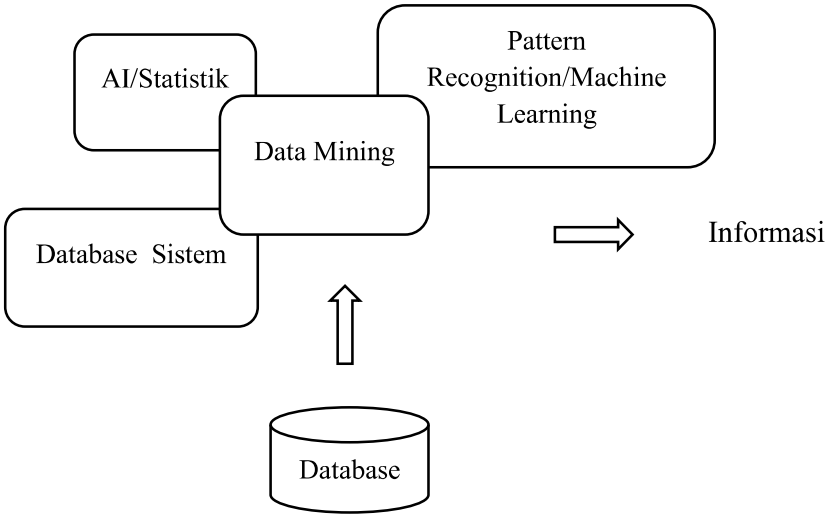
2.2. *Data Mining*

2.2.1. Sejarah *Data Mining*

Istilah *data mining* popular sejak tahun 1990-an di komunitas pengguna basis data. Namun, metode dan teori dasar dari *data mining* telah lahir sebelum era 90. *Data mining* berasal melaui berbagai macam disiplin ilmu, disiplin ilmu data mining memiliki dua ilmu yang paling mendasari yakni *machine learning* dan statistik. Teori-teori statistik yang berberasal dari teori matematika menitikberatkan pada pembentukan model. Model merupakan pendekatan struktur atau asumsi yang mendekati data yang sesungguhnya, melainkan *machine learning* ini lebih memprioritaskan pengembangan algoritme. Awal perkembangan *data mining* dimulai tahun (1763) saat Thomas Bayes mempublikasikan Teorema Bayes. Teori ini merupakan teori yang sangat penting dalam *data mining* karena mengestimasi kemungkinan suatu kejadian berdasarkan sebuah kejadian yang telah terjadi atau sedang berlangsung. Pada tahun (1805) mulai berkembangnya teori regresi yang mempelajari hubungan antar variabel, regresi juga menjadi salah satu bagian sebagai alat penting didalam data mining. Penggunaan komputer untuk mengolah data-data dalam jumlah yang besar dimulai saat Alan Turing memperkenalkan ide mesin pengolah data-data yang bersifat universal pada tahun (1936).

2.2.2. Asal Ilmu *Data Mining*

Asal mula ilmu *data mining* berawal dari irisan berbagai macam disiplin ilmu pengetahuan anrata lain : sistem basis data, kecerdasan buatan atau statistik, serta pattern recognition atau machine learning.



Gambar 2. 2 Asal Ilmu *Data Mining*
Sumber: Gambar Peneliti (2020)

1. *Artificial Intelegence (AI)* atau kecerdasan buatan dan statistik, *AI* adalah salah satu disiplin ilmu yang sangat penting di dalam pembangunan *data mining* dengan teknik pengolahan informasi didasari pada pola piker atau penalaran manusia. Dengan adanya statistik data yang telah diolah dapat dirangkum ke EDA atau *ecploratory data analysis, ecploratory data analysis* ini berfungsi sebagai pengidentifikasi hubungan antarvariabel yang sistematis atau fitur apabila tidak mencukupi informasi alami yang telah dibawanya.

2. *Pattern Recognition/Machine Learning,* pada dasarnya *data mining* merupakan pengenalan pola tetapi hanya terbatas pada pola basis data. Data yang akan diambil tidak dalam bentuk relasi, melainkan dalam bentuk normal pertama.
3. Sistem basis data, sistem basis data ini menyediakan sejumlah informasi berupa data-data yang akan diolah.

2.2.3. Pengertian *Data Mining*

Berbagai macam pendefinisian *data mining*, sebagai berikut:

(Lailil & Dkk, 2018)

a. Penguraian yang tidak sederhana dari sekumpulan data menjadi informasi yang memiliki potensi secara implisit tidak nyata atau jelas yang sebelumnya tidak diketahui.
b. Penggalian dan analisis, dengan menggunakan peranti ototmatis atau otomatis, dari sejumlah besar data yang bertujuan untuk menemukan pola yang memiliki arti.
c. Data mining juga merupakan bagian dari *knowledge discovery* dalam database *Knowledge Discovery in Database* atau KDD.

Data mining ialah sebuah proses penggalian atau penambangan data terhadap data yang tidak dapat diolah dengan cara manual karena jumlah data terlalu besar, menggunakan satu atau beberapa teknik untuk menganalisis sehingga memperoleh informasi yang dapat dimengerti serta berguna untuk

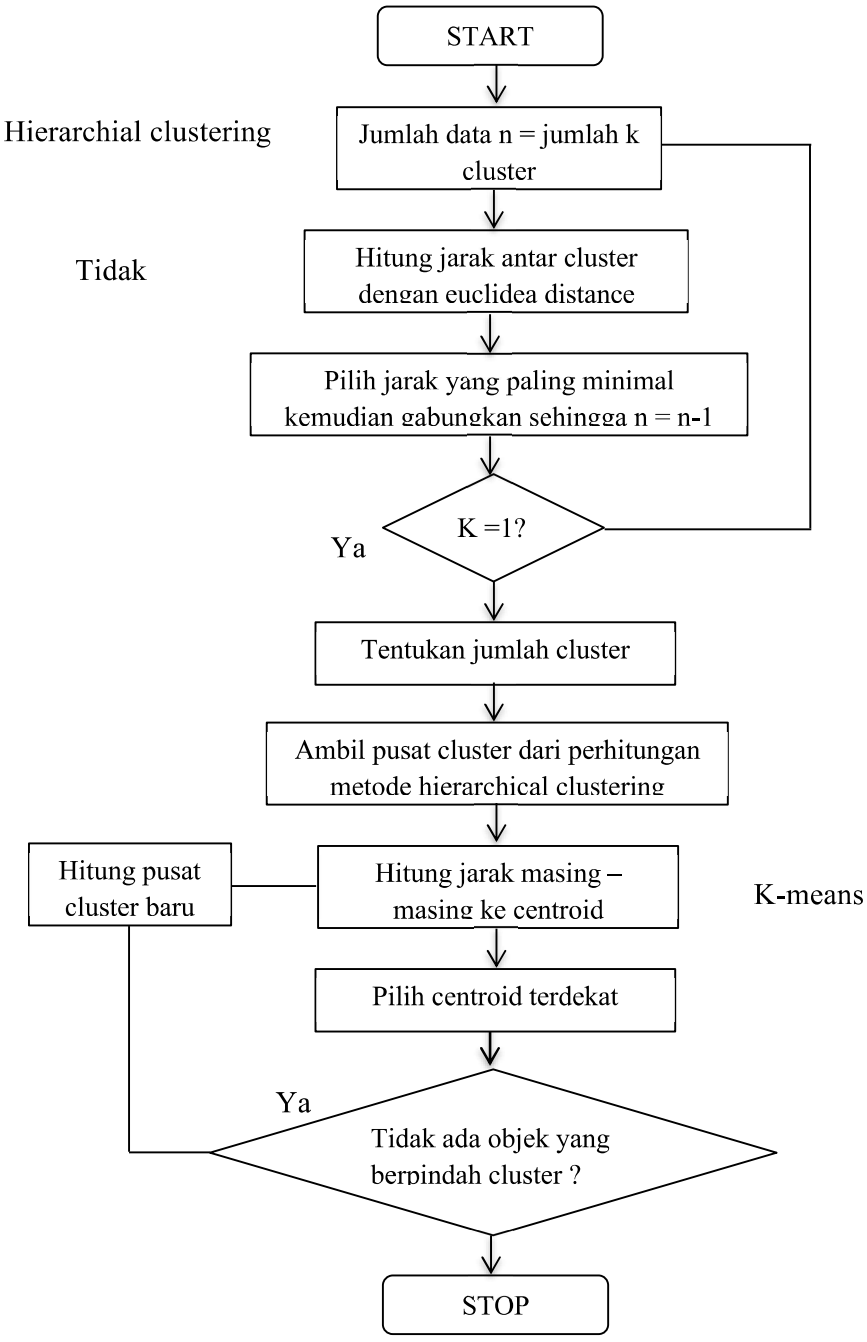
pemilik data yang sebelumnya tidak disadari keberadaan informasi tersebut yang mengandung nilai positif untuk kedepannya.

2.2.4. Algoritma *Data Mining*

2.2.4.1. Algoritma *k-means clustering*

K-means merupakan sebuah algoritma *clustering* yang mengalami pengulangan, yang mengelompokkan sejumlah objek menggunakan variabel tertentu ke dalam kelompok-kelompok, pada algortima *k-means* jumlah *cluster* harus ditentukan terlebih dahulu. Algoritma ini menetapkan nilai-nilai cluster secara random dan sementara nilai cluster tersebut akan menjadi sebuah pusat dari *centroid* atau *means.* Untuk menghitung jarak setiap data pada beberapa *centroid* menggunakan sebuah rumus *Euclidian Distance* , rumus ini digunakan hingga ditemukan jarak paling dekat pada setiap data dengan *centroid* dan perhitungan ini akan dilakukan terus menerus hingga nilai *centroid* tidak lagi berubah.

Berikut merupakan langkah - langkah untuk proses nilai centroid agar nilai centroid tidak berubah:



Gambar 2. 3 Proses Mencari Nilai *Centroid*
Sumber: Gambar Peneliti (2020)

1. Tentukan nilai k sebagai jumlah *cluster* yang akan di bentuk, temukan titik pusat *cluster* / k *centroid* awal secara acak / random dari objek-objek yang ada sebanyak k *cluster.* Untuk menghitung centroid cluster ke-*i* selanjutnya menggunakan rumus berikut:

$$v = \frac{\sum_{i=1}^{n} xi}{n} \; ; \; i = 1,2,3\ldots.n$$

Rumus 2. 1 *Centorid Cluster*

Keterangan :

v = *centroid* pada *cluster*

xi = objek ke-*i*

n = jumlah / banyaknya objek yang menjadi bagian / anggota *cluster*

2. Gunakan rumus *Euclidean Distance* hitung jarak *centroid* setiap objek dari masing-masing *cluster* :

$$d\,(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{n}\{xi - xi\}^2} \; ; \; i = 1,2,3\ldots.n$$

Rumus 2. 2 *Euclidean Distance*

Keterangan :

xi = objek *x* ke-*i*

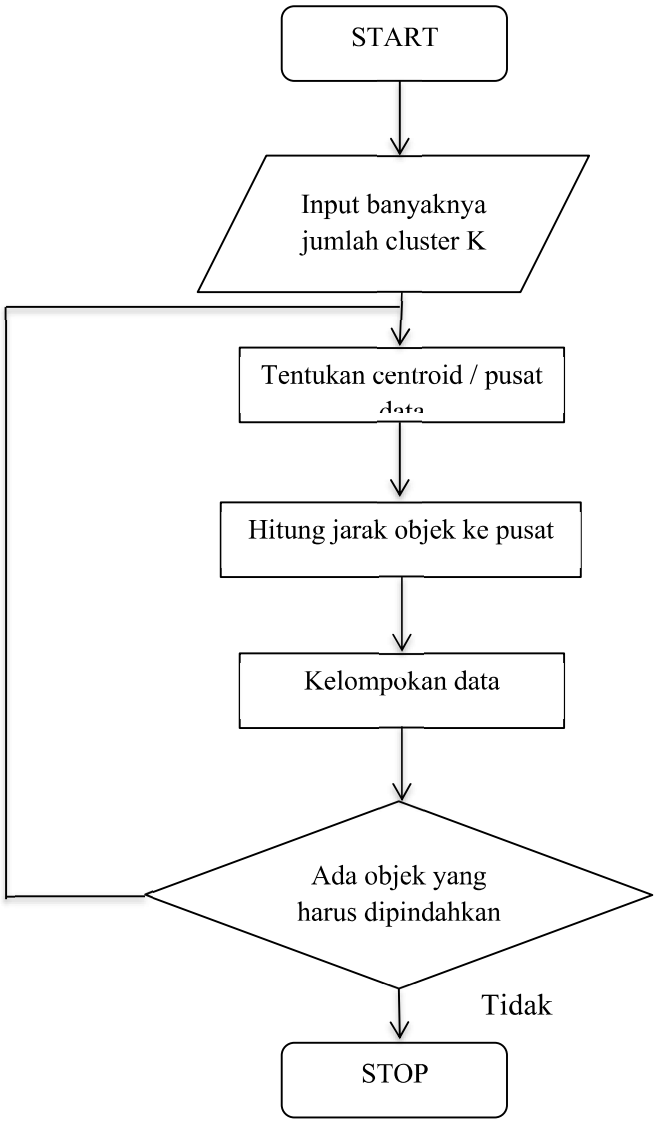
yi = daya y ke-*i*

n = banyaknya objek

3. Selanjutnya melakukan pengalokasikan masing-masing objek ke *centroid* yang terdekat, lakukan iterasi dan tentukan posisi *centroid* baru dengan menggunakan persamaan. Lakukan pengulangan perhitungan langkah ke 3, apabila posisi *centroid* baru memiliki nilai yang tidak sama dengan iterasi

sebelumnya. Ababila nilai yang dihasilkan sama dengan iterasi sebelumnya makan iterasi berhenti.

Berikut merupakan *flowchart* dari algortima *k-means clustering*, sebagai berikut:



Gambar 2. 4 *Flowchart* Proses *K-means*
Sumber: Gambar Peneliti (2020)

Penjelasan langkah-langkah dari gambar 2.4 *flowchart* proses *k-means*, yaitu:

1. Tentukan jumlah *cluster*, langkah pertama pada metode algoritma k-means ialah penentuan jumlah *cluster* didasarkan data yang telah diperoleh.
2. Tentukan pusat *cluster* awal, pusat *cluster* awal diperoleh dari data sendiri dengan mereandom atau acak pusat *cluster* awal melalui data yang telah diperoleh.
3. Hitung jarak objek ke pusat *cluster*, menggunakan rumus *Euclidean Distance* berfungsi mengukur jarak antara data dengan pusat *cluster*.

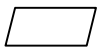
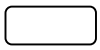
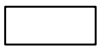
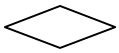
   Tahap algoritma perhitungan jarak data dengan pusat cluster yaitu :

   1. Ambil nilai data dan nilai pusat *cluster* yang sudah diperoleh
   2. Melakukan hitungan menggunakan *Euclidean Distance* data dengan tiap pusat *cluster*
4. Mengelompokan objek-objek data, lakukan perbandingan jarak dari hasil perhitungan kemudian lakukan pemilihan jarak yang terdekat antara data dengan pusat *cluster*, pemilihan jarak ini memperlihatkan data tersebut berada dalam kelompok yang sama atau satu kelompok dengan pusat *cluster* yang terdekat. Tahap algoritma mengelompokan objek-objek data yaitu :

   1. Ambil nilai jarak pada tiap pusat *cluster* pada data-data
   2. Temukan nilai jarak terkecil dari data
   3. Kelompokan data-data dengan pusat *cluster* yang memiliki jarak terkecil.

5. Tentukan pusat *cluster* yang baru, pusat *cluster* yang baru berfungsi untuk melakukan iterasi berikutnya (apabila hasil tidak konvergen). Proses iterasi akan dihentikan apabila hasil telah konvergen antara pusat *cluster* baru dengan pusat *cluster* lama.
6. Hitung jarak pusat *cluster,* melakukan perhitungan menggunakan *Euclidean Distance* dari semua data ketitik pusat yang baru atau ke C1 dan C2 (yang telah dilakukan pada tahap 2). Setelah memperoleh hasil perhitungan berikutnya bandingkan hasil tersebut.
7. Berikut merupakan penjelasan simbol-simbol yang terdapat pada *flowchart* proses *k-means*, sebagai berikut:

Tabel 2. 1 Simbol-Simbol *Flowchart* Proses *K-means*

Simbol	Nama	Deskripsi
	*Flow Direction Symbol*	*Flow direction symbol* ialah simbol yang digunakan untuk mengkoneksikan atau menghubungkan simbol yang satu dengan simbol yang lainya.
	Simbol *Input-Output*	*Input-Output symbol* ialah simbol yang menandakan proses input dan output data.
	*Terminator Symbol*	Termintar ialah simbol untuk menyatakan permulaan begin/mulai dan stop/akhir dari suatu proses.
	*Processing Symbol*	*Processing symbo*l ialah simbol yang menandakan pengolahan yang dijalankan oleh pc.
	*Decision Symbol*	*Decision symbol* ialah simbol pemilihan proses kegiatan berdasarkan kondisi yang telah ada.

Sumber: Tabel Peneliti (2020)

2.2.4.2. Algoritma apriori

Algoritma apriori merupakan algoritma mencari pola hubungan atau kombinasi item satu dengan item lainnya dalam sebuah data, kombinasi ini mengandung sebuah nilai keseringan pengambilan antara item satu dengan item yang lain. Hasil dari agoritma apriori ini dapat dimanfaatkan untuk membantu pihak yang menyediakan suatu barang dalam pengambilan keputusan dalam manajemen penempatan barang mau pun strategi pemasaran berupa diskon untuk kombinasi barang tersebut.

2.2.4.3. Algortima *nearest neighbor*

Algoritma *Nearest Neighbor* merupakan algoritma yang mencari kasus dengan cara menghitung kedekatan antara kasus yang lama dengan kasus yang baru, jadi kasus lama digunakan sebagai pedoman untuk menghasilkan sebuah hasil untuk pemecahan masalah pada kasus baru, berdasarkan pencocokan-pencocokan bobot dari sejumlah fitur yang telah ada.

2.2.4.4. Operasi dasar *data mining*

Operasi dasar dalam *data mining* dikategorikan menjadi dua yaitu metode prediktif dan metode deskriptif. Metode prediktif merupakan metode yang mempunyai tujuan dalam mempresiksi atau memperkirakan nilai suatu variabel tertentu berdasarkan pada nilai variabel-variabel lain. Variabel tak bebas merupakan variabel yang dijadikan target untuk diprediksi, sedangkan variabel

bebas merupakan variabel-variabel yang dijadikan untuk membantu prediksi. Regresi dan klasifikasi termasuk dalam metode prediktif.

Regresi bertujuan untuk melakukan tugas prediksi, regresi ini memiliki konsep dasar pada *data mining* yang berasal dari teori statistika. Regresi akan mengidentifikasi beberapa relasi antar variabel terikat dengan variabel bebas, setelah menghasilkan suatu model matematika berdasarkan identifikasi relasi tersebut, hasil dari model matematika digunakan untuk memperkirakan nilai melalui suatu variabel terikat berdasarkan nilai variabel bebasnya. Dalam *data mining* klasifikasi mempunyai tujuan untuk mengelompokkan data-data menjadi sejumlah kelompok, pengelompokan ini memanfaatkan proses acuan data yang sudah diketahui kelas atau kelompoknya. Apabila terdapat data yang belum memiliki kelompok maka dapat ditentukan kelompoknya menggunakan proses perbandingan melalui data yang sudah diketahui kelompoknya.

Metode deskriptif merupakan metode yang memiliki tujuan untuk menemukan relasi atau pola melalui data yang mudah dipahami oleh manusia, *association rule* dan *clustering* termasuk dalam metode deskriptif.

association rule adalah metode pencarian relasi serta pola antar data dalam sekumpulan data-data, pola yang dihasilkan suatu data dapat diprediksi kemunculannya berdasarkan kemunculan data lainnya. Dan *clustering* bertujuan untuk membagi data-data ke dalam beberapa kelompok, pada *clustering* proses pengelompokkan data tidak membanding data-data lain yang telah diketahui kelompoknya. Data-data pada *clustering* dikelompokkan dengan cara membandingkan seluruh data yang belum memiliki kelompok kemudian membagi

data-data tersebut kedalam beberapa kelompok yang memiliki kemiripan antar data.

2.2.4.5. Tantangan dalam *data mining*

Dalam data mining memiliki tantangan yang akan dihadapi peneliti meliputi:

1. Data mining memiliki data yang kompleks, dalam kesatuan data terdiri dari beberapa bagian. Bagian tersebut saling bergantungan dan saling berhubungan.
2. Heterogen data dalam data mining, memiliki beraneka ragam data dan karakteristik yang tidak sama diantara satu data dengan data yang lainnya.
3. Data mining memiliki jumlah variabel dan *cluster* yang banyak dalam data yang akan diproses.
4. Skalabilitas dalam data mining, kemampuan menampung data yang berskala besar.
5. Distribusi dalam data mining, data dalam jumlah yang besar akan dilakukan proses pembagian data dengan metode data mining.
6. Kepemilikan data, data akan diperoleh dari pihak yang memilik data sesuai dengan topik peneliti.
7. Data mining membutuhkan proteksi data apabila terdapat data yang sensitif.

2.2.4.6. Etika dalam *data mining*

Data mining akan memberikan dampak positif maupun sebaliknya yang sangat bergantung pada penggunaannya, dampak negatif akan muncul apabila tidak diperhatikan etika dalam penggunaan data khususnya data-data yang berhubungan dengan data pelanggan yang bersifat pribadi. Seperti klasterisasi pelanggan berdasarkan golongan, ras, agama, suku bangsa, adat, usia maupun gender jika tidak bijak dalam penggunaan data tersebut akan menimbulkan masalah diskriminasi dan bisa berakibat merugikan kelompok-kelompok terentu. Akan tetapi, pembeda masalah usia atau gender tertentu akan memberikan efek yang positif jika diterapkan dalam masalah medis seperti beberapa jenis penyakit yang rentan diderita oleh kelompok usia tertentu atau kaum pria maupun wanita.

2.2.4.7. Manfaat *data mining*

Pemanfaatan data mining dilihat dari dua sudut pandang, yaitu sudut pandang komersial dan sudut pandang keilmuan. (Vulandari, 2017)

Sudut pandang di lihat dari sudut komersial, data mining dimanfaatkan untuk mengatasi penumpukan data yang berlebihan, bagaimana memanfaatkan dan menyimpan data-data tersebut. Adanya teknik komputasi dapat dimanfaatkan untuk menghasilkan informasi yang digunakan sebagai aset untuk meningkatkan kompetitif suatu perusahaan. Seperti, bagaimana memprediksi tingkat pembelian dan penjualan, mengetahui konsumen dan produk yang memiliki kesamaan atau kemiripan karakteristik, mengetahui bagaimana mengidentifikasi produk-produk yang telah terjual secara bersamaan dalam satu waktu dengan produk lainnya,

mengetahui menyusutnya pelanggan akibat persaingan, memprediksi prilaku bisnis dimasa depan, memprediksi tingkatan resiko dalam menentukan jumlah produksi.

Data mining dalam sudut pandang keilmuan dapat dimanfaatkan untuk menyimpan dan menganalisa data yang sanga besar serta bersifat *real time*. Seperti, pemindaian langit dengan menggunakan teleskop, *remote sensor* yang ditempatkan pada *remote TV*.

2.3. Metode *Data Mining*

Terdapat sejumlah fungsi atau metode dalam *data mining* yang dapat digunakan dalam menemukan, menambang dan menggali infromasi pengetahuan. Dalam metode *data mining* terdapat 5 metode utama, yaitu:

2.3.1. *Clustering*

Clustering yaitu metode pengelompokan yang membentuk kelompok objek dengan memiliki kesamaan atau kemiripan dengan objek-objek lainnya, dan membentuk kelompok objek lainnya jika memiliki ketidaksamaan atau ketidakmiripan. *Clusteing* ini melakukan proses pembagian keseluruh data-data menjadi kelompok yang terdapat kesamaan atau kemiripan.

2.3.2. Estimasi

Metode estimasi ini melakukan estimasi sebuah data atau nilai baru yang belum diketahui atau yang tidak memiliki keputusan berdasarkan histori data atau nilai yang telah ada, seperti melakukan estimasi terhadap pendapatan atau

penghasilan seseorang apabila informasi mengenai orang tersebut telah diketahui.

2.3.3. Prediksi

Metode predisi ini digunakan untuk memperkirakan suatu kejadian atau sebuah peristiwa tertentu terjadi atau memperkirakan suatu kejadian atau sebuah peristiwa yang memiliki nilai masa yang akan datang, seperti memprediksi saham satu tahun ke depan.

2.3.4. Klasifikasi

Metode klasifikasi merupakan proses penemuan model atau sebuah fungsi yang membedakan maupun menjelaskan kelas data atau konsep, yang memiliki tujuan untuk memperoleh perkiraan kelas melalui suatu objek yang labelnya tidak diketahui.

2.3.5. Asosiasi

Metode ini berfungsi untuk menemukan hubungan antar variable-variabel pada sebuah database yang berjumlah banyak. Metode ini secara umum disebut *Market Basker Analysis* atau analisis keranjang belanja, dimana mengidentifikasi hubungan yang bersangkutan kuat antar berbagai jenis produk yang akan diambil bersamaan dalam setiap pembelanjaan.

2.4. Software Data Mining

Pada saat ini telah banyak *software* atau aplikasi *data mining* yang dapat dimanfatkan untuk mempermudah pengguna untuk mendapatkan informasi, mulai

dari *software* atau aplikasi yang gratis hingga yang berbayar. Dibawah ini merupakan *software* atau aplikasi gratis terdiri dari :

1. *RapidMiner*

Gambar 2. 5 Logo *RapidMiner*
Sumber: rapidminer.com

Pada awalnya *RapidMiner* dikenal dengan *Yet Another Learning Environment* atau YALE. Pada tahun 2001 *RapidMiner* dikembangkan bersifat *Open Source* oleh Simon Fischer, Ralf Klinkenberg dan Ingo Mierswa. *RapidMiner* mampu bekerja di semua SO/sistem operasi yang ditulis dalam bahasa *Java. RapidMiner* sebagai *software open soure* untuk *data mining* yang sudah terkemuka di dunia. *RapidMiner* menempati peringkat pertama sebagai *software* data mining pada *polling* oleh KDnuggets, sebuah portal *data mining* pada 2010-2011. *RapidMiner* memberikan solusi dalam melakukan analisis prediksi, dan menggunakan beberapa teknik prediksi dan deskriptif dalam memberikan wawasan untuk para pengguna sehingga dapat mengambil keputusan yang terbaik dari hasil analisa. (Aprillia c & Dkk, 2013)

2. *Weka (Waikato Enviroment for Knowledge Analysis)*

Gambar 2. 6 Logo *Weka*
Sumber : analyticsinsight.net

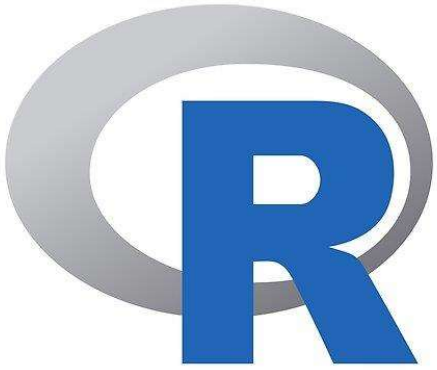
Weka merupaka *software* atau aplikasi *data mining* dengan tampilan yang sederhana bersifat *open source*, ditulis menggunakan bahasa Java. *Weka* pertama kali rilis pada 14 april 2016 oleh Universitas Waikato, *software* ini mendukung OS X, Linux dan *Windows*. Weka memiliki lingkungan untuk membandingkan beberapa algoritma pembelajaran, weka juga memiliki fasilitas untuk menganalisis data seperti perangkat-perangkat *pre-processing* data, metode-metode evaluasi serta algoritma pembelajaran.

3. *Orange*

Gambar 2. 7 Logo *Orange*
Sumber: github.com

Orange merupakan *software data mining* yang bersifat *open source* ditulis dalam bahasa *python, cython*, C++ dan C. *Orange* rilis pada 10 oktober 1996 oleh Universitas Ljubljana, yang mendukung pada *macOS, Linux dan Windows. Orange* memiliki keunggulan dalam hal visualisasi dapat digunakan dengan mudah tidak hanya untuk digunakan oleh para ahli, tetapi dapat digunakan oleh kalangan umum dan para pemula. Dengan menggunakan *software orange* dapat menganalisis suatu data-data penelitian, teks opini dari masyarakat, teks program kerja, membaca data, membandingkan algoritma pembelajaran, memvisualisasikan elemen data dan seterusnya.

4. *Rattle GUI*



Gambar 2. 8 Logo *Rattel GUI*
Sumber: redbubble.com

Rattle GUI dikembangkan oleh Graham Williams, dirilis pada 5 september 2017, *Rattle GUI* adalah *software open source* yang terdapat *user interface* dengan menggunakan bahasa pemrograman statistik R. *Rattle GUI* mampu menyajikan ringkasan-ringkasan data serta menghasilkan visualisasi

data melalui antarmuka pengguna grafis, *Rattle GUI* juga bisa dimanfaatkan sebagai fasilitas pengajaran untuk mempelajari dan mendalami bahasa perangkat lunak R. *Rattle GUI* menawarkan kemudahan untuk pengguna dalam melakukan penambangan data tanpa harus *coding.* Saat ini memiliki sejumlah perusahaan yang telah menggunakan aplikasi *Rattle GUI* untuk penambangan data salah satunya *Bank Commonwealth*.

5. *R studio*

Gambar 2. 9 Logo *R Studio*
Sumber: rstudio.com

R studio merupakan *software open source* yang tersedia untuk *Linux*, *Windows*, *MacOS* yang dikembangkan oleh *RStudio,Inc*. *R studio* merupakan *Integrated Development Environment* (IDE) dari bahasa pemrograman R yang berbahasa pemrograman bersifat standar untuk komputasi statistik dan grafik. *R studio* mewajibkan pengguna untuk menggunakan baris kode dalam melakukan analisis, berbeda dengan *software – software* seperti *Orange*, *Weka* dan lain sebagainya yang hanya mengklik untuk melakukan analisis.

2.5. Tujuan Umum Penelitian

Dari penelitian ini memiliki beberapa tujuan yaitu untuk menambah pengetahuan dan wawasan mengenai *data mining k-means clustering* yang bisa diterapkan untuk memberikan solusi dalam memecahkan masalah pada dunia kerja, kehidupan sehari-hari dan sebagainya.

2.6. Penelitian Terdahulu

Penelitian terdahulu sebagai salah satu alat bantu atau referensi penulis yang berguna untuk memperluas kajian-kajian yang dilakukan dalam penelitian.

Berikut beberapa referensinya :

1. (Windarto, 2017) dengan jurnal ISSN 1412-2693 yang berjudul **"Penerapan Datamining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan *K-Means Clustering Method*"**, Indonesia adalah salah satu negara pengekspor ke negara-negara maju dan berkembang. Tujuan dari eksportir adalah untuk dapat memperoleh keuntungan. Penelitian ini membahas tentang Penerapan Datamining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan *K-Means Clustering Method*. Sumber data penelitian ini dikumpulkan berdasarkan dokumen-dokumen keterangan ekspor impor yang dihasilkan oleh Direktorat Jenderal Bea dan Cukai. Data yang digunakan dalam penelitian ini adalah data dari tahun 2002- 2015 yang terdiri dari 11. Variable yang digunakan (1) jumlah ekspor berat bersih *(netto)* dan (2) *nilai Free On Board (FOB)*. Data akan diolah dengan melakukan clustering dalam 3 *cluster* yaitu *cluster* tingkat

ekspor tinggi, *cluster* tingkat ekspor sedang dan cluster tingkat ekspor rendah. *Centroid* data untuk *cluster* tingkat ekspor tinggi 904.276,5, *Centroid* data untuk *cluster* tingkat ekspor sedang 265.501 dan *Centroid* data untuk *cluster* tingkat ekspor rendah 34.280,1. Sehingga diperoleh penilaian berdasarkan indeks ekspor buah-buahan dengan 2 negara cluster tingkat ekspor tinggi yakni India dan Pakistan, 3 negara *cluster* tingkat ekspor sedang yakni Singapura, Bangladesh dan Negara lainnya dan 6 negara *cluster* tingkat ekspor rendah yakni Hongkong, Tiongkok, Malaysia, Nepal, Vietnam dan Iran. Hal ini dapat menjadi masukkan kepada pemerintah, negara yang menjadi prioritas tertinggi pada kegiatan ekspor buah-buahan berdasarkan klaster yang telah dilakukan.

2. (Novita Sari & Dkk, 2018) dengan jurnal ISSN 2407-1811 yang berjudul **"Penerapan *Metode K-Means Clustering* Dalam Menentukan Predikat Kelulusan Mahasiswa Untuk Menganalisa Kualitas Lulusan"**, Semakin meningkatnya jumlah mahasiswa yang diluluskan setiap tahunnya menyebabkan banyaknya data mahasiswa yang perlu diolah sehingga menyebabkan kesulitan dalam pengelompokan data tersebut. Pada penelitian ini menerapkan Data Mining dengan menggunakan metode *Clustering* untuk mengelompokkan kualitas lulusan mahasiswa Fakultas Ilmu Komputer Universitas Dehasen Bengkulu berdasarkan IPK dan Program Studi. Algoritma yang digunakan yaitu *K-Means Clustering*, dimana data dikelompokkan berdasarkan karakteristik yang sama akan dimasukkan ke dalam kelompok yang sama dan set data yang dimasukkan

ke dalam kelompok tidak tumpang tindih. Informasi yang ditampilkan berupa kelompok – kelompok lulusan mahasiswa yang mendominasi Program Studi, sehingga diketahui kelompok yang memiliki kualitas lulusan terbaik. Hasil penelitian ini akan membantu pihak Universitas dalam menganalisa kualitas mahasiswa yang diluluskan dan program studi yang paling berpotensi diminati. *Software* yang digunakan untuk membantu pengelompokan ini adalah *Rapid Miner*.

3. (Siregar, 2018) jurnal dengan ISSN 2622-1659 yang berjudul **"Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode *K-Means* (Studi Kasus Di Toko Adi Bangunan)"**. Dalam persaingan dunia bisnis saat ini, kita dituntut untuk senantiasa mengembangkan bisnis agar selalu bertahan dalam persaingan. Untuk mecapai hal tersebut, ada beberapa hal yang bisa dilakukan yaitu dengan meningkatkan kualiatas produk, penambahan jenis produk, dan pengurangan biaya operasional perusahaan dengan cara menggunakan analisis data perusahaan. Data mining adalah sebuah teknologi yang mengotomatisasi proses untuk menemukan pola menarik dan sensitif dari kumpulan-kumpulan data yang besar. Ini memungkinkan pemahaman manusia tentang menemukan pola dan skalabilitas teknik. Toko Adi Bangunan merupakan sebuah toko yang bergerak dalam bidang penjualan bahan-bahan bangunan dan peralatan rumah tangga yang memiliki sistem seperti pada swalayan yaitu pembeli mengambil sendiri barang yang akan dibeli. Data-data penjualan, pembelian barang maupun pengeluran tidak

terduga tidak tersusun dengan baik, sehingga data tersebut hanya berfungsi sebagai arsip bagi toko dan tidak dapat dimanfaatkan untuk pengembangan strategi pemasaran. Pada penelitian ini, data mining diterapkan menggunakan model proses *K-Means* yang menyediakan proses standar penggunaan *data mining* pada berbagai bidang digunakan dalam klasifikasi karena hasil metode ini mudah dipahami dan diinterpretasikan.

4. (Mardalius, 2018) jurnal dengan ISSN 2407-1811 yang berjudul **"Pemanfaatan *Rapid Miner* Studio 8.2 Untuk Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma *K-Means*"**. Toko Rafadel Acc menjual berbagai jenis aksesoris yang tersedia yang dijual di toko tersebut. Dari berbagai jenis aksesoris yang dijual tentu tidak semuanya yang laku terjual dan juga ada yang kurang laku serta ada juga yang tidak pernah terjual sama sekali. Dengan adanya masalah ini maka kita perlu melakukan perhitungan untuk menentukan atau mengelompokkan mana kategori aksesoris yang laku, kurang laku dan tidak laku terjual, dalam proses pengelompokan maka akan digunakan sebuah metode pengelompokan menggunakan Algoritma *K-Means Clustering* sebagai metode perhitungan secara manual dan dalam implementasinya maka digunakan sebuah *software Data Mining* menggunakan *RapidMiner* Studio versi 8.2. Dengan adanya aplikasi *RapidMiner Studio* ini pemilik toko dapat melihat hasil pengelompokan aksesoris mana yang paling laku, laku dan kurang laku. Maka, bila terdapat produk yang tidak laku, pemilik toko dapat mencari alternative

lain agar aksesoris yang tidak laku dapat menjadi laku. Metode yang digunakan dalam pengumpulan data adalah observasi dan wawancara kepada pemilik toko Rafadel Acc.

5. (Handoko & Lemana, 2018) jurnal dengan ISSN 2621-8794 yang berjudul **"Pengelompokkan *Data Mining* Pada Jumlah Penumpang di Bandara Hang Nadim"**. *The concept of data mining becomes one of the important tools in information management because the existing information has an increasing number. Data mining has many techniques in practice, one of which is the clustering technique which is the process of grouping data into groups so that data exist in the same group have properties as closely as possible. Clustering has many different methods, one of which is K-Means. By using ata mining clustering on traffic activity data taken from Hang Nadim Airport Batam, it can be obtained by grouping passenger based on clusters according to the nature of each data. The data taken include the number of passengers coming, departing, and transiting. In the process of performing data mining clustering, existing sample data must go through several important stages in order to get the correct cluster results. Stages that must be passed the Stages of Data Processing, Clustering Stage and Stage Algorithm. Based on the results of research that has been done on the existing sample data, it can be concluded the results of data grouping of passengers at Hang Nadim Airport Batam.*
6. (Purba & Dkk, 2018) jurnal dengan doi: 10.1088/1742-6596/1007/1/012049 yang berjudul **" *The effect of mining data k-means***

clustering toward students profile model drop out potential". *The high of student success and the low of student failure can reflect the quality of a college. One of the factors of fail students was drop out. To solve the problem, so mining data with K-means Clustering was applied. K-Means Clustering method would be implemented to clustering the drop out students potentially. Firstly the the result data would be clustering to get the information of all students condition. Based on the model taken was found that students who potentially drop out because of the unexcit ing students in learning, unsupported parents, diffident students and less of students behavior time. The result of process of K-Means Clustering could known that students who more potentially drop out were in Cluster 1 caused Credit Total System, Quality Total, and the lowest Grade Point Average (GPA) compared between cluster 2 and 3*.

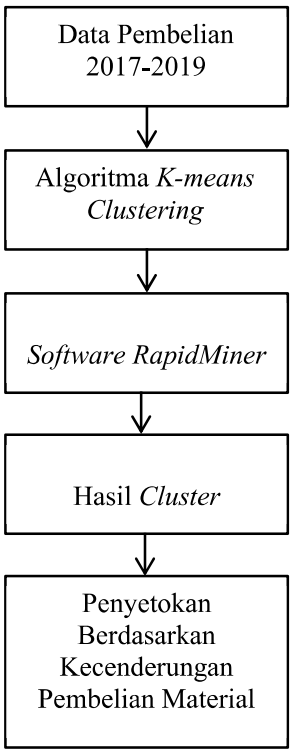
7. (Pernia & Dkk, 2018) jurnal dengan doi: 10.1088/1757-899X/364/1/012045 yang berjudul ***"A data mining approach for indoor air assessment, an alternative tool for cultural heritage conservation"***. *The exposure of cultural heritage to the environment has a significant impact on its degradation process and degradation rate. Consequently, managing the indoor air quality is vital to minimize further damage to historical artefacts and works of art. Despite its potential impact, the traditional assessment of the indoor air quality still represents a challenge for most collection guardians. This approach typically relays on the comparison of measured environmental parameters and corresponding*

acceptable values. However, determining the acceptable values and relative importance of the different environmental parameters turns out to be quite complex since it depends on the material types present in the collection and their preservation state. Furthermore, the significant amount of data generated during the measurements hampers the application of traditional methods of analysis. Considering all these, we propose the use of data mining as an alternative method for the indoor air quality assessment in cultural heritage studies. Data mining can provide knowledge from vast volumes of heterogeneous data, through high-speed processing, detection, and analysis. Here we present its application to identify dynamics and patterns affecting the indoor air quality in a realistic case. Using data from a measuring campaign held at a late Gothic church in Belgium, we show that inappropriate periods can be identified without using standards. In addition, different types of periods can be identified by studying the relation between multiple parameters. For that we use the k-means clustering method, interpreting the results with both visual and statistical tools.

2.7. Kerangka Pemikiran

Kerangka pemikiran ialah gambaran alur berpikir yang dibangun dari dasar teori dan referensi-referensi yang mengarahkan peneliti sampai pada dugaan sementara dari pemecahan masalah yang telah dirumuskan. Bagian ini berisi bukti-bukti empiris, teori-teori, dan pemikiran logis dari peneliti. Kerangka pemikiran akan lebih secara operasional apabila dilengkapi dengan *flow chart* atau

diagram alir yang menggambarkan rangkaian alur berpikir peneliti. Deskripsi kerangka pemikiran ialah landasan bagi perumusan hipotesis. (Toto & Dkk, 2015)



Gambar 2. 10 Kerangka Pemikiran
Sumber: Gambar Peneliti (2020)

Berdasarkan kerangka pemikiran diatas, data pembelian yang telah dikumpulkan dari tahun 2017 hingga 2019 serta telah diseleksi dan ditransformasi kemudian diolah menggunakan metode algoritma *k-means clustering*. Setelah itu, hasil data dari *k-means clustering* diimplementasikan kedalam *software* atau aplikasi *RapidMiner* untuk menemukan *cluster* pada data, setelah ditemukan *cluster* dapat ditentukan kecenderungan pembelian material yang akan digunakan untuk memprediksi penyetokan material pada proyek sejenisnya dimasa yang akan datang.