

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1. Landasan Teori**

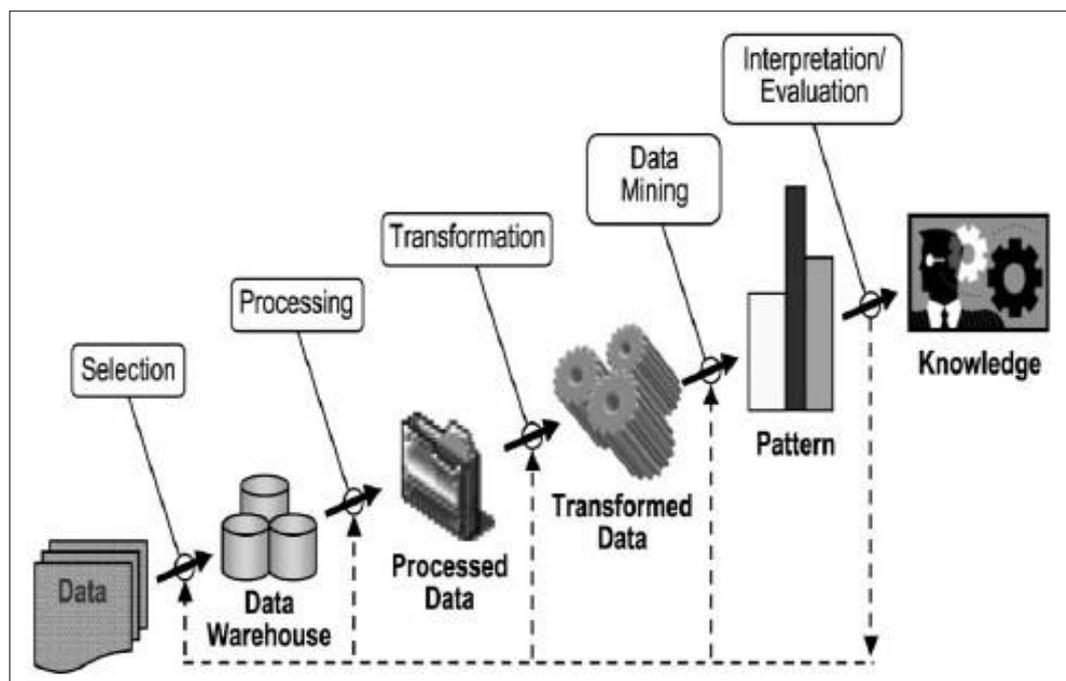
##### **2.1.1. Pengertian *Data Mining***

*Data mining* merupakan istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Turban & dkk, 2005)

Hampir senada dengan Turban, Beni (2014) mendefinisikan *data mining* sebagai kegiatan menemukan pola yang menarik dari data yang berjumlah besar, data dapat disimpan di *database*, *data warehouse*, atau penyimpanan informasi x lainnya. Pola yang ditemukan merupakan pola yang menarik, berharga dan penting dalam memberikan keuntungan kepada pemilik data. Pendapat yang mirip juga dikemukakan oleh Geetha dan Manimekalai (2013) bahwa *data mining* adalah proses menggunakan berbagai alat analisa data untuk menemukan pola dan hubungan dalam data yang dapat digunakan untuk membuat prediksi yang *valid*. Dari berbagai pendapat diatas, dapat disimpulkan bahwa *data mining* adalah serangkaian proses dalam menemukan pengetahuan tersembunyi dan berharga dari sekumpulan data yang dapat digunakan untuk membuat prediksi yang *valid*.

### 2.1.2. Tahapan *Data Mining*

*Data mining* merupakan rangkaian beberapa proses, maka *data mining* dapat dibagi menjadi beberapa tahap. Tahap-tahap data mining dapat diilustrasikan pada gambar 2.1



**Gambar 2.1** Tahapan *Data Mining*

(Sumber: <https://www.kajianpustaka.com/2017/09/data-mining.html>)

Berdasarkan gambar diatas, tahap-tahap *data mining* adalah:

1. *Data selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dimulai. Pada proses ini data diseleksi, hasil

dari data yang sudah diseleksi akan digunakan untuk proses data mining, misalnya data apa saja yang dibutuhkan untuk proses selanjutnya.

#### 2. *Pre-processing (Cleaning)*

Proses pembersihan data untuk membuang data yang tidak konsisten dan *noise*.

#### 3. *Transformation*

Data diubah menjadi bentuk yang sesuai dan diperlukan untuk *di-mining*. Transformasi dan pemilihan data ini juga menemukan kualitas dari hasil *data mining* nantinya karena ada beberapa karakteristik dari teknik-teknik *data mining* tertentu yang tergantung pada tahap ini.

#### 4. *Data Mining*

*Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi.

#### 5. *Interpretation (Evaluation)*

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

### **2.1.3. Pengelompokkan *Data Mining***

Menurut Larose dalam (Setiawan & Tes, 2016), data mining dapat dikelompokkan menjadi 6, yaitu:

1. Deskripsi

Deskripsi merupakan cara untuk mendeskripsikan pola dan kecenderungan yang terdapat dalam data yang dimiliki.

2. Estimasi

Estimasi hampir sama dengan klasifikasi, dalam estimasi variabel target lebih ke arah numerik dari pada ke arah kategorik. Model yang dibangun menggunakan *record* lengkap yang menyediakan nilai variabel target sebagai nilai prediksi.

3. Prediksi

Berbeda dengan klasifikasi dan estimasi, pada prediksi nilai dari hasil akan ada dimasa mendatang

4. Klasifikasi

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

5. Pengklusteran

Pengklusteran merupakan pengelompokkan *record*, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidak miripan dengan *record-record* dalam kluster lain.

6. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

#### **2.1.4. Clustering**

Tujuan utama dari analisis *cluster* adalah untuk mengelompokkan objek- objek berdasarkan karakteristik yang dimilikinya. Analisis cluster mengklasifikasikan objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam cluster yang sama (Sri Mulyati1), 2015).

Klasterisasi banyak digunakan dalam berbagai bidang dengan beragam aplikasi vital yang sangat penting, diantaranya adalah: riset pasar, dimana klasterisasi digunakan untuk segmentasi dan *profiling* pelanggan yang membantu dalam merancang strategi-strategi produk, harga, tempat, dan promosi. Dalam data mining ada dua jenis metode clustering yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical* (Ong, 2013).

*Hierarchical clustering* adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya

akan membentuk sebuah cluster. Dendogram biasanya digunakan untuk membantu memperjelas proses hierarki tersebut (Ong, 2013).

Berbeda dengan metode *hierarchical clustering*, metode *non-hierarchical clustering* justru dimulai dengan menentukan terlebih dahulu jumlah cluster yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan *K-Means Clustering* (Santoso, 2010) dalam (Ong, 2013).

#### **2.1.5. *K-means Clustering***

*K-means* merupakan algoritma *clustering* yang berulang-ulang. Langkah pertama algoritma *K-means* dimulai dengan pemilihan secara acak  $K$ ,  $K$  merupakan banyaknya cluster yang ingin dibentuk. Kemudian secara random tetapkan nilai  $K$ . Nilai  $K$  yang sudah ditetapkan untuk sementara menjadi pusat dari cluster atau biasa disebut dengan centroid, mean atau “means”. Hitung jarak setiap data yang ada terhadap masing-masing centroid menggunakan rumus Euclidean hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Klasifikasikan setiap data berdasarkan kedekatannya dengan centroid (*distance space*). Beberapa cara yang telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan centroid termasuk diantaranya  $L1$  (*Mahanttan/City Block distance space*).  $L2$  (*Euclidean distance space*), dan  $Lp$  (*Minkowski distance space*) (Studi et al., 2016).

Menurut Santosa (2007) dalam (Ong, 2013), langkah-langkah melakukan *clustering* dengan metode *K-Means* adalah sebagai berikut:

1. Tetapkan jumlah cluster  $k$ .
2. Inisialisasi  $k$  pusat cluster ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat cluster diberiduberi nilai awal dengan angka-angka random.
3. Alokasikan semua data/ objek ke cluster terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke cluster tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat cluster. Jarak paling antara satu data dengan satu cluster tertentu akan menentukan suatu data masuk dalam cluster mana. Untuk menghitung jarak semua data ke setiap titik pusat cluster dapat menggunakan teori jarak *Euclidean* yang dirumuskan sebagai berikut:

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

Dimana:

$D(i, j)$  = jarak data ke  $i$  ke pusat *cluster*  $j$

$X_{ki}$  = data ke  $i$  pada atribut data ke  $k$

$X_{kj}$  = titik pusat ke  $j$  pada atribut  $k$

4. Hitung kembali pusat cluster dengan keanggotaan cluster yang sekarang. Pusat cluster adalah rata-rata dari semua data/ objek dalam cluster tertentu. Jika dikehendaki bisa juga menggunakan median dari cluster tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai.

5. Tugaskan lagi setiap objek memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai. Atau, kembali ke langkah nomor 3 sampai pusat *cluster* tidak berubah lagi

### **2.1.6. Pengertian Nasabah**

Nasabah adalah aset atau kekayaan utama perusahaan karena tanpa pelanggan perusahaan tidak berarti apa-apa. Ada istilah “pelanggan adalah raja” yang semua kebutuhan dan keinginan pelanggan harus dipenuhi, dalam banyak hal perusahaan memang harus mengikuti kebutuhan dan keinginan pelanggannya jika ingin produk yang ditawarkan laku dipasaran.

### **2.1.7. Pengertian Asuransi**

Menurut Ketentuan Pasal 246 KUHD dalam (Guntara, 2016), Asuransi atau Pertanggungan adalah Perjanjian dengan mana penanggung mengikatkan diri kepada tertanggung dengan menerima premi untuk memberikan penggantian kepadanya karena kerugian, kerusakan atau kehilangan keuntungan yang diharapkan yang mungkin dideritanya akibat dari suatu evenemen (peristiwa tidak pasti).

Asuransi atau pertanggungan adalah *perjanjian antara dua pihak atau lebih dengan mana pihak penanggung mengikatkan diri kepada tertanggung dengan menerima premi asuransi untuk memberikan penggantian kepada tertanggung*

*karena kerugian, kerusakan atau kehilangan keuntungan yang diharapkan, atau tanggung jawab hukum kepada pihak ketiga yang mungkin akan diderita tertanggung yang timbul dari suatu peristiwa yang tidak pasti, atau untuk memberikan suatu pembayaran yang didasarkan atas meninggal atau hidupnya seseorang yang dipertanggungjawabkan*, menurut Ketentuan Undang-undang No.2 tahun 1992 tertanggal 11 Februari 1992 tentang Usaha Perasuransian (“UU Asuransi”) yang sudah dicabut oleh Undang-undang No. 40 tahun 2014 tertanggal 17 Oktober 2014 tentang Perasuransian yang memuat pengertian asuransi sebagai berikut : Asuransi adalah perjanjian antara dua pihak, yaitu perusahaan asuransi dan pemegang polis, yang menjadi dasar bagi penerimaan premi oleh perusahaan asuransi sebagai imbalan untuk:

- a. memberikan penggantian kepada tertanggung atau pemegang polis karena kerugian, kerusakan, biaya yang timbul, kehilangan keuntungan, atau tanggung jawab hukum kepada pihak ketiga yang mungkin diderita tertanggung atau pemegang polis karena terjadinya suatu peristiwa yang tidak pasti; atau
- b. memberikan pembayaran yang didasarkan pada meninggalnya tertanggung atau pembayaran yang didasarkan pada hidupnya tertanggung dengan manfaat yang besarnya telah ditetapkan dan/atau didasarkan pada hasil pengelolaan dana.

### **2.1.8. Rapid Miner**

Software yang digunakan dalam penelitian ini adalah Rapid Miner. Tujuan dari penggunaan software ini adalah membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan dengan proses di Rapid Miner.

Rapid Miner adalah perangkat lunak yang digunakan untuk mengolah data. Rapid miner mengekstrak pola-pola dari dataset yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan serta database (Dumai et al., 2017). Dengan menggunakan Rapid Miner, pengguna dapat mengolah data dengan jumlah yang banyak dengan menggunakan operator-operator. Modifikasi data dilakukan dengan memanfaatkan operator-operator tersebut. Mekanismenya untuk melihat hasil, data tinggal hanya dihubungkan dengan node hasil. Rapid Miner mampu menampilkan hasil secara visual dan grafik.

## **2.2. Penelitian Terdahulu**

Beberapa hasil penelitian terdahulu yang memanfaatkan teknik *clustering* dalam berbagai bidang yaitu:

**Tabel 2.1** Penelitian Terdahulu

Judul	Nama Pengarang	ISSN	Hasil
Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University	Johan Oscar Ong (2013)	1412-6869	Penelitian ini digunakan untuk menentukan strategi maketing President University dengan mengelompokkan data mahasiswa yang telah lulus mulai dari nama mahasiswa hingga kota asal
Application of k-Means Clustering algorithm for prediction of Students' Academic Performance	Oyelade, O. J Oladipupo, O. O Obagbuwa, I. C	1947-5500	Lembaga swasta di Nigeria mengimplementasikan algoritma k-means untuk memprediksi dan memantau perkembangan kinerja akademik siswa di insitusi yang lebih tinggi yang bertujuan untuk membuat keputusan yang lebih efektifk dalam membuat perencanaan akademik dimasa yang akan datang.

**Tabel 2.1** Lanjutan

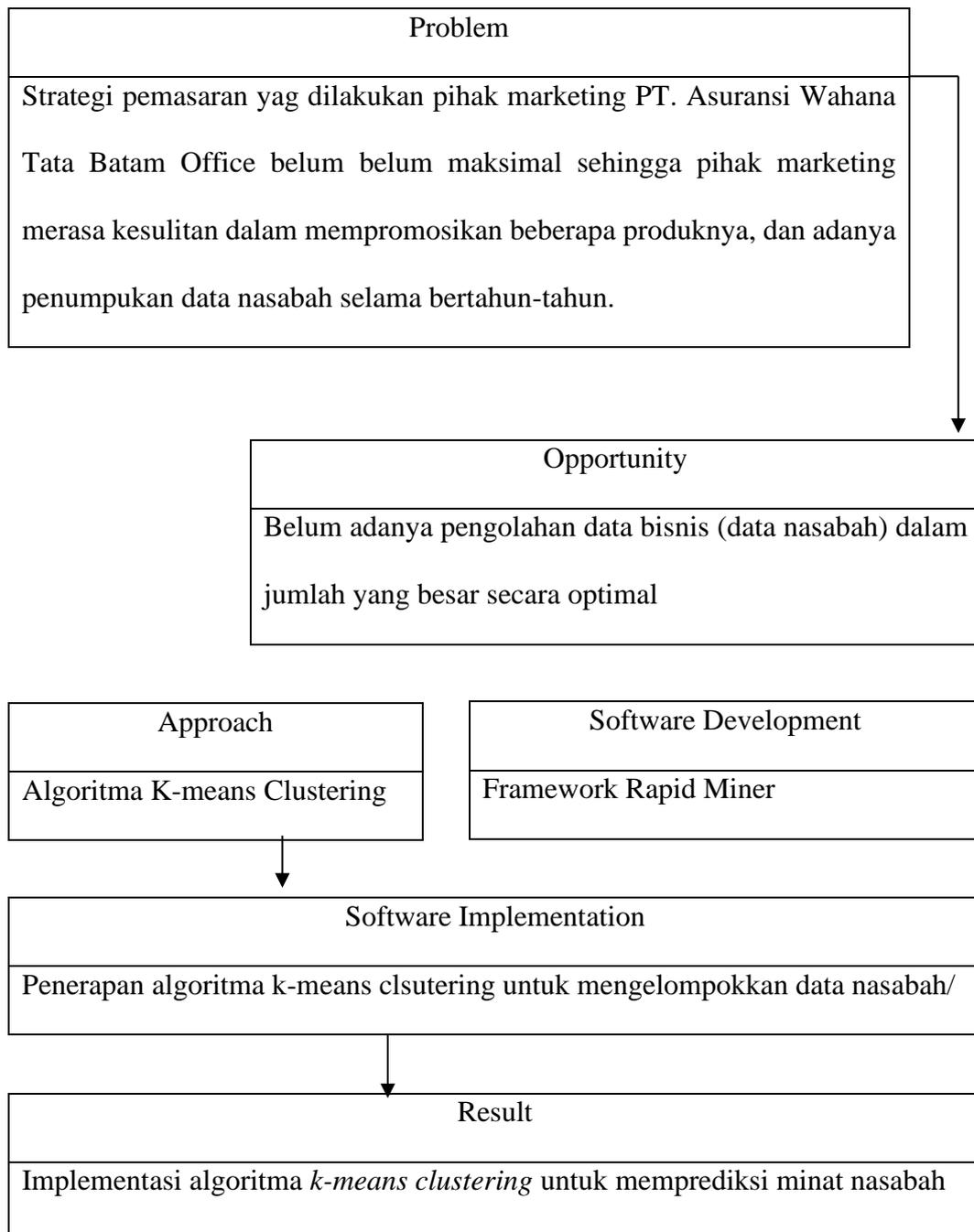
<p>Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa Baru (Studi Kasus: Universitas Islam Indragiri)</p>	<p>Fitri Yunita</p>	<p>2302-8149</p>	<p>Algoritma K-means clustering digunakan untuk membantu menentukan strategi mempromosikan masing – masing program studi yang ada di Universitas Islam Indragiri.</p>
<p>Penerapan Data Mining Dengan Metode Interpolasi Untuk Memprediksi Minta Konsumen Asuransi (Studi Kasus Asuransi Metlife)</p>	<p>Sandy Kurniawan, Taufiq Hidayat</p>	<p>0854-4743</p>	<p>Penelitian ini menggali informasi dari basis data nasabah dan transaksi pembayaran premi asuransi sehingga dapat memprediksi seberapa besar angsuran premi asuransi yang terbaik sesuai dengan kondisi calon nasabah.</p>

**Tabel 2.1 Lanjutan**

Implementasi Data Mining Pelanggan Potensial Menggunakan Algoritma K-Means	Randi Rian Putra, Cendra Wadisman	2614-1574	K-means diimplementasikan dalam pemilihan pelanggan potensial di MC Laundry.
--	-----------------------------------	-----------	--

### 2.3. Kerangka Pemikiran

Kerangka pemikiran adalah narasi (uraian) atau pernyataan (proposisi) tentang kerangka konsep pemecahan masalah yang telah diidentifikasi atau dirumuskan. Melalui uraian dalam kerangka berpikir, peneliti dapat menjelaskan secara komprehensif variabel-variabel apa saja yang diteliti dan dari teori apa variabel-variabel itu diturunkan, serta mengapa variabel-variabel itu saja yang diteliti. Uraian dalam kerangka berpikir harus mampu menjelaskan dan menegaskan secara komprehensif asal-usul variabel yang diteliti, sehingga variabel-variabel yang tercantum di dalam rumusan masalah dan identifikasi masalah semakin jelas asal-usulnya (<https://id.scribd.com/id>, n.d.).



**Gambar 2.2** Kerangka Pemikiran

(Sumber: Data Penelitian, 2018)