# BAB II

## LANDASAN TEORI

### Tinjauan Pustaka

Dari berbagai aspek dan sudut pandang yang berbeda-beda mencakup sesuai dengan hal-hal yang berkaitan dengan penelitian sebagai berikut:

#### Algoritma C.45

“Algoritma C.45 Merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan. Aturan dapat dengan mudah di pahami dengan bahasa alami. Dan mereka juga dapat di ekpresikan dalam bentuk bahasa basis data seperti *Structure Query Language* untuk mencari *record* pada kategori tertentu.” [Kusrini dan Luthfi, 2009: 13]. Tahapan algoritma C.45 untuk membangun pohon keputusan, yaitu:

1. Pilih atribut sebagai akar.

Berikut data keputusan bermain tenis

| **No** | **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| --- | --- | --- | --- | --- | --- |
| 1 | Sunny | Hot | High | False | No |
| 2 | Sunny | Hot | High | True | No |
| 3 | Cloudy | Hot | High | False | Yes |
| 4 | Rainy | Mild | High | False | Yes |
| 5 | Rainy | Cool | Normal | False | Yes |
| 6 | Rainy | Cool | Normal | True | Yes |
| 7 | Cloudy | Cool | Normal | True | Yes |
| 8 | Sunny | Mild | High | False | No |
| 9 | Sunny | Cool | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | False | Yes |
| 11 | Sunny | Mild | Normal | True | Yes |
| 12 | Cloudy | Mild | High | True | Yes |
| 13 | Cloudy | Hot | Normal | False | Yes |
| 14 | Rainy | Mild | High | True | No |

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* menggunakan rumus persamaan 1 berikut:

Sumber: Kusrini (2009: 16)

Dimana:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut a

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Untuk menghitung nilai entropi menggunakan persamaan 2 berikut:

Sumber: Kusrini (2009: 16)

Dimana:

S : himpunan kasus

A : fitur

n : jumlah partisi S

Pi : proporsi dari Si terhadap S

Hitung jumlah kasus, jumlah kasus untuk keputusan **yes**, jumlah keputusan **no,** dan entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut ***OUTLOOK, TEMPERATURE, HUMIDITY,*** dan ***WINDY.*** dan lakukan perhitungan *Gain* untuk setiap atribut.

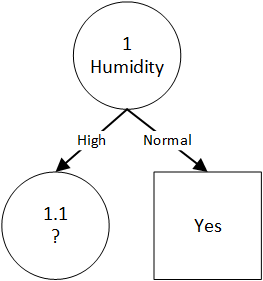
| **Node** |  |  | **Jumlah**  **Kasus (S)** | **Tidak**  **(S1)** | **Ya**  **(S2)** | **Entropy** | **Gain** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | TOTAL |  | 14 | 4 | 10 | 0.863120569 | 0.258532037 |
|  | OUTLOOK |  |  |  |  |  |  |
|  |  | CLOUDY | 4 | 0 | 4 |  |  |
|  |  | RAINY | 5 | 1 | 4 | 0.721928095 |  |
|  |  | SUNNY | 5 | 3 | 1 | 0.970950594 |  |
|  | TEMPERATURE |  |  |  |  |  | 0.183850925 |
|  |  | COOL | 4 | 0 | 4 | 0 |  |
|  |  | HOT | 4 | 2 | 2 | 1 |  |
|  |  | MILD | 6 | 2 | 4 | 0.918295834 |  |
|  | HUMIDITY |  |  |  |  |  | 0.370506501 |
|  |  | HIGH | 7 | 4 | 3 | 0.985228136 |  |
|  |  | NORMAL | 7 | 0 | 7 | 0 |  |
|  | WINDY |  |  |  |  |  | 0.005977711 |
|  |  | FALSE | 8 | 2 | 6 | 0.811278124 |  |
|  |  | TRUE | 6 | 4 | 2 | 0.918295834 |  |

Tabel II.2

Perhitungan Node 1.1

Dari hasil tabel diatas atribut dengan Gain tertinggi adalah ***HUMIDITY,*** yaitu sebesar 0.37. Dengan demikian ***HUMIDITY*** dapat menjadi node akar. Ada dua nilai atribut tersebut, nilai atribut ***NORMAL*** sudah mengklasifikasikan kasus menjadi 1, yaitu keputusan ***Yes***, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut ***HIGH*** masih perlu dilakukan perhitungan lagi.

Dari hasil tersebut dapat digambarkan pohon keputusan sementara sebagai berikut:



Gambar II.1

Pohon Keputusan Hasil Perhitungan Node 1

1. Buat cabang untuk tiap nilai-nilai.

Menghitung jumlah kasus, jumlah kasus untuk keputusan ***Yes,*** jumlah kasus untuk keputusan ***No,*** dan entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut ***OUTLOOK, TEMPERATURE, HUMIDITY,*** dan ***WINDY*** yang dapat menjadi node akar dari nilai atribut ***HIGH.*** Dan lakukan perhitungan *Gain* untuk setiap atribut.

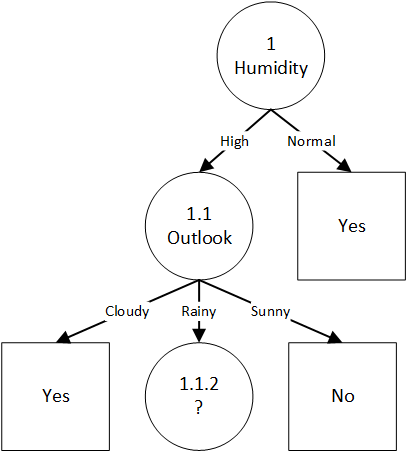
| **Node** |  |  | **Jumlah**  **Kasus (S)** | **Tidak**  **(S1)** | **Ya**  **(S2)** | **Entropy** | **Gain** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1.1** | HUMIDITY-HIGH |  | 7 | 4 | 3 | 0.985228136 |  |
|  | OUTLOOK |  |  |  |  |  | 0.69951385 |
|  |  | CLOUDY | 2 | 0 | 2 | 0 |  |
|  |  | RAINY | 2 | 1 | 1 | 1 |  |
|  |  | SUNNY | 3 | 3 | 0 | 0 |  |
|  | TEMPERATURE |  |  |  |  |  | 0.020244207 |
|  |  | COOL | 4 | 0 | 4 | 0 |  |
|  |  | HOT | 4 | 2 | 2 | 1 |  |
|  |  | MILD | 6 | 2 | 4 | 0.918295834 |  |
|  | WINDY |  |  |  |  |  | 0.005977711 |
|  |  | FALSE | 8 | 2 | 6 | 0.811278124 |  |
|  |  | TRUE | 6 | 4 | 2 | 0.918295834 |  |

Tabel II.2

Perhitungan Node 1.1

Dari hasil diatas diketahui bahwa atribut dengan gain tertinggi adalah ***OUTLOOK,*** yaitu sebesar 0.67. dengan demikian ***OUTLOOK*** dapat menjadi node cabang dari nilai atribut ***HIGH.*** Ada 3 nilai atribut dari ***OUTLOOK***, yaitu ***CLOUDY, RAINY,*** dan ***SUNNY*** sudah mengklasifikasikan kasus menjadi satu dengan keputusan ***No***, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut ***RAINY*** masih perlu dilakukan perhitungan lagi

Pohon keputusan yang terbentuk sampai tahap ini adalah sebagai berikut:



Gambar II.2

Pohon keputusan hasil perhitungan node 1.1

1. Bagi kasus dalam cabang.

Menghitung jumlah kasus jumlah kasus untuk keputusan ***Yes***, jumlah kasus untuk keputusan ***No***, dan entropy dari semua kasus yang dibagi berdasarkan atribut ***TEMPERATURE*** dan ***WINDY*** yang dapat menjadi cabang dari nilai atribut ***RAINY.*** Lakukan perhitungan gain untuk tiap atribut

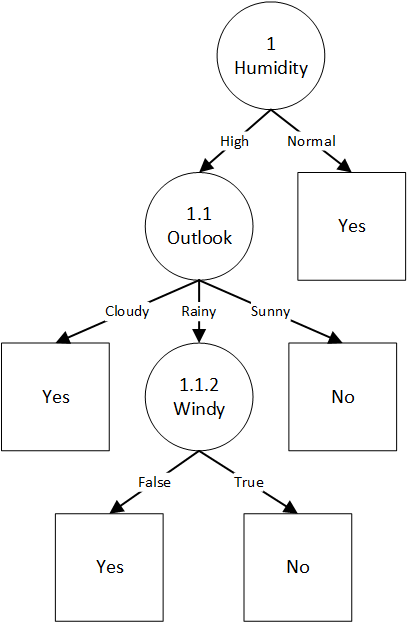
| **Node** |  |  | **Jumlah**  **Kasus (S)** | **Tidak**  **(S1)** | **Ya**  **(S2)** | **Entropy** | **Gain** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1.1.2** | HUMIDITY - HIGH dan  OUTLOOK - RAINY |  | 2 | 1 | 1 | 1 |  |
|  | TEMPERATURE |  |  |  |  |  | 0 |
|  |  | COOL | 0 | 0 | 0 | 0 |  |
|  |  | HOT | 0 | 0 | 0 | 0 |  |
|  |  | MILD | 2 | 1 | 1 | 1 |  |
|  | WINDY |  |  |  |  |  | 1 |
|  |  | FALSE | 1 | 0 | 1 | 0 |  |
|  |  | TRUE | 1 | 1 | 0 | 0 |  |

Tabel II.3

Perhitungan Node 1.1.2

Dari hasil diatas dapat diketahui bahwa atribut dengan Gain tertinggi adalah ***WINDY***, yaitu sebesar 1. Dengan demikian ***WINDY*** dapat menjadi node cabang dari nilai atribut ***RAINY***. Ada 2 nilai dari atribut ***WINDY*** yaitu ***FALSE*** and ***TRUE*** sudah mengklasifikasikan kasus menjadi satu dengan keputusan ***No,*** sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini.

Pohon keputusan yang terbentuk sampai tahap ini adalah sebagai berikut:



Gambar II.3

Pohon keputusan hasil perhitungan node 1.1.2

Berdasarkan gambar II.3 semua kasus sudah masuk dalam kelas, dengan demikian, pohon keputusan pada gambar II.3 merupakan pohon keputusan terakhir yang terbentuk

#### Pengertian Klasifikasi

“Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan” [Nofriansyah, 2014: 10]

#### Pengertian *Gain*

“Gain (S,A) merupakan Perolehan informasi dari atribut A relative terhadap output data S. Perolehan informasi didapat dari output data atau variabel dependent S yang dikelompokkan berdasarkan atributA, dinotasikan dengan gain (S,A) “ [Nofriansyah, 2014: 22]

#### Pengertian *Entropy*

“Entropy(S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S.” [Nofriansyah, 2014: 21]

#### Pengertian Kredit

“Kredit adalah kemampuan untuk melaksanakan suatu pembelian atau Mengadakan suatu pinjaman dengan suatu janji, pembayaran akan dilaksanakan pada jangka waktu yang telah disepakati.” [ Jusuf, 2014: 310]

#### Pengertian Weka

*WEKA* merupakan sebuah perangkat lunak yang menerapkan berbagai algoritma *machine learning* untuk melakukan beberapa proses yang berkaitan dengan sistem temu kembali informasi atau *data mining.*

#### Pengertian *Data Mining*

“*Data mining is the process of discovering useful patterns and trends in large*

*data sets. “*[Larose, 2015: 4]

Berdasarkan pengertian diatas dapat disimpulkan Data Mining adalah proses menemukan pola dan tren yang bermanfaat dalam data yang luas.

“*Data Mining* adalah serangkaian proses ekstraksi atau penggalian data dan informasi yang besar yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari *database* yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting.” [Indrajani, 2011: 289]

“Istilah *data mining* dan *knowledge discovery in databases* (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 satu sama lain.” [Kusrini dan Luthfi, 2009: 7]

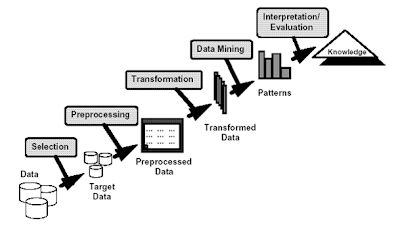
Berdasarkan pengertian-pengertian di atas dapat disimpulkan bahwa Data Mining adalah proses menggali dan mengekstrak informasi yang selama ini tidak diketahui atau tersembunyi dari suatu basis data untuk menemukan pengetahuan, yang penting untuk kepentingan perusahaan atau organisasi tersebut.

* + 1. **Proses Tahapan *Data Mining***

“*Data Mining* merupakan salah satu dari rangkaian *Knowledge Discovery in Database (KDD)*. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, intepretasi dan visualisasi dari pola-pola sejumlah data. Serangkaian proses tersebut memiliki tahapan sebagai berikut:” [Retno, 2017: 2]

1. Pembersihan data (untuk membuang data yang tidak konsisten dan *noise*)
2. Integrasi data (penggabungan data dari beberapa sumber )
3. Transformasi Data (data diubah menjadi bentuk yang sesuai untuk di-*mining*)
4. Aplikasi teknik *Data Mining*, proses ekstrasi pola daridata yang ada.
5. Evaluasi pola yang ditemukan (proses intepretasi pola menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan)
6. Presentasi pengetahuan (dengan teknik visualisasi)

Langkah terkahir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna.



Sumber: Retno (2017: 33)

Gambar II.5

Tahapan *Knowledge Discovery in Database* (KDD)

* + 1. **Pengelompokan *Data Mining***

*Data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu: [Kusrini dan Luthfi, 2009:10-12]

1. **Deskripsi**

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. [Kusrini dan Luthfi, 2009:10-12]

1. **Estimasi**

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. [Kusrini dan Luthfi, 2009:10-12]

1. **Prediksi**

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam predikasi nilai dari hasil akan ada dimasa mendatang. Contoh prediksi bisnis dan penelitian adalah:

a. Prediksi harga beras dalam tiga bulan yang akan datang.

b. Prediksi persentasi kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi. [Kusrini dan Luthfi, 2009:10-12]

1. **Klasifikasi**

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh penggolongan pendapatan dapat dipisahkan dalam tiga kategori , yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis danpenelitian adalah:

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau tidak.

b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.

c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit apa. [Kusrini dan Luthfi, 2009:10-12]

1. **Pengklusteran**

Pengkluteran merupakan pengelompokan *record,* pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster 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. Contoh pengklusteran dalam bisnis dan penelitian adalah:

a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari satu suatu produk bagi perusahaan yang tidak memiliki dana pemesaran yang besar.

b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku financial dalam baik dan mencurigakan.

c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar. [Kusrini dan Luthfi, 2009:10-12]

1. **Asosiasi**

Tugas asosiasi dalam data mining adalah menemukan attribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran upgrade layanan yang diberikan.

b. Menentukan barang dalam supermarket yang dibeli secara bersamaan dan yang tidak pernah dibeli secara bersamaan. [Kusrini dan Luthfi, 2009:10-12]

### Penelitian Terkait

Beberapa penelitian yang terkait dalam penulisan skripsi ini adalah sebagai berikut:

1. Kecelakaan merupakan suatu kejadian yang tidak terencana begitupun pada sebuah proyek konstruksi dimana kecelakaan sering terjadi hal ini disebabkan oleh berbagai faktor. Kita lihat pada Industri jasa konstruksi yang merupakan salah satu sektor industri yang memiliki risiko kecelakaan kerja yang cukup tinggi. Banyaknya kecelakaan kerja yang terjadi tidak terlepas dari faktor Human Error, tentunya berdampak pada kinerja dan pekerjaan yang dilaksanakan, Metode yang digunakan dalam analisis ini adalah Algoritma C4.5 yang merupakan salah satu algoritma modern untuk melakukan Data Mining, Algoritma C4.5 disebut juga dengan pohon keputusan (decision tree) yaitu merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon, dan pada setiap node merepresentasikan atribut,cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas, Konsep dari pohon keputusan ini adalah dengan mengumpulkan data selanjutnya dibuatkan decision tree yang kemudian akan dihasilkan rule-rule solusi permasalahan. dari hasil penelitian faktor-faktor yang menjadi penyebab terjadinya kecelakaan kerja kontrusksi yang sering terjadi adalah Lingkungan Tempat Kerja, Rambu-Rambu Keselamatan dan Pekerja dan Cara kerja. [Elisa, 2017:36].
2. Lemahnya pengawasan dalam proses pemberian kredit kepada karyawan PT. X Group menyebabkan tingginya kredit macet. Dalam menyalurkan kreditnya, PT. X Group haruslah pintar dalam menilai para nasabah dimasa yang akan datang apakah akan menguntungkan atau tidak. Faktor ini sangatlah penting bagi pihak perusahaan karena hal ini akan menunjukkan bahwa layak atau tidaknya suatu usaha atau individu yang akan diberikan pinjaman atau kredit, pada penelitian ini digunakan teknik data mining klasifikasi dengan metode C4.5 untuk mengetahui apakah nasabah tergolong nasabah lancar ataupun tidak. Dengan penelitian ini diharapkan dapat membantu pihak perusahaan dalam membaca pola pembayaran dari nasabahnya sehingga dapat menentukan apakah nasabah tersebut layak mendapatkan kredit atau tidak dan menghasilkan rule dari pohon keputusan yang diterapkan pada implementasi sistem klasifikasi data nasabah kredit di PT. X Group. [Pratama dkk, 2018: 121].
3. Kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan pinjam meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam melunasi hutangnya setelah jangka waktu tertentu dengan pemberian bunga, pada koperasi permasalahan kredit merupakan permasalahan manajemen, dimana jika banyak nasabah yang menunggak dalam pembayaran maka akan mengganggu system keuangan yang ada, untuk itulah penelitian ini menerapkan proses analisa kredit nasabah terlebih dahulu sebelum diambil sebuah keputusan pemberian kredit, analisa keputusan memberikan kredit menggunakan algoritma klasifikasi C4.5 dan Naïve Bayes dimana kedua algoritma tersebut dilakukan penilaian, mana algoritma yang paling akurat dalam menganalisa kemampuan nasabah dalam membayar kredit, analisa berdasarkan data history. Hasil yang didapatkan dari perbandingan kedua algoritma tersebut, bahwa tingkat akurasi yang lebih baik adalah menganalisa menggunaka algoritma klasifikasi C4.5 yaitu 88.90 % sedangkan untuk tingkat akurasi menggunakan algortima klasifikasi Naïve Bayes yaitu 80.00%.[Marsipah, 2016: 187]
4. Data rekam medis adalah catatan khusus pasien yang sedang, seringkali data rekam medis hanya menjadi data yang menumpuk dan tidak dilakukan penelusuran untuk menghasilkan pengetahuan yang berguna bagi rumah sakit. Penelitian ini bertujuan mengolah tumpukan data rekam medis untuk mengklasifikasikan jenis penyakit yang terjadi pada rumah sakit berdasarkan pada kode penyakit internasional (International Classification of Disease) ICD-10. Metode yang digunakan dalam penelitian ini adalah metode algoritma C4.5 dengan menggunakan atribut kode penyakit internasional sebagai atribut label tujuan sebanyak 21 kelompok penyakit internasional, yaitu: A00-B99 sampai dengan Z00-Z99. Penilitian ini menghasilkan pohon keputusan dari 21 berupa value atribut tujuan kode penyakit, algoritma C4.5 dapat merepresentasikan sebanyak 14 value atribut tujuan kode penyakit dan persentasi data yang terbaca lebih dari 66%. Kesimpulan dari penelitian ini yaitu algoritma C4.5 membantu mengklasifikasikan kode penyakit internasional berdasarkan ICD-10 dan membentuk pohon keputusan yang dapat memberikan informasi penyakit apa saja yang sering terjadi pada rumah sakit [Fiandra Dkk, 2018:82]

Berdasarkan penelitian-penelitian diatas dapat disimpulkan bahwa data transaksi perusahaan maupun organisasi masih belum dimanfaatkan secara maksimal sehingga menghasilkan tumpukan data yang memakan memori penyimpanan. Jika dianalisa dengan baik kumpulan data tersebut dapat menghasilkan informasi yang bermanfaat. Salah satunya analisa menggunakan algoritma c.45 yang dapat menemukan informasi bermanfaat.