BAB II
TINJAUAN PUSTAKA

2.1 Data Mining

Data mining merupakan salah satu teknik dalam pengolahan data yang menemukan hubungan dari data yang tidak diketahui oleh pengguna serta menyajikan nya kedalam bentuk yang mudah dipahami sehingga dari hubungan data tersebut dapat dijadikan sebagai dasar dalam pengambilan keputusan (Ridwan et al., 2013). Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu : Deskripsi, Estimasi, Prediksi, Klasifikasi, Clustering, dan Asosiasi. (Muslehatin et al., 2017).

Definisi umum dari data mining itu sendiri adalah proses pencarian pola-pola yang tersembunyi (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam database, data werehouse, atau media penyimpanan informasi yang lain.

Data mining dilakukan dengan tool khusus, yang mengeksekusi operasi data mining yang telah didefinisikan berdasarkan model analisis. Data mining merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah data besar yang disimpan ketika menjalankan bisnis perusahaan. Kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa faktor antara lain: 1). Pertumbuhan yang cepat dalam kumpulan data. 2). Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan memiliki akses ke dalam database yang andal. 3). Adanya peningkataan akses data melalui navigasi web dan internet. 4). Tekanan kompetisi bisnis untuk meningkatkan

penguasaan pasar dalam globalisasi ekonomi. 5). Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi. 6). Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan (Rahmawati & Merlina, 2018)

Secara umum, metode data mining dapat dibagi menjadi dua :

deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakanuntuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi (Suyanto, 2017)

Metode yang ada dalam data mining adalah sebagai berikut :

1. *Classification*

   Klasifikasi merupakan proses untuk menemukan sekumpulan model yang dijelaskan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah objek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih. Sedangkan data uji digunakan untuk mengetahui tingkat akurasi dan model yang telah dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai dari suatu objek data.

2. *Clustering*

Pengelompokan data yang tidak diketahui label kelasnya kedalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripanya. Metode inilah yang digunakan dalam tugas akhir ini.

3. *Association*

Tujuan dari metode ini yaitu untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang terhubung kuat dengan yang lainnya.

4. *Regression*

Regression mirip dengan klasifikasi. Perbedaan utamanya adalah terletak pada atribut yang diproduksi nilai yang kontinyu.

5. *Forecasting*

Prediksi (forecasting) berfungsi untuk melakukan prediksi kejadian yang akan diproses berdasarkan data sejarah yang ada.

6. *Sequence Analiysis*

Tujuan dari metode ini adalah untuk mengenali pola dari data diskrit sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.

7. *Deviation Analiysis*

Tujuan dari metode ini adalah untuk menemukan penyebab perbedaan antara data yang satu dengan data yang lain dan biasa disebut sebagai outlier detection. Sebagai contoh adalah apakah sudah terjadi penipuan terhadap pengguna kartu kredit dengan melihat catatan transaksi yang tersimpan dalam basis data perusahaan tersebut.

1.1.1 Klasifikasi

Klasifikasi data merupakan suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan obyek di dalam sebuah basis data dan mengklasifikasikannya ke dalam kelas-kelas yang berbeda menurut model klasifikasi yang ditetapkan. Tujuan dari klasifikasi adalah untuk menemukan model dari data latih yang akan membedakan atribut ke dalam kategori atau kelas yang sesuai model.(Ente et al., 2020).

Untuk menggunakan metode klasifikasi tentunya harus menerapkan Algortima dalam Impelementasinya. Algortima yang akan digunakan adalah Decision Tree. Algortima C4.5 adalah ekstensi Quinlan untuk algoritma ID3 untuk menghasilkan pohon keputusan (Decision Tree), algoritma C4.5 rekursif mengunjungi setiap node keputusan, memilih split optimal sampai tidak ada perpecahan lanjut yang memungkinkan (Larose, 2005 dalam Novandya, 2017).

Klasifikasi adalah salah satu prediksi teknik data mining yang membuat prediksi tentang data nilai menggunakan hasil yang diketahui yang ditemukan dari kumpulan data yang berbeda. Masalah akurasi dari banyak algoritma klasifikasi adalah diketahui mengalami penurunan informasi saat dihadapi dengan data yang tidak seimbang, misalnya ketika distribusi sampel lintas kelas sangat miring (Misdram, 2021). Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, yang, misalnya, dapat dipartisi menjadi tiga kelas atau kategori: berpenghasilan tinggi, menengah pendapatan, dan pendapatan rendah. Model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi

tentang variabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnisdan penelitian meliputi: (Larose & Larose, 2014).

a. Menentukan apakah transaksi kartu kredit tertentu adalah penipuan
b. Menempatkan siswa baru pada jalur tertentu yang berkaitan dengan kebutuhan khusus
c. Menilai apakah aplikasi hipotek adalah risiko kredit yang baik atau buruk
d. Mendiagnosis apakah ada penyakit tertentu
e. Menentukan apakah surat wasiat ditulis oleh almarhum yang sebenarnya, atau curangoleh orang lain
f. Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkan kemungkinan ancaman teroris

Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, SupportVector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan (Wibawa, 2018).

1.1.2 Decision Tree C45

Pada dasarnya konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (rule). C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining. C4.5 memetakan nilai atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru (Xindong, 2009 dalam Novandya, 2017).

Berikut adalah rumus perhitungan entropy :

Menghitung Algoritma C4.5

Entrophy (S)= $\sum_{i=1}^{n} -pi\ log_2\ pi$

Keterangan :

S = Himpunan Kasus
n = Jumlah partisi S
p_i = probabilitas yang didapat dari jumlah kelas dibagi total kasus

Setelah menghitung nilai entropy dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan Information Gain. Untuk menghitung gain, yang bisa dihitung dengan formula sebagai berikut :

$$\textit{Gain (S,A)} = \textit{Entrophy (S)} = \sum_{i=1}^{n} * \textit{Entrophy}\ (Si)$$

Keterangan :
S = Himpunan kasus
A = Atribut
n = Jumlah atribut
|Si| = Jumlah partisi ke -i
|S| = jumlah kasus dalam S

Apabila ada atribut yang mempunyai banyak nilai atribut perlu untuk menghitung gain ratio, sebelumnya perlu kita ketahui suatu istilah baru yang disebut split information, yang bisa dihitung dengan formula sebagai berikut :

$$\textit{Split Info (S,A)} = \sum_{i=1}^{c} \frac{Si}{S} \log_2 \frac{Si}{S}$$

Keterangan :
S = ruang (data) sampel yang digunakan untuk training
A = atribut
S_i = jumlah sampel untuk atibut i

Dimana S_i sampai S_c adalah subset c yang dihasilkan dari pemecahan S dengan menggunakan atribut A yang mempunyai sebanyak c nilai. Selanjutnya gain ratio dihitung dengan cara :

$$\textit{Gain Ratio (S,A)} = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

1.1.3 Naïve Byes

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan ”naive” dimana diasumsikan kondisi antar atribut saling bebas. Pada sebuah dataset, setiap baris/dokumen diasumsikan sebagai vector dari nilai-nilai atribut dimana tiap nilai-nilai menjadi peninjauan atribut(Naafian et al., 2016) . Pendekatan dari Teorema Naïve Bayes adalah sebagai berikut :

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Keterangan : ? P(Ci|X) : peluang dokumen X pada kategori Ci.

- P(X|Ci) : peluang pada kategori Ci, dimana kata pada dokumen X muncul pada kategori tersebut.

- P(Ci) : peluang dari kategori yang diberikan, dibandingkan dengan kategori-kategori lainnya yang dianalisa.

- P(X) : peluang dari dokumen tersebut secara spesifik. Pada pengembangannya, -- P(X) dapat dihilangkan karena nilainya tetap, sehingga saat dibandingkan dengan tiap kategori, nilai ini dapat dihapus.

2.2 Penelitian Terkait

Berikut ini adalah Penelitian yang berkaitan dengan Klasifikasi Resiko Penyakit Obesitas

Table 1.1 Penelitian terkait

No	Judul, Penulis, Tahun	Dataset	Metode	Hasil	Kekurangan	Kelebihan
1	Klasifikasi Risiko Penyakit pada Ibu Hamil menggunakan Metode Modified K-Nearest Neighbor (MKNN) Yogi Pinanda1, Wayan Firdaus Mahmudy2, Edy Santoso3	Data sampling 102 data	Metode Modified K-Nearest	Berdasarkan jumlah pengujian terhadap jumlah data latih, nilai akurasi tertinggi adalah 85% dengan 82 data latih dan 20 data uji, kemudian terendah dengan 40 data latih dan 20 data uji dengan akurasi 65%.	Data yang sedikit mempengaruhi hasil dari akurasi jadi tidak memungkinkan menggunakan metode ini dengan data yang sedikit.	Kesimpulannya adalah semakin banyak data latih maka semakin besar nilai akurasinya karena semakin banyak data yang harus diperiksa dalam proses klasifikasi.
2	Sistem Informasi Posyandu Ibu Hamil dengan Penerapan Klasifikasi Resiko Kehamilan Menggunakan Metode Naïve Bayes	Dataset yang digunakan untuk klasifikasi dalam sistem ini Dinas Kesehatan Kabupat	Metode Naïve Bayes	tingkat akurasi ketika menggunakan 17 atribut didapatkan 53.913%, 19 atribut didapatkan 54.348%, , 21 atribut	Akurasi yang masih sangat kecil dalam pengambilan keputusan	semakin banyak atribut yang digunakan maka akurasi klasifikasi benar akan semakin tinggi dan metode naïve bayes memiliki kecenderungan

	(Implementing Qomariyatul Hasanah, Anang Andrianto, Muhammad Arief Hidayat	en Jember dan Puskesmas Mangli sebanyak 230		didapatkan 54.783%, dan 22 atribut didapatkan 56.957%		akurasi yang lebih tinggi jika
3	Sistem Prototype Klasifikasi Risiko Kehamilan Dengan Algo-ritma k-Nearest Neighbor (k-NN) Atma Deharja1* , Maya Weka Santi2 , Muhammad Yunus3 , Ervina Rachmawati4	data Kohort ibu hamil yang diukur ber-dasarkan standar Kartu Skor Poedji Rochjati (KSPR)	algoritma k-NN	Data KSPR dengan tingkat akurasi mencapai 80%		Hasil uji coba menunjukkan bahwa prototype sistem mampu melakukan klasifikasi dengan tepat dengan membandingkan hasil sistem menggunakan algoritma k-NN

2.3 Kehamilan

Kehamilan adalah suatu proses pembuahan dalam rangka melanjutkan yang terjadi secara alami menghasilkan janin yang tumbuh di rahim ibu. Kehamilan adalah sebuah proses yang dimulai dari tahap konsepsi sampai lahirnya janin. (Depkes RI (2016))

2.4 Akurasi

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal,akurasi adalah sebagian kecil dari prediksi model kami yang benar.

Secara formal, akurasi memiliki definisi sebagai berikut :

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \quad (2)$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Dimana

TP =

True

Positif TN = True Negatif

FP = False Positif

FN = False Negatif