BAB II

TINJAUAN PUSTAKA

II.1 Data Mining

Menurut (Suntoro, 2019) data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Data mining adalah suatu 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. Fungsi Data mining adalah mengidentifikasi fakta- fakta atau kesimpulan-kesimpulan yang di sarankan berdasarkan penyaringan melalui data untuk menjelajahi pola-pola atau anomali-anomali data. Data Mining mempunyai 5 fungsi :

- Classification, yaitu menyimpulkan definisi-definisi karakteristik sebuah grup. Contoh: pelanggan-pelanggan perusahaan yang telah berpindah kesaingan perusahaan yang lain.
- Clustering, yaitu mengindentifikasikan kelompok-kelompok dari barang-barang atau produk-produk yang mempunyai karakteristik khusus(clustering berbeda dengan classification, dimana pada clustering tidak terdapat definisi-definisi karakteristik awak yang di berikan pada waktu classification.)

- Association, yaitu mengidentifikasikan hubungan antara kejadian-kejadian yang terjadi pada suatu waktu, seperti isi-isi dari keranjang belanja.
- Sequencing Hampir sama dengan association, sequencing mengidentifikasikan hubungan-hubungan yang berbeda pada suatu periode waktu tertentu, seperti pelanggan-pelanggan yang mengunjungi supermarket secara berulang-ulang.
- Forecasting memperkirakan nilai pada masa yang akan datang berdasarkan pola-pola dengan sekumpulan data yang besar, seperti peramalan permintaan pasar.

Tujuan Data Mining adalah sebagai berikut :

- **Explanatory** adalah Untuk menjelaskan beberapa kondisi penelitian, seperti mengapa penjualan truk pick up meningkat di colorado.
- **Confirmatory** Untuk mempertegas hipotesis, seperti halnya 2 kali pendapatan keluarga lebih suka di pakai untuk membeli peralatan keluarga, di bandingkan dengan satu kali pendapatan keluarga.
- **Exploratory** Menganalisis data untuk hubungan yang baru yang tidak di harapkan, seperti halnya pola apa yang cocok untuk kasus penggelapan kartu kredit.

Komponen-komponen dari rencana data mining adalah sebagai berikut.

1. Analisa Masalah (Analyzing the Problem) Data asal atau data sumber harus bisa ditaksir untuk dilihat apakah data tersebut memenuhi kriteria data mining.

2. Mengekstrak dan Membersihkan Data (Extracting dan Cleansing The Data) Data pertama kli diekstrak dari data aslinya, seperti dari OLTP basis data,

text file, Microsoft Acces Database, dan bahkan dari spreadsheet, lalu data tersebut diletakan dalam data warehouse yang mempunyai sruktur yang sesuai dengan data model secara khas.

3. Validitas Data (Validating the Data) Sekali data telah diekstrak dan dibersihkan, ini adalh latihan yang bagus untuk menelusuri model yang telah kita ciptakan untuk memastikan bahwa semua data yang ada adalah data sekarang dan tetap.

4. Membuat dan Melatih Model (Creatig and Training the Model) Ketika algoritma diterapkan pada model, struktur telah dibangun.

5. Query Data dari Model Data Mining (Querying the Model Data) Ketika model yang telah cocok diciptakan dan dibangun, data yang telah dibuat tersedia untuk mendukung keputusan.

6. Evaluasi Validitas dari Mining Model (Maintaining the Validity of the Data Mining Model) Setelah moddel data mining terkumpul, lewat bebrapa waktu, karakteristik data awal seperti granularitas dan validitas mungkin berubah. Karena model data mining dapat terus berubah seiring perkembangan waktu.

II.2 Klasifikasi Data Mining

Klasifikasi adalah proses menemukan model yang menggambarkan dan membedakan kelas atau konsep data. Klasifikasi adalah salah satu teknik yang paling berguna dan penting karena mampu menangani data dengan jumlah skala besar (Gorade et al., 2017). Tujuan klasifikasi adalah untuk memprediksi dengan tepat kelas target untuk setiap kasus dalam data (Menaka & Kesavaraj, 2019). Model klasifikasi adalah salah satu domain tujuan umum dari data mining. Model

klasifikasi digunakan untuk mengklasifikasikan data yang baru tersedia menjadi label kelas(Menaka & Kesavaraj, 2019).

II.3 Algoritma Naïve Bayes

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Menurut Olson Delen (2008) menjelaskan Naïve Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari ” master ” tabel keputusan. Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variable independent, maka hanya varians dari suatu variable dalam sebuah kelas yg dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

Kegunaan Naïve Bayes

- Mengklasifikasikan dokumen teks seperti teks berita ataupun teks akademis
- Sebagai metode machine learning yang menggunakan probabilitas
- Untuk membuat diagnosis medis secara otomatis
- Mendeteksi atau menyaring spam

Kelebihan Naïve Bayes

- Bisa dipakai untuk data kuantitatif maupun kualitatif
- Tidak memerlukan jumlah data yang banyak
- Tidak perlu melakukan data training yang banyak
- Jika ada nilai yang hilang, maka bisa diabaikan dalam perhitungan.
- Perhitungannya cepat dan efisien
- Mudah dipahami
- Mudah dibuat
- Pengklasifikasian dokumen bisa dipersonalisasi, disesuaikan dengan kebutuhan setiap orang
- Jika digunakan dalaam bahasa pemrograman, *code*-nya sederhana
- Bisa digunakan untuk klasifikasi masalah biner ataupun *multiclass*

Kekurangan Naïve Bayes

- Apabila probabilitas kondisionalnya bernilai nol, maka probabilitas prediksi juga akan bernilai nol
- Asumsi bahwa masing-masing variabel independen membuat berkurangnya akurasi, karena biasanya ada korelasi antara variabel yang satu dengan variabel yang lain.
- Keakuratannya tidak bisa diukur menggunakan satu probabilitas saja. Butuh bukti-bukti lain untuk membuktikannya.
- Untuk membuat keputusan, diperlukan pengetahuan awal atau pengetahuan mengenai masa sebelumnya. Keberhasilannya sangat bergantung pada

pengetahuan awal tersebut Banyak celah yang bisa mengurangi efektivitasnya Dirancang untuk mendeteksi kata-kata saja, tidak bisa berupa gambar.

Algoritma *Naive Bayes* adalah penggolong probabilitas sederhana yang menghitung sekumpulan probabilitas dengan menghitung frekuensi dan kombinasi nilai dalam satu set data yang diberikan (Paquin et al., 2015). NBC adalah metode klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan kelas. NBC telah terbukti memiliki akurasi tinggi dan waktu pemrosesan yang cepat ketika diterapkan pada database dengan data besar (Gerhana et al., 2019a). Metode ini hanya membutuhkan sejumlah kecil data pelatihan untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi.

Bentuk umum Algoritma *Naïve Bayes* dapat dilihat pada persamaan 2.1. (Gerhana et al., 2019a)

$$P(C|X) = \frac{P(x|c).P(c)}{P(X)} \quad \text{(II.1)}$$

dimana :

X : Data dengan class yang belum diketahui

C : Hipotesis data merupakan suatu class spesifik

P(C|X) : Probabilitas hipotesis C berdasar kondisi X (posteriori probabilitas)

P(c) : Probabilitas hipotesis c (prior probabilitas)

P(x|c) : Probabilitas X berdasarkan kondisi pada hipotesis C

P(x) : Probabilitas X (Prdiktor prior probabilitas)

II.4 Algoritma C4.5

Algoritma C4.5 adalah **algoritma** yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe diskret dari record yang baru. Pohon keputusan telah menjadi metode yang terkenal dalam penambangan data. Meningkatnya penggunaan metode tersebut karena sejumlah manfaat antara lain, adalah kenyataan bahwa itu mudah dipahami dan ditafsirkan (Kartiwi et al., 2018). Decision tree juga membutuhkan upaya minimal persiapan data, mampu menangani data numerik dan kategorikal, dan mereka berkinerja sangat baik dengan set data besar dalam waktu yang cukup singkat (Kartiwi et al., 2018).

Pohon keputusan dimulai dengan simpul akar yang digunakan oleh pengguna untuk mengambil tindakan. Dari simpul akar ini, pengguna memecahnya sesuai dengan algoritma pohon keputusan. Hasil akhirnya adalah pohon keputusan dengan masing-masing cabang menunjukkan skenario yang memungkinkan dari keputusan yang diambil. Metode pohon keputusan memiliki beberapa keunggulan kemudian metode lain untuk *database* besar yang memiliki waktu pemrosesan yang relatif cepat, dapat dikonversi menjadi aturan klasifikasi dengan mudah dan sederhana (Gerhana et al., 2019a).

Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain. Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk

membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain (Berry dan Linoff, 2004). Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuannya. Sebuah pohon keputusan mungkin dibangun dengan seksama secara manual atau dapat tumbuh secara otomatis dengan menerapkan salah satu atau beberapa algoritma pohon keputusan untuk memodelkan himpunan data yang belum terklasifikasi.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

a. Pilih atribut sebagai akar
b. Buat cabang untuk masing-masing nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

II.5 Gain

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

$$Gain(S,A)=Entropy(S)-\sum_{i=1}^{n}\frac{|Si|}{|S|}*Entropy(Si) \quad \text{.....(II.2)}$$

Keterangan

- **S** : Himpunan Kasus
- **A** : Atribut
- **n** : Jumlah partisi atribut A
- $|S_i|$: Jumlah Kasus pada partisi ke-i
- **|S|** : Jumlah Kasus dalam S

II.6 Entropy

Sedangkan perhitungan nilai entropy dapat dilihat pada persamaan berikut ini :

$$Entropy(S)=\sum_{i=1}^{n}(-pi)*\log2(pi) \quad \text{.....(II.3)}$$

Keterangan

- **S** : Himpunan Kasus
- **n** : Jumlah partisi atribut S
- p_i : Proporsi dari S_i terhadap S

Pada tahap pembelajaran algoritma C4.5 memiliki 2 prinsip kerja yaitu:

1. Pembuatan **pohon keputusan**. Tujuan dari algoritma penginduksi pohon keputusan adalah mengkontruksi struktur data pohon yang dapat digunakan untuk memprediksi kelas dari sebuah kasus atau *record baru* yang belum memiliki kelas. C4.5 melakukan konstruksi pohon keputusan dengan metode *divide and conquer*. Pada awalnya hanya dibuat *node* akar dengan menerapkan algoritma *divide and conquer*. Algoritma ini memilih pemecahan kasus-kasus yang

terbaik dengan menghitung dan membandingkan *gain ratio*, kemudian *node-node* yang terbentuk di level berikutnya, algoritma *divide and conquer* akan diterapkan lagi sampai terbentuk daun-daun.

2. Pembuatan **aturan-aturan** (*rule set*). Aturan-aturan yang terbentuk dari pohon keputusan akan membentuk suatu kondisi dalam bentuk *if-then.* Aturan-aturan ini didapat dengan cara menelusuri pohon keputusan dari akar sampai daun. Setiap node dan syarat percabangan akan membentuk suatu kondisi atau suatu *if*, sedangkan untuk nilai-nilai yang terdapat pada daun akan membentuk suatu hasil atau suatu *then.*

Dewasa ini seiring dengan perkembangan dan kemajuan di bidang teknologi, yang diantaranya adalah artificial intellegence, database, statistik, pemodelan matematika, pengolahan citra, rekayasa perangkat lunak, data mining, dan lain sebagainya. Pemanfaatan IT sendiri telah masuk dan digunakan dalam banyak bidang, salah satunya adalah bidang kesehatan. Kemudian mulailah berkembang sistem informasi rumah sakit yang dapat memudahkan masyarakat serta rumah sakit itu sendiri. Kemudian salah satu perkembangan dalam bidang IT saat ini yang dapat diterapkan dalam bidang kesehatan adalah Data Mining. Secara umum, penerapan IT ataupun data mining dalam bidang kesehatan berperan untuk mencegah kejadian medical error yang terbagi dalam 3 mekanisme besar, yaitu :

1. Pencegahan (Adverse Event)
2. Respon Cepat
3. Umpan Balik

Salah satu penerapan data mining yang paling dikenal pada saat ini adalah penggunaan data mining untuk sistem atau aplikasi yang dapat memprediksi atau memperhitungkan penyakit seseorang. Sebagai contohnya adalah Mode Data Mining sebagai prediksi Penyakit Diabetes Mellitus dengan Teknik Decision Tree. Decision Tree merupakan salah satu teknik atau metode yang sangat banyak digunakan dalam penerapan data mining. Decision Tree adalah teknik klasifikasi dan prediksi yang amat kuat serta terkenal. Decision Tree dapat mengubah fakta yang sangat besar dan rumit menjadi pohon keputusan yang mampu mempresentasikan dan menggambarkan aturan, sehingga dapat mudah dipahami. Kemudian dalam decision tree sendiri terdapat lagi algoritma yang dikenal dengan nama algoritma C4.5 sebagai algoritma yang mendukung sistem keputusan dalam data mining. Secara umum langkah-langkah algoritma C4.5 dalam arsitektur decision tree adalah sebagai berikut :

1. Pilih atribut sebagai akar.
2. Buat dan kembangkan cabang untuk tiap nilai yang ada.
3. Klasifikasi dan bagi kasus ke dalam beberapa kelompok cabang.
4. Ulang kembali proses untuk setiap langkah cabang, hingga semua kasus memiliki kelas dan klasifikasi yang sama.

Gambaran dalam penggunaan ini adalah, nantinya pengguna atau pasien akan dihadapkan dengan beberapa pertanyaan. Kemudian dari jawaban yang diberikan itulah akan didapatkan apakah pasien tersebut mengalami diabetes atau tidak, lebih kurang seperti itulah gambaran penerapan data mining dalam bidang kesehatan.

II.7 Diabetes Mellitus

Diabetes mellitus merupakan penyakit kronis yang disebabkan oleh gagalnya organ pankreas memproduksi jumlah hormon insulin secara memadai sehingga menyebabkan peningkatan kadar glukosa dalam darah. Diabetes Mellitus merupakan salah satu penyakit tidak menular dan merupakan salah satu masalah kesehatan masyarakat yang penting. Diabetes Mellitus adalah suatu penyakit metabolik yang diakibatkan oleh meningkatnya kadar glukosa atau gula darah. Gula darah sangat vital bagi kesehatan karena merupakan sumber energi yang penting bagi sel-sel dan jaringan. Penyakit ini dibagi menjadi beberapa tipe, yaitu:

- Diabetes tipe 1, di mana sistem daya tahan tubuh menyerang dan menghancurkan sel beta di pankreas yang memproduksi insulin.
- Diabetes tipe 2, di mana sel beta di pankreas tidak memproduksi insulin dalam jumlah yang cukup, atau sel-sel tubuh tidak menunjukkan respons terhadap insulin yang diproduksi.
- Diabetes gestasional, yakni diabetes yang terjadi saat kehamilan.

Jika tidak dikelola dengan baik, diabetes dapat menyebabkan terjadinya berbagai komplikasi, seperti penyakit jantung koroner, stroke, obesitas, serta gangguan pada mata, ginjal, dan saraf.Selain itu, diabetes juga dapat menyebabkan terjadinya fluktuasi kadar gula darah dalam tubuh. Hal ini dapat mengakibatkan penurunan (hipoglikemia) atau peningkatan kadar gula darah (hiperglikemia) secara tiba-tiba. Beberapa tanda dan gejala diabetes adalah sebagai berikut :

- Sering buang air kecil
- Rasa haus berlebihan
- Penurunan berat badan
- Kelaparan
- Kulit bermasalah
- Penyembuhan luka lambat
- Infeksi jamur
- Mudah letih
- Pandangan kabur
- Kesemutan atau mati rasa

Kriteria penegakkan diabetes mellitus adalah dengan melakukan pemeriksaan kadar gula darah kapiler vena. Dikatakan diabetes jika :

- Kadar glukosa darah puasa ≥ 126 mg/dl, atau
- Kadar glukosa darah 2 jam setelah makan ≥ 200 mg/dl, atau
- HbA1C ≥ 6,5%
- Terdapat gejala klasik dan GDS ≥ 200 mg/dL

Salah satu pemanfaatan data mining pada bidang kesehatan adalah penerapan data mining untuk melakukan klasifikasi mendiagnosa penyakit Diabetes Mellitus. Data pasien yang terkena Diabetes Mellitus dapat digunakan untuk menunjukkan gejala penyakit Diabetes Mellitus yang diderita pasien. Klasifikasi ini bertujuan untuk membentuk model pohon keputusan untuk memprediksi penyakit pasien dan melihat variabel yang paling mempengaruhi penyakit pasien dengan kategori Diabetes Mellitus. Objek penelitian ini ialah data

pasien, jenis kelamin, usia, kadar gula darah dan gejala-gejala Diabetes Mellitus.

II.8 *Waikato Environment for Knowledge Analysis (Weka)*

Tool yang digunakan untuk menentukan hasil akurasi dalam melakukan perbandingan terhadap kedua algoritma menggunakan tool weka v3.9. *Waikato Environment for Knowledge Analysis* (Weka) adalah seperangkat perangkat lunak pengetahuan mesin. Weka berisi kompilasi alat visualisasi dan algoritma untuk analisis data dan pemodelan prediktif, dengan antarmuka pengguna grafis (S.Menaka & G.Kesavara, 2019).

Berikut ini merupakan beberapa review dari penelitian terdahulu terkait keunggulan dari *tools Weka* dalam data mining, tampak pada tabel II.1 :

Tabel II.1 Penelitian Terkait Terhadap Weka

No	The Title Of Papers	Penulis	Publisher	*Kelebihan Weka*
1	Data Mining: WEKA Software (an Overview)	(Ibrahim & Shiba, 2019)	Journal of Pure & Applied Sciences; ISSN 2521-9200; Vol.18 No. 3 2019	Menurut Ibrahim & Shiba di tahun 2019 dalam penelitiannya menyatakan bahwa weka memiliki kelebihan diantaranya yaitu kemudahan dalam pengoperasian karena antarmuka menggunakan grafis; weka memiliki banyak kumpulan teknik preprocessing serta pemodelan data yang komprehensif; dan weka juga tersedia secara freeware.

No	The Title Of Papers	Penulis	Publisher	*Kelebihan Weka*
2	Perbandingan Kinerja Tool Data Mining Weka dan Rapid Miner Dalam Algoritma Klasifikasi	(Faid et al., 2019)	TEKNIKA; DOI: 10.34148/teknika. v8i1.95; Volume 8, Nomor 1, Juli 2019 ISSN: 2549-8037, EISSN: 2549-8045	Penelitian yang dilakukan oleh Faid et al, ditahun 2019, dimana dalam penelitiannya membandingkan antara tools data mining yakni Weka dan Rapid Miner dalam kasus klasifiasi, mendapatkan kesimpulan bahwa tingkat akurasi yang dihasilkan terhadap kedua tools yang lebih unggul adalah Tools Weka dari pada Rapid Miner.
3	*Performance Comparison of Tree based classifier using WEKA*	(Ghogare, 2019)	Journal Of The Gujarat Research Society; ISSN: 0374-8588 Volume 21 Issue 3	Pada penelitian yang dilakukan oleh Ghogare ditahun 2019 dapat diambil kesimpulan terkait keunggulan dari tools weka diantaranya Weka dirancang dalam bahasa pemrograman Java sehingga tools dapat berjalan di berbagai platform, Weka juga mendukung berbagai format extension file diantaranya .csv, .arff, .arff.gz, .names, .data, .json, .json.gz, dan sebagainya yang berjumlah sebnyak 13 jenis file extension.

No	The Title Of Papers	Penulis	Publisher	*Kelebihan Weka*
4	Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News	(Kalmegh, 2015)	IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.; ISSN 2348 – 7968	Keuntungan dari Weka menurut Kalmegh tidak jauh berbeda dengan penelitian terdahulu yang lainnya, dalam penelitian kalmegh menyebutkan bahwa Weka dapat diunduh secara gratis. Weka dapat dioperasionalkan dalam semua sistem operasi yakni Windows, Linux maupun Mac. Weka berisi banyak macam metode dan pemodelan data yang sangat komprehensif dan yang sangat membantu dalam pengelolaan datanya Userinterfacenya sangat mudah untuk digunakan.
5	REVIEW OF VARIOUS INTRUSION DETECTION METHODS FOR TRAINING DATA SETS;	(Khan & Lilhore, 2017)	Internasioanl Journal Of Modern Trends in Engineering and Research; ISSN: 2349-9745; DOI:10.21884/ IJ MTER.2016.3170 .FI71A	Berdasarkan penelitian yang dilakukan oleh Khan dan Lilhore di tahun 2017 menyatakan bahwa weka merupakan salah satu tools canggih untuk develop dalam teknik pembelajaran mesin; selain itu Skema pembelajaran mesin yang baru juga bisa dikembangkan di dalam tools ini; WEKA juga merupakan perangkat lunak open source yang dikeluarkan di bawah lisensi GNU General Public Lisensi
6	*A Study on WEKA Tool for Data Preprocessing, Classification and Clustering;*	(Singhal & Jena, 2013)	International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013	Menurut Singhal dan jena bahwa, tools weka merupakan tools data mining yang tersedia secara gratis (freeware), weka juga mendukung beberapa dari data mining seperti preprocessing data, pengelompokan, klasifikasi, regresi, visualisasi, dan pemilihan fitur.