BAB II

TINJAUAN PUSTAKA

2.1 Penyakit Jantung

Penyakit jantung merupakan penyakit yang sangat berbahaya karena salah satu penyakit yang mematikan, terutama penyakit jantung koroner yang jumlah penderitanya sangat banyak dibandingkan dengan jenis penyakit jantung lainnya. Penyakit jantung koroner terjadi akibat adanya penyumbatan pembuluh arteri oleh plak yang menghambat suplai oksigen dan nutrisi ke jantung [8]. Kemunculan plak yang berupa timbunan lemah atau kalsium melalui proses secara bertahap. Biasanya, diawali dengan kekakuan pembuluh darah atau biasa di sebut dengan aterosklerois, kemudian penyempitan pembuluh darah.

2.2 Gejala Penyakit Jantung

Penyakit jantung koroner atau arteri koroner dianggap sangat berbahaya karena dapat menimbulkan serangan jantung mendadak yang berujung kematian. Serangan jantung terjadi akibat terhambatnya aliran darah menuju jantung sehingga suplai oksigen dan nutrisi di otot jantung dan jaringan disekitar jantung berkurang.

Tidak seperti otot tubuh lainnya, otot jantung tidak memiliki kemampuan beregenerasi. Apabila terdapat saja kerusakan maka akan berakibat fatal bagi tubuh. Semaikin lama serangan jantung terjadi semakin banyak pula kerusakan pada jantung. Karena itu penting bagi kita untuk mengenali gejala-gejala dari penyakit jantung sehingga dapat memberikan pertolongan dengan segera. Gejala penuakit jantung koroner secaa umum tidak dikenali oleh orang awam. Mereka terkadang menyepelekan dan menganggapnya wajar. Penderita baru menyadari bahwa dirinya terkena penyakit jantung koroner ketika kondisinya sudah parah. Bahkan, tak jarang dari mereka pada akhirnya harus meregang nyawa karena keterlambatan penanganan [9].

2.3 Tipe Penyakit Jantung

Beberapa tipe dari penyakit jantung antara lain[10]:

a) Gagal jantung

Gagal jantung disebut juga gagal jantung kongestif, yang berarti jantung tidak memompa darah sebagaimana mestinya. Gagal jantung tidak berarti jantung berhenti bekerja, tetapi kebutuhan tubuh akan darah dan oksigen tidak terpenuhi.

b) Arrhythmia

Aritmia adalah detak jantung yang tidak normal, jantung berdetak terlalu lambat, terlalu cepat, atau tidak teratur. Ketika denyut jantung kurang dari 60 denyut permenit hal ini dinamakan bradiakardia. Takikardia adalah ketika denyut jantung lebih dari 100 denyut permenit. Aritmia dapat mempengaruhi kinerja jantung.

c) Masalah Katup Jantung

Beberapa jenis masalah katup jantung diantaranya adalah stenosis, regurgitasi, dan prolaps katup mitral. Stenosis adalah keadaan dimana katup jantung tidak cukup terbuka untuk memungkinkan darah mengalir sebagaimana mestinya. Regurgitasi terjadi ketika katup jantung tidak menutup dengan benar dan memungkinkan darah bocor. Prolaps katup mitral terjadi ketika katup leaflet menonjol atau prolaps kembali ke ruang atas.

2.4 Data Mining

Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu[11]. Kaka's data mining meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang

dilakukan oleh data mining melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. Secara khusus, koleksi metode yang dikenal sebagai 'data mining' menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi prediksi model [12]. Berdasarkan tugas dan tujuan analisis, proses data mining dapat dibagi menjadi dua kategori utama, Tergantung pada adanya target variabel dan metode belajar (learning) yaitu antara proses belajar yang diawasi (supervised) dan tanpa pengawasan (unsupervised) [13].

2.5 Klasifikasi.

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri dapat berupa aturan jika-maka (if-then), berupa pohon keputusan (decision tree), jaringan saraf tiruan (neural network).

Klasifikasi adalah urutan yang sangat penting dalam data komunitas pertambangan. 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[14]

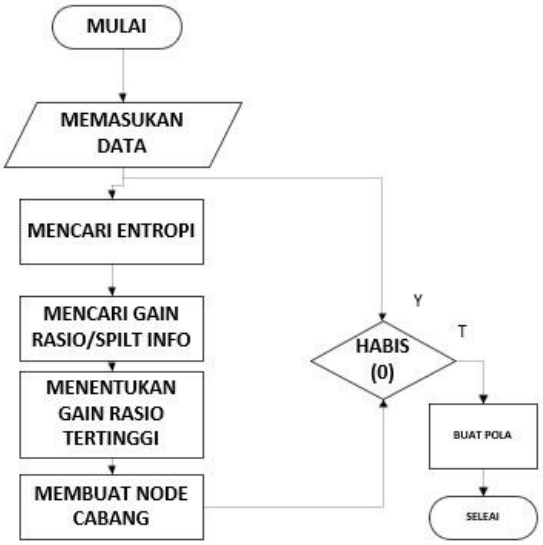
Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, yang, misalnya, dapat dipartisi menjadi tiga kelas atau kategori: berpenghasilan tinggi, menengah pendapatan, dan pendapatan rendah. Model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi tentang variabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnis dan penelitian meliputi:

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 curang oleh 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, Support Vector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan [15].

2.6 Algoritma C4.5

Algoritma C4.5 adalah sebuah algoritma yang berfungsi untuk membangun decision tree (pohon keputusan). Algoritma C4.5 dan pohon keputusan merupakan dua model yang tidak terpisahkan. Algoritma C4.5 adalah salah satu dari algoritma klasifikasi yang kuat dan banyak digunakan atau di implementasikan untuk pengklasifikasian dalam berbagai hal. Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan (1996) sebagai versi perbaikan dari algoritma Iterative Dichotomiser 3 (ID3). Serangkaian perbaikan dilakukan pada algoritma ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan simple yang berpengaruh untuk pembentukan pohon keputusan. Perbaikan tersebut meliputi metode untuk menangani data kontinew, mengatasi missing data, dan melakukan pemangkasan pohon [16].

Berikut adalah flowchart dari Algoritma C4.5 untuk membentuk sebuah pohon keputusan yang dapat dilihat pada Gambar 2.1



Gambar 2.1 Flowchart Pembangunan Sebuah Pohon Keputusan

Pada Gambar 2.1 memasukan data yang telah dimasukkan ke beberapa atribut, kemudian melakukan perhitungan nilai entropy dan gain untuk mendapat gain tertinggi. Nilai tersebut yang akan menjadi atribut akar atau root dari pohon keputusan. Kemudian dalam proses pembuatan node cabang untuk masing – masing nilai. Jika setiap kasus dalam cabang tersebut telah berada di dalam satu kelas yang sama maka proses perhitungan sudah selesai, tapi jika kasus berbeda kelas maka kembali ke perhitungan entropy dan begitu seterusnya hingga semua kasus berada di dalam kelas yang sama. Dalam memilih satu atribut menjadi akar, dilakukan perhitungan nilai dari atribut yang ada. Nilai gain yang paling tinggi dijadikan root di pohon keputusan. Untuk menghitung nilai gain rumus yang digunakan adalah [2] persamaan 2.1

$$Gaint\ (S.A\ ) = Entropy(S)\frac{Si}{S}*\text{ Entropy (Si)} \quad \textbf{(2.1)}$$

Keterangan:

S: Himpunan kasus

A: Data Atribut

n: Jumlah partisi di dalam atribut

|Si|: Jumlah kasus pada partisi ke-i

|S|: Jumlah kasus

Sedangkan untuk menghitung nilai entropy dapatt dihitung dengan rumus [17], persamaan 2.2

$$Entropy\ (S) = \sum_{pi}^{n} -\text{pi}=\log_2 \text{pi} \quad \textbf{(2.2)}$$

Keterangan:

S: Himpunan kasus

n: Jumlah partisi dalam atribut

pi: Proposi dari Si terhadap S

Gain

Gain adalah Ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan *entropy* dari variabel tersebut. *Gain* dapat dirumuskan dengan:

$$Gain\ (A) = \text{Entropy (S)} - entropy_A(\text{S}) \quad \textbf{(2.3)}$$

Pada algoritma C4.5, nilai *gain* digunakan untuk menentukan variabel mana yang menjadi *node* dari suatu pohon keputusan. Suatu variabel yang memiliki *gain* tertinggi akan dijadikan *node* di pohon keputusan.

Split Info

Split info digunakan sebagai pembagi dari *Gain(*A) yang akan menghasilkan *Gain Ratio*.

$$SplitInfo_A\,(D) = \sum_{J=1}^{V} \binom{DJ}{D} \text{Log}_2 \binom{DJ}{D} \quad \textbf{(2.4)}$$

Gain Ratio

Gain Ratio merupakan salah satu ukuran lain yang digunakan untuk mengatasi masalah pada atribut yang memiliki nilai sangat bervariasi. *Gain Ratio* tertinggi dipilih sebagai atribut test untuk simpul.

$$\text{GainRatio} \frac{Gain(A)}{SplitInfo_{A(D)}} \quad \textbf{(2.5)}$$

2.6.1 Algoritma *Naïve Bayes*

Algoritma *Naïve Bayes* merupakan metode yang dapat digunakan untuk mengklasifikasikan sekumpulan data. Algoritma ini memanfaatkan metode probabilitas dan Statistik yang dikemukanan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya[15].

Naïve Bayes merupakan *machine learning* yang menggunakan perhitungan probabilitas yang menggunakan konsep pendekatan *Bayesian.* Kata Naïve, yang terkesan merendahkan, berasal dari asumsi *independensi* pengaruh nilai suatu atribut dari probalilitas pada kelas yang diberikan terhadap nilai atribut lainnya Penggunaan teorema Bayes pada algoritma *Naïve Bayes* yaitu dengan mengkombinasikan prior probability dan probabilitas bersyarat dalam sebuah rumus yang bisa digunakan untuk menghitung probabilitas tiap klasifikasi yang mungkin.

Rumus *Naïve Bayes* nya adalah:

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad \textbf{2.6)}$$

Keterangan:

X = data dengan kelas yang belum diketahui

H = hipotesis data X, merupakan suatu kelas yang spesifik

P(H|X) = probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) = probabilitas hipotesis H (posteriori probability)

P(X|H) = probabilitas X berdasar kondisi H

P(X) = probabilitas dari X

atau

$$Posterior\ Probability = \frac{Prior\ Probability\ X\ likelihood}{evidance} \quad \textbf{(2.7)}$$

2.6.2 Forward Selection

Forward Selection merupakan salahsatu metode permodelan (pembangunan model linier) untuk menemukan kombinasi peubah yang "terbaik" dari suatu gugus peubah. Dalam prosedur forward Selection, sekalinya variabel masuk kedalam persamaan maka tidak bisa dihilangkan. Selain itu, forward selection dapat berarti memasukan variabel bebas yang memiliki korelasi yang paling erat dengan variabel tak bebasnya (variabel yang paling potensial untuk memiliki hubungan linier dengan Y)[8]

F-measure dihitung dengan menggabungkan precision dan recall menjadi satu nilai dengan menggunakan rumus berikut:

F-measure = 2 * (precision * recall) / (precision + recall)

precision adalah rasio dari jumlah prediksi yang benar dalam kelas positif terhadap jumlah total prediksi dalam kelas positif.

recall adalah rasio dari jumlah prediksi yang benar dalam kelas positif terhadap jumlah total item dalam kelas positif.

Nilai F-measure berkisar antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan kinerja yang lebih baik dari sistem klasifikasi.

2.6.3 Pemilihan Variable /Fitur

Pemilihan variabel yang juga disebut sebagai pemilihan atribut, digunakan pada *dataset* untuk menemukan pola yang penting dalam *data mining*. Pemilihan variabel digunakan untuk pengurangan dimensi pada *dataset*. Pemilihan variabel digunakan untuk melakukan eliminasi variabel yang tidak *relevan* dan *redundan*, yang dapat menyebabkan kebingungan dalam penggunaan variable.

Pemilihan variabel dapat mengurangi dimensi data, hal ini memungkinkan lebih efektif dalam operasi agar lebih cepat dari beberapa algoritma data mining. Dengan adanya pemilihan variabel membuat algoritma data mining lebih cepat dan lebih efektif.

Penggunaan pemilihan variabel pada *dataset* yang menggunakan variabel bebas dapat meningkatkan performa model. Pemilihan variabel juga merupakan proses yang cukup memakan biaya, dan juga berten.tangan dalam asumsi awal, bahwa semua informasi diperlukan untuk mencapai akurasi yang maksimal. Metode yang dapat digunakan untuk pemilihan variabel antara lain *Backward Elimination, Forward Selection*, *Genetic Algorithm*, dan yang lainnya. Metode-metode tersebut digunakan dalam penelitian *data mining* agar dapat menghasilkan variabel yang relevan dalam penelitian.

Pemilihan variabel dengan filter model ini lebih murah dalam komputasi karena tidak melibatkan induksi algoritma dalam prosesnya.

2.7 Tinjauan Studi

Berikut adalah ringkasan dari beberapa penelitian sebelumnya yang terkait dengan klasifikasi *data mining*.

Tabel 2.1 Ringkasan Tinjauan Studi

No	Judul	Penulis	Objek	Algoritma	Hasil
1	Model Data Mining sebagai	Ari Muzakir, Rika	Data penyakit Hipertensi	Decision Tree 4.5	Setelah dilakukan penelitian mendapatkan

	Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree	Anisa Wulandari (2016)	kehamilan pada RSIA YK Madira Palembang		decision tree dan rules yang dapat memprediksi penyakit hipertensi dalam kehamilan, dilakukan evaluasi dengan supplied test set menggunakan WEKA dihasilkan kesalahan (error) 7.3427% dan tingkat akurasi 92.6573%. Data training yang berjumlah 286 instances, hal ini menunjukkan bahwa terdapat 265 instances yang akurat dan

					21 instances yang error atau prediksinya salah
2	Klasifikasi penyakit diabetes mellitus tipe 2 Dengan metode algoritma c4.5	Prosiding Moh. Jasri (2017)	Data penyakit diabetus melitus pada Rumah Sakit Waluyojati Kraksan Probolinggo (RSML)	Decision Tree 4.5	Tingkat akurasi ratarata pada klasifikasi penyakit yang berkaitan dengan DM2 sebesar 90 %.
3	Klasifikasi Risiko Klasifikasi Penyakit Diabetes Mellitus dengan Menggunakan	Susanto (2018)	Data hasil laboratorium dan rekam medik dari pasien diabetes rumah sakit BP Batam	Decision Tree 4.5	1. Hasil klasifikasi data mining bahwa algoritma C4.5 dapat digunakan untuk mengklasifika si penyakit diabetes mellitus menjadi

	Algoritma Decision Tree C4.5				diabetes 1, diabetes 2 atau normal. 2. Dari metode klasifikasi data mining ini dengan algoritma C4.5 dan pengaplikasia n pohon keputusan yang membentuk aturan tersebut terdapat akurasi pada data training yang berjumlah 80 dari 100 data pasien sebesar 100% sedangkan akurasi pada data testing yang berjumlah 20 dari 100 data pasien sebesar 100%. Perhitungan

					keduanya dengan menggunakan confusion matrix

Berdasarkan hasil penelitian-penelitian terdahulu dapat disimpulkan bahwa penerapan algoritma klasifikasi C4.5 dan Algotitma Naïve Bayes dapat diimplementasikan pada berbagai bentuk keperluan klasifikasi, selain itu tingkat *accuracy* penerapan algoritma ini dapat mencapai lebih dari 80%. Dengan demikian, hasil penelitian ini bahwa penerapan algoritma C4.5 dan Algotitma Naïve Bayes tingkat *accuracy 80%* ini diharapkan dapat digunakan secara optimal dalam memprediksi Penderita Penyakit Gagal Jantung.

2.8 Confusion Matrix

Pengujian dengan Confusion Matrix Pada tahap ini pengujian model penelitian dilakukan dengan metode Confusion Matrix yang mempresentasikan hasil evaluasi model dengan menggunakan tabel matrik, Jika dataset terdiri dari 2 kelas, kelas pertama dianggap positif dan kelas kedua dianggap negatif. Evaluasi menggunakan confusion matrix menghasilkan nilai Akurasi, Precision, Recall, serta F-Measure. Akurasi dalam klasifikasi merupakan presentasi ketepatan record data diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. Precision merupakan proposikasi yang diprediksi positif yang juga positif benar pada data sebenarnya. Recall merupakan proporsi kasus positif yang sebenarnya diprediksi positif secara benar. True Positive (TP) merupakan jumlah record postif dalam dataset yang diklasifikasikan positif. True Negative (TN) merupakan jumlah record negative dalam dataset yang diklasifikasikan positif.

False Positive (FP) merupakan jumlah record negatif dalam dataset yang diklasifikasikan positif. False Negative (FN) merupakan jumlah record positif dalam dataset yang diklasifikasikan negatif. Berikut adalah persamaan model Confusion Matrix. [13]

Accuracy adalah jumlah perbandingan data yang benar dengan jumlah keseluruhan data.

$$\text{Akurasi} = \frac{(\text{TN}+\text{TP})}{(\text{TN}+\text{FN}+\text{FP}+\text{TP})} \quad \textbf{(2.9)}$$

Precision digunakan untuk mengukur seberapa besar proporsi dari kelas data positif yang berhasil diprediksi dengan benar dari keseluruhan hasil prediksi kelas positif.

$$precision = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad \textbf{(2.10)}$$

Recall digunakan untuk menunjukkan presentase kelas data positif yang berhasil diprediksi benar dari keseluruhan data kelas positif.

$$Recall = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad \textbf{(2.11)}$$