BAB II
LANDASAN TEORI

2.1 *DATA MINING*

Data Mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak terduga dan meringkas data dengan berbagai cara dengan cara yang berbeda dari sebelumnya, yang dapat dipahami dan berguna bagi pemilik data [4].

Data Mining adalah proses menemukan pola dan pengetahuan yang menarik dari sejumlah besar data. Sumber data dapat mencakup *database*, *data warehouses*, *Website*, repositori informasi lainnya, atau data yang dimasukkan secara dinamis ke dalam sistem [5].

Data mining atau bisa juga disebut dengan *Knowledge Discovery in Database* (KDD) adalah suatu kegiatan yang berkaitan dengan pengumpulan data, menggunakan data historis untuk menemukan pengetahuan, informasi, pola, atau hubungan dalam data yang berukuran besar atau banyak. Keluaran yang dihasilkan dari *data mining* dapat digunakan sebagai alternatif dalam pengambilan keputusan atau untuk memperbaiki pengambilan keputusan di masa yang akan datang [6].

Data mining didefinisikan sebagai proses menemukan pola dalam data. Proses ini harus otomatis atau (biasanya) semi otomatis. Pola yang ditemukan harus bermakna karena mengarah pada beberapa keuntungan, biasanya ekonomi. Data selalu hadir dalam jumlah yang substansial [7].

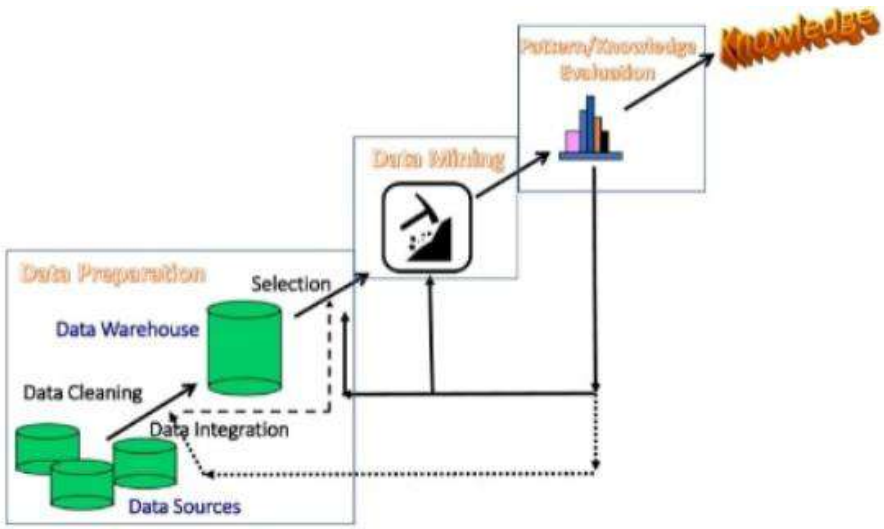
Jadi, dari beberapa pendapat ahli di atas dapat disimpulkan bahwa *data mining* adalah teknik yang digunakan untuk mengeksplorasi data yang ada dan menemukan hubungan yang sampai sekarang tidak diketahui dari sejumlah data besar atau banyak untuk menghasilkan informasi yang penting dan dapat digunakan untuk tujuan tertentu.

Pengolahan data menjadi informasi/pola/pengetahuan yang berguna dibutuhkan peranan *data mining* di dalamnya. Secara umum terdapat 5 (lima) peranan dalam *data mining*, yaitu estimasi, prediksi, klasifikasi, klustering, dan asosiasi. Tipe data yang digunakan pada *data mining* secara sederhana dibedakan menjadi 3 (tiga), yaitu tipe data numerik, tipe data kategorial, dan tipe data rentang waktu. Tipe data numerik dibagi menjadi dua bagian yaitu ratio dan interval. Tipe data kategorial juga dibagi menjadi dua bagian yaitu ordinal dan nominal [8].

Proses pengolahan data pada *data mining* diperlukan algoritme-algoritme buat melakukan ekstraksi sebagai informasi/pola/pengetahuan.. Penggunaan algoritme pada *data mining* diklasifikasi berdasarkan masing-masing peranan *data mining*. Pada peranan estimasi dan prediksi, algoritme yang banyak digunakan adalah Linear Regression, Support Vector Machine, Neural Network, dll. Algoritme yang banyak digunakan pada peranan klasifikasi adalah k-Nearest Neighbos (k-NN), *Naïve Bayes* , ID3, C4.5, CART, dll. Peranan klustering digunakan algoritme K-*Mean*s, Fuzzy C-*Mean*s, K-Medoid, Self-Organization Map (SOM), dll. Sedangkan pada peranan asosiasi digunakan algoritme FP-Growth, A Priori, Chi Square, Coefficient of Correlation, dll [8].

2.1.1 Tahapan Proses Dalam *Data mining*

Tahapan proses *data mining* ada beberapa yang sesuai dengan proses *Knowledge Discovery in Database* (KDD) sebagaimana seperti yang digambarkan pada Gambar 2.1 [5], [9] :



Gambar 2. 1 Proses *knowledge discovery* [5]

1. *Cleaning and Integration*

   a. Data Cleaning (Pembersihan Data)

Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan *noise* pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari *database* suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses *cleaning*. Proses *cleaning* dapat berpengaruh terhadap performa dari teknik *data mining*.

b. Data Integration (Integrasi data)

Integrasi data merupakan proses penggabungan data dari berbagai *database* sehingga menjadi satu *database* baru. Data yang diperlukan pada proses *data mining* tidak hanya berasal dari satu *database* tetapi juga dapat berasal dari beberapa *database*.

2. *Selection and Transformation*

a. *Data Selection* (Seleksi Data)

Tidak semua data yang terdapat dalam *database* akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari *database*. Misalnya pada sebuah kasus *market basket analysis* yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

b. *Data Transformation* (Transformasi Data)

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu. *Data mining* membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima *input* data yang bersifat kategorikal. Karenanya data yang berupa angka numerik apabila mempunyai sifat kontinyu perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut dengan transformasi data.

3. *Proses Mining*

Proses *mining* dapat disebut juga sebagai proses penambangan data. Proses *mining* merupakan proses utama yang menggunakan metode untuk menemukan pengetahuan berharga yang tersembunyi dari data.

4. *Evaluation and Precentation*

a. Evaluasi Pola (*Pattern Evaluation*)

Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam *knowledge based* yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternatif yang bisa diambil seperti menjadikannya umpan balik untuk memperbaiki proses *data mining*, atau mencoba metode *data mining* lain yang lebih sesuai.

b. Presentasi Pengetahuan (*Knowledge Presentation*)

Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses *data mining* adalah memformulasikan keputusan dari hasil analisis yang didapat.

2.1.2 Pengelompokan *Data mining*

Data mining dapat dibagi dalam beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu [4]:

1. Deskripsi

Terkadang peneliti dan analis hanya mencoba mencari cara untuk menggambarkan pola dan tren yang ada dalam data. Misalnya, lembaga survei mungkin menemukan bukti bahwa orang yang dipecat cenderung tidak mendukung calon presiden saat ini. Deskripsi pola dan tren sering kali memberikan kemungkinan penjelasan untuk pola dan tren tersebut. Misalnya, mereka yang diberhentikan sekarang lebih buruk secara finansial dari pada sebelum petahana terpilih dan cenderung memilih alternatif.

2. Estimasi

Estimasi hampir sama dengan klasifikasi, hanya saja variabel target yang diestimasi lebih bersifat numerik daripada kategoris. Model dibangun menggunakan record lengkap yang memberikan nilai variabel target sebagai nilai prediksi. Selain itu, pada pemindaian berikutnya, nilai estimasi variabel target didasarkan pada nilai variabel prediktor. Misalnya, kita mungkin tertarik untuk memperkirakan pembacaan tekanan darah sistolik pasien rumah sakit, berdasarkan usia pasien, jenis kelamin, indeks massa tubuh, dan kadar natrium darah. Hubungan antara tekanan darah sistolik dan variabel prediktor dalam set pelatihan akan memberi kita model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

3. Prediksi

Prediksi hampir seperti dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Contoh prediksi dalam bisnis dan penelitian adalah :

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

b. Prediksi tingkat penganguran lima tahun akan datang.

c. Predisksi persentase kanaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi juga bisa digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

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

5. Pengklasteran

Pengklasteran atau *clustering* adalah proses pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Pengklasteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklasteran. Pengklasteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklasteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul secara bersamaan. Dalam dunia bisnis sering disebut sebagai market basket analysis. Sehingga peranan *data mining* dalam hal ini adalah mencari aturan yang tidak tercakup untuk mendapatkan hubungan antara dua atau lebih atribut. *Association rule* adalah bentuk *if antacendant, then consequent* bersama-sama dengan suatu ukuran *support* dan *confidence*.

2.1.3 Jenis-Jenis Atribut

Atribut adalah suatu simbol yang menggambarkan identitas atau karakteristik objek. Sebagai contoh atribut yang menggambarkan objek pasien rumah sakit adalah nama, umur, golongan darah, dan tekanan darah. Berikut penjelasan dari empat macam atribut berdasarkan contohnya [9].

1. Atribut Nominal

Atribut nominal adalah nilai atribut yang diperoleh dengan cara kategorisasi karena nilainya menggambarkan kategori, kode, atau status yang tidak memiliki urutan. Misalnya, atribut golongan darah yang mempunyai empat kemungkinan nilai yaitu A, B, AB, dan O. Contoh lainnya seperti atribut jenis kelamin yang bisa bernilai pria dan wanita.

2. Atribut Ordinal

Atribut ordinal adalah atribut yang memiliki nilai dengan menggambarkan urutan atau peringkat. Namun, ukuran perbedaan antara dua nilai yang berurutan tidak diketahui. Atribut ordinal sangat berguna dalam survei, yaitu untuk penilaian subjektif (kualitatif) yang tidak dapat diukur secara objektif. Misalnya, kepuasan

pelanggan yang menghasilkan atribut bernilai ordinal, yaitu 0 (Tidak Puas), 1 (Cukup Puas), 2 (Puas), 3 (Sangat Puas).

3. Atribut Interval (Jarak)

Atribut interval adalah atribut numerik yang diperoleh dengan melakukan pengukuran, di mana jarak dua titik pada skala sudah diketahui dan tidak mempunyai titik nol yang absolut. Misalnya, suhu 0°C-100°C atau tanggal 1 sampai tanggal 31.

4. Atribut Rasio (Mutlak)

Atribut rasio adalah atribut numerik dengan titik nol absolut. Artinya, jika sistem pengukuran menggunakan rasio, dapat dihitung perkalian atau perbandingan antara suatu nilai dengan nilai yang lain. Misalnya, berat badan Doni 20 kg, berat badan Amanah 40 kg, berat badan Faiz 60 kg dan berat badan Udin 80 kg. Jika diukur dengan skala rasio maka berat badan Udin dua kali berat badan Amanah.

2.2 KLASIFIKASI

Klafikasi merupakan sebuah proses training (pembelajaran) suatu fungsi (target) yang digunakan untuk memetakan tiap himpunan atribut suastu objek ke satu dari label kelas tertentu yang di definisikan sebelumnya. Teknik klasifikasi ini cocok digunakan dalam mendeskripsikan data-set dengan tipe data dari suatu himpunan data yaitu biner atau nominal. Adapun kekurangan dari teknik ini yaitu biner atau nominal. Adapun kekurangan dari teknik ini yaitu tidak tepat untuk himpunan data ordinal karena pendekatan-pendekatan yang digunakan secara implisit dalam kategori data. Terdapat beberapa teknik klasifikasi yang digunakan

sebagai pemecahan kasus diantaranya yaitu Algoritma C4.5, Algoritma *K-Nearest Neighbor*, ID3, *Naïve Bayesian Clasification*, CART (*Clasification* And *Regression Tree*), dan lain-lain. Keluaran dari metode klasifikasi ini biasanya dalam bentuk "*Decision Tree* (pohon keputusan)" [10].

Klasifikasi adalah suatu teknik dengan melihat pada pola 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 beberapa aturan [11].

Klasifikasi adalah teknik yang digunakan untuk memprediksi kelas atau atribut dari setiap *instance* data. Model prediktif memprediksi nilai variabel yang tidak diketahui berdasarkan nilai variabel lain. Mengklasifikasikan pemetaan data ke dalam grup kelas yang telah ditentukan sebelumnya. Klasifikasi juga dikenal sebagai *supervised learning* karena kelas data telah ditentukan sebelumnya [12].

Jadi, dari beberapa pendapat ahli di atas dapat disimpulkan bahwa klasifikasi adalah metode pembelajaran terawasi yang mencoba menemukan hubungan antara atribut *input* dan atribut target. Klasifikasi juga merupakan teknik yang dapat digunakan untuk memprediksi atribut berdasarkan data contoh. Model prediktif memprediksi nilai variabel yang tidak diketahui berdasarkan nilai variabel lain. Mengklasifikasikan data pemetaan ke dalam kelompok kelas yang telah ditentukan.

2.3 ALGORITMA *NAÏVE BAYES*

Bayes adalah teknik prediksi probabilistik sederhana berdasarkan penerapan teorema *Bayes* atau aturan *Bayesian* dengan asumsi independensi kuat (*naïve*).

Dengan kata lain, *naïve bayes* adalah model fitur independen. Dalam *Bayes* (khususnya *naïve bayes*), arti independensi fitur adalah bahwa fitur dalam data tidak terkait dengan ada atau tidak adanya fitur lain dalam data yang sama. Prediksi *Bayesian* didasarkan pada teorema *Bayes* dengan rumus umum dengan persamaan berikut [9].

$$\mathrm{P}(H|\mathrm{E}) = \frac{P(\mathrm{E}|H).\mathrm{P}(H)}{\mathrm{P}(\mathrm{E})}$$

Keterangan persamaan diatas sebagai berikut:

E : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik

P(H|E) : Probabilitas bebas bersyarat (*conditional probability)* suatu hipotesis H jika diberikan bukti *(Evidence)* E terjadi.

P(E|H) : Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis H.

P(H) : Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun.

P(E) : Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

Ide dasar dari aturan *Bayes* adalah bahwa hasil dari hipotesis atas peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dalam aturan *Bayes* tersebut yaitu,

a) Sebuah probabilitas awal/prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
b) Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Algoritma naïve bayes atau yang biasa dikenal dengan *naïve bayes classifier* (NBC) merupakan salah satu algoritma dalam metode klasifikasi yang dapat memprediksi probabilitas atau probabilitas kepunyaan suatu kelas. NBC mengasumsikan bahwa nilai atribut suatu kelas adalah independen atau independen dari nilai atribut lainnya [13].

Metode *naïve bayes* adalah pengklasifikasi statistik yang dapat memprediksi kelas elemen probabilitas untuk pengklasifikasi *Bayesian* sederhana, atau disebut *naïve bayesian Classifie*r. Dapat diasumsikan bahwa efek dari nilai atribut dari kelas tertentu tidak tergantung pada atribut lainnya. Asumsi ini disebut independensi kelas bersyarat, dibuat untuk kemudahan perhitungan. Penafsiran ini dianggap "naive" [14].

Pengklasifikasi *naïve bayes* memiliki fungsi menghitung dan mencari nilai probabilitas tertinggi untuk mengklasifikasikan data eksperimen ke dalam kategori yang benar. Teknik prediksi probabilistik sederhana berdasarkan penerapan teorema *Bayes* atau aturan *Bayes* adalah teknik yang diimplementasikan dalam algoritma *naïve bayes* [10].

Jika X adalah vektor masukkan yang berisi fitur dan Y adalah label kelas, *naïve bayes* dituliskan dengan P(X|Y). Notasi tersebut berarti probabilitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas

akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan P(Y'|X') yang didapat.

Formulasi *naïve bayes* untuk klasifikasi yaitu pada persamaan berikut :

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

P(X|Y) adalah probabilitas data dengan vektor X pada kelas Y. P(Y) adalah probabilitas awal kelas Y. $\prod_{i=1}^{q} P(X_i|Y)$ adalah probabilitas independen kelas Y dari semua fitur dalam vetor X. Nilai P(X) selalu tepat sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian P(Y) $\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan Persamaan berikut.

$$P(X|Y = y) = \prod_{i=1}^{q} P(X_i|Y = y)$$

Setiap set fitur X=$\{x_1, x_2, x_3, \dots, x_q\}$ terdiri atas *q* atribut (*q* dimensi).

Umumnya, *Bayes* mudah dihitung untuk fitur bertipe ketegoris seperti pada kasus klasifikasi hewan dengan fitur "penutup kulit" dengan nilai {bulu, rambut, cangkang} atau kasus fitur "jenis kelamin" dengan nilai {pria, wanita}.

Namun untuk fitur dengan tipe numerik (kontinyu) ada perlakuan khusus sebelum dimasukan dalam *naïve bayes* . Caranya yaitu:

a) Melakukan diskretisasi pada setia fitur kontinyu dan mengganti nilai fitur kontinyu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasikan fitur kontinyu ke dalam fitur ordinal.

b) Mengasumsi bentuk tertentu dari distribusi probabilitas untuk fitur kontinyu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi Gaussian biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinyu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi Gaussian dikarakteristikkan dengan dua parameter, yaitu: *mean* (μ) dan standard deviasi σ .

Untuk setiap kelas Y_j, probabilitas bersyarat kelas Y_j untuk fitur X_i adalah seperti pada persamaan berikut .

$$P(X_i = x_i | Y_j = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Dimana dihitung terlebih dahulu nilai *mean* (μ) sesuai persamaan berikut :

$$\mu = \frac{\sum_i^n x_i}{n}$$

Dan standard deviasi σ sesuai persamaan berikut :

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{n-1}}$$

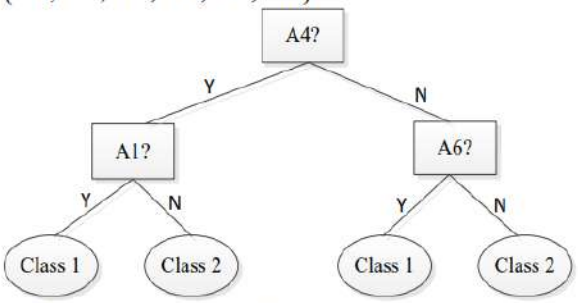
Parameter μ_{ij} bisa didapat dari *mean* sampel X_i ($\bar{x}$) dari semua data latih yang menjadi milik kelas Y_j, sedangkan σ_{ij}^2 dapat diperkirakan dari *varian* sampel (s^2) dari data latih.

Berdasarkan pemahaman di atas, dapat disimpulkan bahwa algoritma *naïve bayes* adalah salah satu metode algoritma yang sangat independen (naif) yang digunakan untuk memprediksi probabilitas masa depan melalui pengalaman masa lalu.

2.4 *GREEDY FORWARD SELECTION*

Greedy Forward Selection adalah salah satu pendekatan dalam algoritma *greedy* pada metode seleksi atribut atau seleksi fitur. Metode seleksi atribut *greedy* memiliki beberapa pendekatan seperti: *Greedy Forward Selection*, *Greedy backward elimination* dan kombinasi *Forward Selection* dan *backward elimination.* Mengenai metode seleksi fitur *greedy*, dimana *greedy* mencari atribut terbaik atau terburuk dari suatu data dengan menggunakan pengukuran *information gain* [5].

Algoritma *greedy* dapat disebut juga dengan *stepwise*, dan *RapidMiner* menggunakan algoritma ini dengan sebutan optimize selection [5]. Pada Gambar 2.2 dijelaskan mengenai metode seleksi fitur dengan *greedy* dimana *information gain* menjadi pengukur untuk mencari atribut yang terbaik maupun terburuk dapat mengunakan.

Forward Selection	Backward Elimination	Decision Tree *Induction*
Atribut asli: {A1, A2, A3, A4, A5, A6} Pengurangan atribut asli: {} =>{A1} =>{A1, A4} =>Pengurangan atribut: {A1, A4, A6}	Atribut asli: {A1, A2, A3, A4, A5, A6} =>{A1, A3, A4, A5, A6} =>{A1, A4, A5, A6} =>Pengurangan Atribut: {A1, A4, A6}	Atribut asli: {A1, A2, A3, A4, A5, A6}  =>Pengurangan Atribut: {A1, A3, A6}

Gambar 2. 2 Metode *Heuristic* (*Greedy*) untuk Seleksi Subset Atribut

Algoritma *greedy* dengan seleksi subset atribut sebagai berikut [5]:

1. Stepwise *Forward Selection*

Prosedur dimulai dengan himpunan kosong dari atribut sebagai set yang dikurangi, atribut yang terbaik dari atribut asli ditentukan dan ditambahkan pada set yang kurang. Pada setiap iterasi berikutnya yang terbaik dari atribut asli yang tersisa ditambahkan ke set.

2. Stepwise backward elimination

Prosedur dimulai dengan full set atribut. Pada setiap langkah, teknik ini dapat menghilangkan atribut terburuk yang tersisa di dataset.

Atribut yang terbaik dan terburuk dapat ditentukan dengan menggunakan tes signifikansi statistik yang berasumsi bahwa atribut tidak saling berhubungan satu sama lain (independen) [5].

Seleksi atribut menggunakan *information gain* adalah memilih *gain* tertinggi dan formula yang digunakan adalah sebagai berikut dimana langkah yang pertama adalah dengan mencari nilai *entropy* sebagai berikut:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

$Info(D)$ adalah untuk mengetahui nilai dari *entropy*, kemudian jika suatu atribut mempunyai nilai yang berbeda beda maka menggunakan formula sebagai berikut:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ Info(D_j)$$

Kemudian untuk mendapatkan hasil *gain*, dimana formulanya menjadi: $Gain(A) = Info(D) - Info_A(D)$

2.5 *TOOLS DATA MINING*

2.5.1 *RapidMiner*

RapidMiner adalah perangkat lunak sumber terbuka (*open source*). *RapidMiner* merupakan sebuah solusi untuk melakukan analisis terhadap *data mining*, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediktif untuk memberikan informasi kepada pengguna guna membantu mereka membuat keputusan terbaik. *RapidMiner* memiliki sekitar 500 operator *data mining*, termasuk operator untuk *input*, *output*, data preprocessing, dan visualisasi. *RapidMiner* adalah perangkat lunak yang berdiri sendiri untuk analisis data dan merupakan alat penambangan data yang dapat diintegrasikan ke

dalam produknya sendiri. *RapidMiner* ditulis dalam bahasa *Java*, sehingga dapat bekerja pada sistem operasi apa pun [11].

RapidMiner adalah platform perangkat lunak ilmu data yang dikembangkan oleh perusahaan dengan nama yang sama yang menyediakan lingkungan terpadu untuk pembelajaran mesin (*machine learning*), pembelajaran mendalam (*deep learning*), penambangan teks (*text mining*), dan analitik prediktif (*predictive analytics*). Aplikasi ini dapat digunakan dalam penerapan bisnis dan komersial serta penelitian, pendidikan, pelatihan, pembuatan prototipe cepat dan pengembangan aplikasi, dan mendukung semua fase proses pembelajaran mesin, termasuk persiapan, data, visualisasi hasil, validasi dan pengoptimalan. *RapidMiner* dikembangkan dengan model inti terbuka (*open core*) [15].

RapidMiner adalah perangkat lunak pengolah *data mining*. Pekerjaan yang dilakukan oleh *RapidMiner text mining* berkisar pada menganalisis teks, mengekstraksi pola dari kumpulan data besar, dan menggabungkannya dengan metode statistik, kecerdasan buatan, dan basis data [16].

Jadi, dari beberapa pendapat ahli di atas dapat disimpulkan bahwa *RapidMiner* adalah platform perangkat lunak ilmu data yang digunakan sebagai pengolah *data mining*. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediktif untuk memberikan informasi kepada pengguna guna membantu mereka membuat keputusan terbaik.

RapidMiner menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML (*Extensible Markup Language*) yang mendefenisikan proses analitis keingginan pengguna untuk

diterpkan ke data. File ini kemudian dibaca oleh *RapidMiner* untuk menjalankan analis secara otomatis.

RapidMiner memiliki beberapa sifat sebagai berikut [11]:

1. Ditulis dengan bahasa pemrograman *Java* sehingga dapat dijalankan di berbagai sistem operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator trees
3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
5. Konsep *multi-layer* untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki *GUI*, *comm*and *line mode*, dan *Java API* yang dapat dipanggil dari program lain.

Beberapa Fitur dari *RapidMiner*, antara lain:

1. Banyaknya algoritma *data mining*, seperti *Decision Tree*e dan self-organization map.
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.
3. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks.
4. Menyediakan prosedur *data mining* dan machine learning termasuk: ETL (*extraction, transformation, loading*), data *preprocessing*, visualisasi, *modelling* dan evaluasi

5. Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI
6. Mengintegrasikan proyek *data mining Weka* dan statistika R.

2.5.2 *Google Colaboratory*

Google Colaboratory atau bisa disebut *Colab*, adalah platform *machine learning* gratis yang disediakan oleh Google yang memungkinkan pengguna untuk membuat, menjalankan, dan berbagi kode pada lingkungan cloud tanpa memerlukan pengaturan atau konfigurasi perangkat lunak pada komputer lokal. Colab menyediakan banyak fitur untuk memudahkan para peneliti, pengembang, dan praktisi dalam menganalisis data, membuat model machine learning, serta melakukan eksperimen yang berkaitan dengan *artificial intelligence* (AI) dan *deep learning* [17].

Google Colaboratory merupakan sebuah layanan *cloud computing* gratis dari Google yang dirancang khusus untuk mendukung pengolahan data dan pelatihan model machine learning. Colab memberikan akses ke sumber daya cloud yang kuat seperti GPU dan TPU, serta lingkungan pengembangan berbasis web yang terintegrasi dengan Google Drive. Selain itu, fitur kolaborasi real-time pada Colab memungkinkan pengguna untuk berbagi kode dan hasil secara langsung dengan anggota tim, sehingga mempermudah proses pengembangan dan analisis data [18].

Colaboratory memiliki beberapa fitur yang membuatnya populer di kalangan para praktisi AI, yaitu:

1. Colab dapat diakses langsung melalui peramban web, tidak memerlukan instalasi apa pun di komputer lokal, dan tidak memerlukan akun Google Cloud Platform.
2. Colab terintegrasi dengan layanan penyimpanan awan Google Drive, sehingga pengguna dapat dengan mudah mengunggah, men-download, dan berbagi data.
3. Colab menyediakan lingkungan Python yang sama dengan Jupyter Notebook, sehingga pengguna dapat dengan mudah mengakses dan menggunakan berbagai paket Python yang tersedia.
4. Colab menyediakan akses ke GPU gratis untuk mempercepat proses pelatihan model machine learning.
5. Colab memiliki fitur kolaborasi real-time yang memungkinkan pengguna untuk berbagi kode dan hasil secara langsung dengan anggota tim.
6. Colab memungkinkan pengguna untuk menyesuaikan lingkungan kerja mereka dengan mudah dengan menginstal paket Python tambahan dan menyesuaikan pengaturan.

Pada penelitian kali ini penulis menggunakan *tools data mining RapidMiner* dan *Google Colaboratory* sebagai alat bantu yang digunakan untuk melakukan analisis terhadap *data mining* dan analisis prediksi untuk memberikan informasi yang berguna untuk membantu membuat keputusan terbaik.

2.6 GAGAL JANTUNG

Penyakit jantung adalah kondisi dimana jantung sebagai organ vital manusia mengalami gangguan sehingga menyebabkan tidak berfungsi dengan baik. Gejala yang menunjukkan masalah jantung, yaitu adanya penyempitan pembuluh darah di jantung, kelainan pada irama jantung, kelainan jantung yang didapatkan sejak lahir, terganggunya otot jantung, adanya infeksi oleh bakteri, virus, atau parasit, serta gangguan pada katup jantung bisa salah satunya atau keseluruhan. Faktor yang paling beresiko dari penyakit jantung adalah makanan yang tidak sehat, kurangnya aktifitas fisik, konsumsi tembakau dan alkohol [19].

Gagal jantung yaitu suatu kondisi di mana jantung tidak lagi mampu memompa suplai darah yang cukup terkait dengan aliran balik vena dan terkait dengan kebutuhan metabolisme jaringan tubuh pada saat itu. Semua bentuk penyakit jantung dapat menyebabkan dekompensasi dan kegagalan [20].

Gagal jantung merupakan salah satu alasan mengapa orang sering dirawat di rumah sakit, usia lanjut merupakan salah satu faktor yang menyebabkan pasien harus dirawat inap lagi dengan penyakit yang sama. Gagal jantung merupakan salah satu penyakit kronis dengan waktu pengobatan terlama di Indonesia. Pasien dengan gagal jantung berisiko tinggi untuk dirawat kembali di rumah sakit dan bahkan dapat kembali ke ruang gawat darurat dalam waktu 30 hari setelah keluar dari rumah sakit [21].

Cardiovascular diseases (CVDs) atau penyakit jantung merupakan salah satu penyakit paling mematikan di dunia dan penyebab utama kematian di seluruh dunia, dengan angka kematian total sekitar 17,9 juta orang setiap tahun [19]. Indonesia

merupakan negara yang perlu mendapat perhatian khusus terhadap penyakit jantung. Menurut data Riset Kesehatan Dasar (Riskesdas), kejadian penyakit jantung terus meningkat 1,5% per tahun.

2.7 PENELITIAN SEJENIS

Penulis memulai penelitian ini dengan terlebih dahulu melakukan studi kepustakaan berdasarkan kajian dan sumber terkait. Beberapa penelitian yang serupa telah dilakukan dengan menggunakan teknik *data mining* untuk mengungkap berbagai informasi.

Tabel 2. 1 Penelitian Sejenis

Peneliti dan Tahun	Masalah	Metode	Hasil Penelitian
Firza Novaldy, Asti Herliana, 2021 [22]	Mengingat tingginya angka penderita gagal jantung dan pentingnya organ vital seperti jantung, memprediksikan gagal jantung telah menjadi prioritas bagi dokter. Beberapa penelitian tentang penyakit gagal jantung telah dilakukan oleh peneliti sebelumnya namun penelitian yang dilakukan masih mendapatkan nilai akurasi yang belum sempurna. Untuk itu perlu dilakukan penelitian menggunakan algoritma lainnya agar bisa mendapatkan hasil akurasi terbaik.	*Naïve Bayes* , *Particle Swarm Optimization* (PSO)	Hasil dari perhitungan dengan algortima *naïve bayes* yang diterapkan pada Heart Failure Clinical Records Dataset mendapatkan nilai akurasi confusion matrix sebesar 75.00% dan AUC sebesar 0.847. Kemudian setelah diterapkan Particle Swarm Optimization untuk seleksi fitur pada dataset yang digunakan, Nilai akurasi meningkat menjadi 91.67% dan AUC sebesar 0.908. Menggunakan optimasi PSO pada algoritma *naïve bayes* diperoleh beberapa atribut-atribut yang berpengaruh terhadap bobot atribut.

Elin Nurlia, Ultach Enri, 2021 [23].	Gagal jantung merupakan penyakit yang terjadi karena kegagalan jantung dalam memompa darah dan menyebabkan tingkat kematian yang tinggi. Penelitian ini bertujuan untuk memprediksi kematian akibat penyakit gagal jantung menggunakan metode klasifikasi dengan algoritma C4.5 berbasis *Forward Selection*.	C4.5 berbasis *Forward Selection*	Pengujian algoritma C4.5 menghasilkan akurasi sebesar 77,89% dan nilai AUC 0,750 yang termasuk kategori fair classification, sedangkan algoritma C4.5 berbasis *Forward Selection* memperoleh akurasi sebesar 84,29% dan nilai AUC 0,785 yang termasuk kategori fair classification.
Duwi Cahya Putri Buani, 2021 [24].	Jantung merupakan organ penting yang dimiliki oleh manusia, sehingga jika terjadi masalah pada jantung otomatis akan terjadi masalah juga pada organ tubuh yang lain. Penelitian ini bermaksud melakukan prediksi gagal jantung menggunakan *naïve bayes* dan algoritma genetika sebagai seleksi fitur agar akurasi dari prediksi lebih optimal.	*Naïve Bayes* dan Algoritma Genetika	Hasil penelitian ini menunjukkan dimana dengan menggunakan data yang sama tanpa menggunakan algortima genetika akurasi *naïve bayes* hanya mencapai 69,60%, setelah dilakukan seleksi fitur dengan algoritma genetika hasil akurasi meningkat menjadi 96,67%.
Dede Andri Muhammad Reza, Amril Mutoi Siregar, Rahmat 2022 [25].	Mengingat berharganya organ vital seperti jantung, memprediksi gagal jantung telah menjadi prioritas bagi tenaga medis, tetapi hingga saat ini prediksi kejadian terkait gagal jantung dalam praktik klinis biasanya gagal mencapai akurasi yang tinggi. Penelitian ini	*K-Nearest Neighbord*	Hasil yang didapat pada peneltian ini yaitu menghitung manual dengan menggunakan nilai k=7 yang menghasilkan kategori peristiwa kematian data testing adalah **Tidak**. Pengujian pada *RapidMiner* dilakukan dengan menggunakan pergantian nilai k, akurasi tertinggi didapat

	menggunakan algoritma K-Nearest Neighbord yaitu merupakan algoritma klasifikasi berdasarkan kedekatan jarak suatu data dengan data yang lain.		pada nilai k=7 dengan nlai akurasi 94,92%. Kemudian pengujian pada bahasa pemrograman python menghasilkan nilai akurasi 68%.
Yuri Yulian, 2022 [26]	Gagal jantung yang merupakan masalah kesehatan yang global yang tidak hanya menimbulkan masalah fisik, dampak lain seperti psikologis, sosial dan ekonomi, meningkatkan tingkat rawat inap hingga kematian. Menggunakan machine learning untuk mempresiksi kelangsungan hidup pasien penderita gagal jantung agar dapat melakukan pencegahan dari awal.	Algoritma *random forest, random subspace, logitboost* dan Seleksi Fitur *Bestfirst.*	Hasil penelitian menggunakan aplikasi *Weka* dengan melakukukan seleksi fitur bestfirst serta metode class balancer untuk menangani class yang tidak balance dan perbandingan terhadap 3 algoritma yang menunjukan performa terbaik yaitu pada algoritma *random* forest dengan metode percentage split 80%, accuracy 91,45%, *mean* absolute error 0.1874, incorrectly classified instances 8.55%, precision 0.915, recall 0.914, AUC 0.953.

Berdasarkan penelitian sejenis diatas, yang membedakan penelitian ini dengan penelitian sebelumnya adalah sebagai berikut :

Pada penelitian yang dilakukan oleh Firza Novaldy, Asti Herliana, 2021 [22] memiliki perbedaan yaitu pada metode yang digunakan, penulis akan menerapkan algoritma *Greedy Forward Selection* untuk menyeleksi atribut. Sedangkan persamaan terletak pada meneliti mengenai prediksi penyakit gagal jatung menggunakan algortima *naïve bayes.*

Pada penelitian yang dilakukan oleh Elin Nurlia, Ultach Enri, 2021 [23] memiliki perbedaan yaitu pada metode yang digunakan, penulis akan menerapkan algoritma *naïve bayes* dan *Greedy Forward Selection*. Sedangkan persamaan terletak pada meneliti mengenai prediksi penyakit gagal jatung.

Pada penelitian yang dilakukan oleh Duwi Cahya Putri Buani, 2021 [24] memiliki perbedaan yaitu pada metode yang digunakan, penulis akan menerapkan algoritma *Greedy Forward Selection* untuk menyeleksi atribut. Sedangkan persamaan terletak pada meneliti mengenai prediksi penyakit gagal jatung menggunakan algortima *naïve bayes.*

Pada penelitian yang dilakukan oleh Dede Andri Muhammad Reza, Amril Mutoi Siregar, Rahmat 2022 [25] memiliki perbedaan yaitu pada metode yang digunakan, penulis akan menerapkan algoritma *Naïve Bayes* dan *Greedy Forward Selection*. Sedangkan persamaan terletak pada meneliti mengenai prediksi penyakit gagal jatung.

Pada penelitian yang dilakukan oleh Yuri Yulian, 2022 [26] memiliki perbedaan yaitu pada metode yang digunakan, penulis akan menerapkan algoritma *Naïve Bayes* dan *Greedy Forward Selection*. Sedangkan persamaan terletak pada meneliti mengenai prediksi penyakit gagal jatung.

Hasil penelitian ini diharapkan dapat menjadi acuan oleh peneliti kedepannya dalam memprediksi data penderita gagal jantung agar lebih akurat dan optimal.