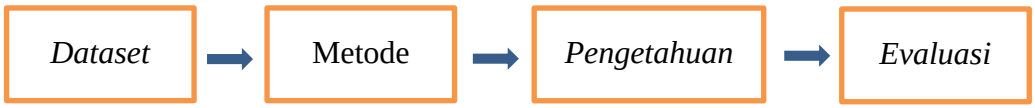
BAB II
TINJAUAN PUSTAKA

1.1 Data Mining

Data minig adalah seni dan ilmu untuk menemukan pola inovatif yang berguna dari data (Purwati, 2021). Apa yang kita ketahui bahwa dalam basis data merupakan sekumpulan data yang saling berhubungan, perti basis data pelanggan dengan data produk, data karyawan, data pelayanan, data penjualan, dan semua itu saling berhubungan.

Fungsi data mining sendiri proses untuk menemukan hubungan antara atribut. Konsep data mining sendiri melibatkan penerapan teknik analisis untuk mengidentifikasi pola tersembunyi, hubungan, dan wawasan berharga dari dataset yang besar, dengan tujuan mendukung pengambilan keputusan yang informasional dan strategis.

Proses *data mining* diantaranya ada himpunan data, metode *data mining*, pengetahuan (*knowledge*), dan evaluasi. Pemilihan algoritma yang tepat akan bergantung pada proses *Knowledge Discovery in Database* (KDD) secara keseluruhan. Bagan proses *data mining* dapat dilihat pada Gambar 2.1.



Gambar 2.1. Proses data mining 1

1.2 Teknik *Data Mining*

Teknik klasifikasi merupakan teknik pendekatan analitis yang memanfaatkan pola-pola yang ada dalam data untuk mengelompokkan atau mengkategorikan entitas berdasarkan atribut-atributnya.

2.4.1 *Classification* (Klasifikasi)

Klasifikasi dapat didefinisikan suatu proses yang elakukan pembelajaran terhadap fungsi target f yang memetakan setiap set fitur x ke dalam satu dari sejumlah label kelas y yang tersedia. Model klasifikasi dapat dilihat pada gambar

Gambar 1.2 model klasifikasi 1

Metode klasifikasi yang sering di gunuakan pada umumnya yaitu, *Suport Vektor Machine, Multilayer Perceptron, Naive bayes*, dll. Tujuan utamanya adalah untuk mengenali pola atau hubungan antara atribut-atribut yang ada dalam data sehingga dapat memprediksi kelas atau kategori yang sesuai untuk data baru yang belum pernah dilihat sebelumnya. Proses klasifikasi melibatkan pelatihan model menggunakan dataset yang sudah diklasifikasikan sebelumnya, dan kemudian menggunakan model ini untuk mengklasifikasikan data baru berdasarkan fitur-fiturnya. Dalam esensinya, klasifikasi adalah alat yang kuat dalam analisis data yang membantu dalam pengambilan keputusan berdasarkan informasi yang terkandung dalam pola-pola data yang teramati.

1. *Naive Bayes* untuk Klasifikasi

(Prasetyo, 2012) menjelaskan kaitan anatara *Naive Bayes* dengan klasifikasi korelasi hipotesis dan bukti klasisfikasi adalah hipotesis dalam teorema *bayes*k merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan beukti merupakan fitur-fitur yang menjasikan masukan dalam model klasifikasi. jika X adalah vektor masukan yang berisi fiturn Y adalah labe kelas, *Naive Bayes* dituliskan dengan P(X|Y). Notasi tersebut berarti probabilitas label kelas y didapatkan setelah fitur-fitur X diamati.

Selama proses pelatihan harus dilakukan pemeberajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi Xdan Y berdasarkan informasi yang didapat daari data latih. Dengan mebangun model tersebut, suatu data uji X’ dapat

diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan P(Y'|X') yang didapat. Formulasi *Naive Bayes* untuk klasivikasi yaitu pada persamaan berikut.

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

P(X|Y) adalah probabilitas data dengan vektor X pada kelas Y. P(Y) adalah probabilitas awal kelas Y. $\prod^{q}_{i=1}P(X_i|Y)$ adalah probabilitas independen kelas Y dari semua fitur dalam vektor X. Nilai P(X) selalu taepat sehingga dalalam perhitungan prediksi nantinya tinggal mengghitung bagian P(Y) $\prod^{q}_{i=1}P(X_i|Y)$ dengan memilih yang terbesar sebagian kelas yang dipilih sebagian hasil prediksi. Sementara probabilitas independen $\prod^{q}_{i=1}P(X_i|Y)$ tersebut merupak pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan pesamaan berikut.

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

Setiap set fitur X= {x_1, x_2, x_3……..x_q} terdiri atas q atribut (q dimensi). Umumnya, *bayes* mudah dihitung untuk fitur bertipe kategori seperti pada kasus klasisfikasi hewan dengan fitur "penutup kulit" dengan nilai {bulu, rambur, cangkang} atau kasus fitur "jenis kelamin" dengan nilai {pria, wanita}. Namun untuk fitur dengan tipe numerik

2. *Decision Tree* C4.5

Algoritma C4.5 adalah salah satu pendekatan populer dalam membangun pohon keputusan untuk masalah klasifikasi. Fokus utamanya adalah memilih atribut yang paling signifikan untuk membagi dataset menjadi subset yang lebih kecil. Setiap langkah algoritma didasarkan pada konsep informasi gain dan split information, yang membantu dalam pemilihan atribut yang optimal. Rumus-rumus yang terlibat dalam C4.5 adalah sebagai berikut:

Entropy (Entropi)Entropi mengukur tingkat ketidakpastian dalam dataset. Dalam konteks klasifikasi, entropi mengukur seberapa campur aduk kelas dalam dataset.

Rumus Entropi:

$$Entropy(S) = \sum_{i=1}^{n} - Pi * log2\ Pi \qquad (1)$$

Di mana:

n adalah jumlah kelas.

p_i adalah proporsi jumlah sampel dari kelas (*i*) terhadap total sampel.

Information Gain (Gain Informasi) :

Gain informasi mengukur pengurangan entropi setelah membagi dataset berdasarkan atribut tertentu. Semakin tinggi gain informasi, semakin baik atribut tersebut dalam membagi dataset.

Rumus Gain Informasi:

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

Dimana:

A adalah atribut yang dipertimbangkan untuk pembagian.

D adalah dataset saat ini.

Values (*A*) adalah nilai-nilai yang mungkin dari atribut (*A*).

D_v adalah subset dataset (*D*) yang memiliki nilai atribut (*A*) sama dengan (*v*).

Split Information:ni adalah pengukuran seberapa merata nilai-nilai atribut tersebar dalam dataset.

Rumus Split Information:

$$Entropy(S) = \sum_{t=1}^{i} \frac{S1}{S} log2 \frac{S1}{S} \qquad (3)$$

Setelah mendapatkan nilai Gain Informasi dan *Split Information*, dapat dihitung *Gain Ratio* yang menggambarkan informasi gain yang dinormalisasi terhadap tingkat dispersi atribut.

Rumus *Gain Ratio* :

$$Gainration(S, A) = \frac{Gain\ (S,A)}{SplitInformation\ (S,A)} \qquad (4)$$

Contoh kasus:

Misalkan Anda memiliki dataset yang ingin Anda gunakan untuk membangun pohon keputusan untuk memprediksi apakah suatu hewan adalah mamalia atau bukan berdasarkan atribut "Berbulu", "BerkakiEmpat", dan "Bertelinga".

No	Berbulu	Berkaki Empat	Bertelinga	Kelas
1	Ya	Ya	Ya	Mamlia
2	Tidak	Ya	Tidak	Bukan
3	Ya	Tidak	Tidak	Bukan
4	Ya	Ya	Tidak	Mamalia
5	Tidak	Ya	Ya	Mamlia

Table 2.1. Contoh kasus 1

Pertama, kita perlu menghitung entropi dari kelas (Mamalia dan Bukan) dalam dataset keseluruhan. Hitung proporsi masing-masing kelas:

Jumlah Mamalia : 2

Jumlah Bukan : 2

Total sampel : 5

Proporsi Mamalia : $p$mamalia = (2/5)

Proporsi Bukan : $p$buka = (2/5)

Entropy : *entropy (D)* = $-p_{\text{mamalia}} \cdot \log2(p_{\text{mamalia}}) - p_{\text{bukan}} \cdot \log2\,(p_{\text{bukan}})$

Hitung entropi:

Entropy(D) = $-(5/2)\cdot\log2(5/2)-(5/2)\cdot\log2(5/2)=0.97095$

Selanjutnya, akan menghitung *gain ratio untuk setiat atribut* (“Berbulu”, “Berkaki Empat”,”Bertelinga”) dan memilih atribut dengan *gain ratio* tertigggi Misalnya, menghitung *gain rasio* untuk atribut “Bebulu” :

- Hitung entropi untuksetiap nilai atribut “Berbulu” (“Ya” dan “Tidak”):
  Jumlah sampel “Berbulu” =5
  Jumlah sampel “Berbulu” dan Mamalia =2
  Jumlah sampel “Berbulu” dan Bukan =1
  Entropi (“Ya”) = -(2/5) . log2(2/5) – (3/5) . log2(3/5)
  Entropy (“Tidak”) = - (1/1) . log2(1/1)-0 (karena hanya ada satu sampel dan kelasnya Bukan)

- Hitung Gain Informasi
  Gain Informasi (“Bebulu”)= *entropy* (*D*) – (5/5). Entropy (“Ya”) – (1/5) . entropy (“Tidak”)
- Hitung Split Information
  Split Information (“Berbulu”) = -(3/5) . log2 (3/5) – (2/5) . log2(2/5)
- Hitung Gain Rasio
  Gain rasio (“Berbulu”) = $\frac{\text{Gain information (“Berbulu”)}}{\text{Split Informastion (“Berbulu”)}}$

Lakukan langkah yang sama untuk atribut “Berkaki Empat” dan “Bertelinga”. Pilih dengan Gain ratio tertinggi sebagai atribut yang akan membagi dataset.

2.4.2 *Association Rule*

Association rule mining adalah teknik dalam data mining yang difokuskan pada penemuan aturan kesamaan dalam suatu kejadian atau kumpulan data. Contoh penerapan aturan asosiasi yang sering dijumpai adalah dalam analisis pola pembelian di pusat perbelanjaan. Dengan menggunakan data historis transaksi, kita dapat mengidentifikasi hubungan antara pembelian berbagai produk. Sebagai contoh, kita mungkin menemukan bahwa pembelian susu sering kali diikuti dengan pembelian roti. Informasi ini dapat dimanfaatkan oleh pemilik toko untuk mengoptimalkan strategi penempatan produk dan memberikan diskon yang menarik, sehingga mendorong penjualan kedua produk tersebut.

Dalam kasus ini, metode-metode asosiasi digunakan untuk mengidentifikasi aturan-aturan seperti "jika membeli produk A, maka kemungkinan besar juga akan membeli produk B." Tujuan utama dari teknik ini adalah untuk mengungkap pola-pola yang tidak terlihat atau tidak terduga dalam data yang ada, yang nantinya dapat digunakan untuk pengambilan keputusan yang lebih baik.

Metode umum yang digunakan dalam penggalian aturan asosiasi meliputi FP-*Growth, Coefficient of Correlation, Chi Square, dan A Priori*. Setiap metode memiliki kelebihan dan kelemahan masing-masing, dan pemilihan metode tergantung pada karakteristik data serta tujuan analisis yang diinginkan. Kesimpulannya, penggalian aturan asosiasi memberikan wawasan berharga

tentang keterkaitan antara entitas dalam *dataset*, yang dapat diaplikasikan dalam berbagai konteks bisnis dan ilmiah untuk meningkatkan efisiensi dan pengambilan keputusan yang lebih cerdas.

2.4.3 Klasterisasi (*Clutering*)

Teknik klasterisasi berbeda dengan teknik klasifikasi, di mana dalam klasifikasi, kelas atau label data sudah ditentukan sebelumnya. Klasterisasi, atau yang dikenal juga sebagai *clustering*, adalah teknik yang bertujuan untuk mengelompokkan data secara otomatis berdasarkan kesamaan atau pola yang ada di dalamnya. Yang membedakan klasterisasi adalah bahwa dalam metode ini, tidak ada label kelas yang diberikan sebelumnya kepada data.

Pendekatan klasterisasi melibatkan pembentukan kelompok-kelompok data yang memiliki kemiripan, di mana setiap kelompok mencakup data dengan pola-pola serupa. Salah satu kegunaan utama klasterisasi adalah memberikan label pada data yang belum memiliki kelas yang ditentukan sebelumnya. Oleh karena itu, klasterisasi sering dikelompokkan sebagai metode pembelajaran tanpa pengawasan (*unsupervised learning*).

Berbagai metode klasterisasi dapat digunakan untuk mencapai tujuan ini. Beberapa metode yang sering digunakan meliputi:

1. K-*Medoids* : Metode ini mengelompokkan data dengan menemukan titik-titik tengah kelompok (*medoids*) yang mewakili setiap klaster.
2. K-*Means* : Mirip dengan K-*Medoids*, tetapi menggunakan *mean* (rata-rata) sebagai representasi kelompok.
3. *Fuzzy C-Means* : Menggunakan pendekatan fuzzy untuk menunjukkan sejauh mana data berada dalam setiap kelompok.
4. *Self-Organizing Map* (SOM) : Metode ini menggunakan jaringan saraf tiruan untuk memetakan data ke dalam kelompok.

Pilihan metode klasterisasi tergantung pada karakteristik data dan tujuan analisis. Klasterisasi memberikan wawasan tentang struktur internal data, mengidentifikasi kelompok-kelompok yang mungkin tidak terlihat pada pandangan pertama, dan membantu dalam pengenalan pola-pola yang relevan.

1.3 Pengetahuan (*Knowledge*)

Presentasi pengetahuan mengacu pada proses visualisasi dan penyampaian informasi serta pengetahuan yang dihasilkan dari eksplorasi data kepada pengguna. Pengetahuan yang diperoleh melalui data mining mencakup cara-cara untuk merumuskan keputusan spesifik atau langkah-langkah tindakan selanjutnya berdasarkan hasil analisis yang telah diperoleh. Tahap ini penting untuk mengkomunikasikan hasil pengolahan data mining kepada individu yang mungkin tidak memiliki pemahaman mendalam tentang konsep data mining.

Informasi yang dihasilkan dari pengolahan data mining perlu disajikan kepada pihak yang tidak memiliki latar belakang dalam bidang data mining. Oleh karena itu, kemampuan untuk menyajikan informasi secara jelas dan dapat dimengerti oleh berbagai pihak merupakan salah satu tahapan penting dalam proses data mining. Penggunaan visualisasi dalam bentuk grafik atau representasi visual lainnya sangat membantu dalam memfasilitasi komunikasi efektif mengenai hasil analisis data mining kepada berbagai pihak.

Dalam eksplorasi data, pengetahuan berperan sebagai jembatan untuk menghubungkan hasil analisis yang rumit dengan keputusan praktis yang harus diambil. Cara informasi ini disajikan dan disampaikan memiliki dampak signifikan pada pemahaman dan pengambilan keputusan yang akurat berdasarkan wawasan yang diperoleh melalui data mining.

1.4 Evaluasi

Evaluasi pola adalah langkah penting dalam proses eksplorasi data, yang melibatkan identifikasi dan penilaian terhadap pola-pola menarik yang ditemukan. Evaluasi memiliki dua dimensi utama: pertama, penilaian terhadap pola menarik atau model prediksi untuk memeriksa apakah mereka memenuhi hipotesa awal; kedua, jika hasil evaluasi tidak sesuai dengan ekspektasi, langkah-langkah alternatif dapat diambil untuk memperbaiki proses data mining.

Dalam konteks evaluasi data mining, khususnya dalam tipe klasifikasi, evaluasi dilakukan dengan menguji proses prediksi kebenaran objek. Metode evaluasi ini mengandalkan penggunaan matriks konfusi (*confusion matrix*), di mana kelas prediksi ditempatkan di bagian atas matriks, dan kelas yang diamati

ditempatkan di bagian kiri matriks. Matriks ini memuat angka-angka yang mencerminkan jumlah kasus aktual dari kelas yang diamati.

Sebagai contoh, dalam kasus evaluasi model klasifikasi dua kelas, seperti yang dijelaskan pada Tabel 2.9, kita memiliki matriks konfusi dengan nilai-nilai kelas 0 dan 1. Setiap elemen matriks (fij) mencerminkan jumlah rekaman dari kelas i yang diprediksi masuk ke kelas j selama pengujian.

Evaluasi kinerja model data mining juga melibatkan pengukuran akurasi (*accuracy*) atau tingkat kesalahan (*error rate).* Dalam proses ini, *confusion matrix* adalah alat yang umum digunakan untuk menilai performa model klasifikasi. Oleh karena itu, langkah evaluasi tidak hanya membantu memahami sejauh mana model dapat memprediksi dengan benar, tetapi juga memberikan informasi tentang kelas-kelas yang mungkin mengalami kesalahan prediksi.

Melalui langkah evaluasi yang teliti, kita dapat mengidentifikasi aspek-aspek yang memerlukan perbaikan atau peningkatan dalam model data mining, dan dengan demikian, mengarah pada pengembangan model yang lebih akurat.

1.5 Pohon Keputusan

Algoritma C4.5 adalah salah satu metode pembuatan pohon keputusan yang populer dalam pembelajaran mesin. Algoritma ini dikembangkan oleh Ross Quinlan dan merupakan pengembangan dari algoritma ID3. C4.5 bekerja dengan membangun pohon keputusan dari data pelatihan dengan cara memilih atribut yang paling informatif pada setiap langkahnya. Pemilihan atribut ini dilakukan berdasarkan teknik gain informasi dan split informasi untuk mengukur relevansi dan kemurnian atribut dalam membedakan kelas target.

Salah satu keunggulan C4.5 adalah kemampuannya untuk menangani data kategorikal dan kontinu, serta dapat mengatasi masalah atribut yang hilang. Algoritma ini juga mampu menghasilkan pohon keputusan yang lebih kecil dengan teknik pemangkasan, sehingga mengurangi overfitting dan meningkatkan generalisasi model. C4.5 merupakan algoritma yang relevan dan efektif untuk prediksi produksi ikan nila di kelompok budidaya karena dapat memodelkan hubungan kompleks antara atribut lingkungan dengan produksi ikan nila.

Dengan menggunakan algoritma C4.5, dapat di identifikasi atribut lingkungan yang paling berpengaruh terhadap produksi ikan nila dan membangun model prediktif berbasis pohon keputusan untuk memprediksi produksi ikan nila berdasarkan kondisi lingkungan yang diberikan.

1.6 Konsep Pohon Keputusan

Pohon keputusan adalah metode analisis data yang menggambarkan struktur hierarkis berbentuk diagram cabang-cabang (nodes) yang terdiri dari simpul-simpul dan daun (leaves). Setiap simpul pada pohon mewakili keputusan berdasarkan nilai atribut tertentu, sedangkan cabang menghubungkan simpul-simpul dan daun untuk menggambarkan alur keputusan.

Dalam prediksi produksi ikan nila, atribut-atribut lingkungan seperti suhu, kadar oksigen, nilai protein, populasi, pH air, akan menjadi simpul pada pohon keputusan. Setiap simpul akan memisahkan data berdasarkan nilai atribut tertentu untuk mengidentifikasi pola dan hubungan yang berpengaruh terhadap produksi ikan nila.

1.7 Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai acuan refrensi pada penelitian yang dilakukan :

Nama Peneliti dan Tahun	Judul	Metode	Hasil Penelitian
(Matondang, 2021)	Analisa data mining dengna metode algoritma c4.5 pada klasifikasi kenaikan rata-rata volume perikanan	Metode yang digunakan adalah algoritma c4.5	Hasil pengujian atribut nelayan dengan akurasi 100%
(Karyadiputra, 2022)	Rancang Bangun penerapan data mining utuk	Metode algoritma c4.5 dengan penjuian model	Hasil penelitian ini adalah pengujian didapatkan akurasi *k*-

	klasifikasi spesies ikan di lingkunaga akuatik air tawar	*naive bayes* dan *k-nears neighbor* (KNN)	*nearest neighbor* sebesar 82,32% dengan nilai kappa 0,737

Tabel 2.2. Penelitian Terdahulu 1