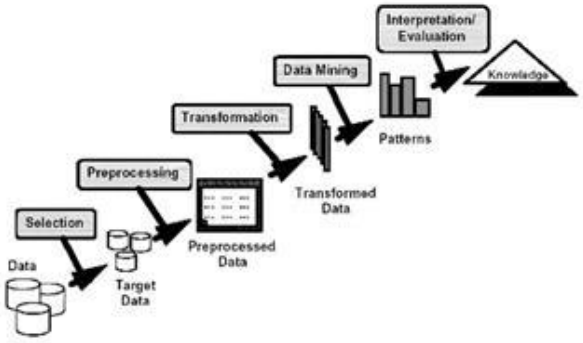
BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining merupakan istilah yang yang sering dikatakan sebagai salah satu cara untuk menguraikan serta mencari penemuan berupa pengetahuan didalam suatu *database* atau yang bisa disebut dengan Knowledge Discovery In Database (KDD). Data Mining merupakan tahapan untuk menemukan sebuah pola yang ada didalam hubungan antar dua data dalam sebuah database yang berukuran besar. Data mi ning pada prinsipnya yaitu proses pengalian data/informasi dari database yang tidak terbatas dan besar, faktanya Data mining merupakan langkah menuju *Knowladge Discovery in Database.* Data Mining dan *Knowledge in Database* (KDD) digunakan secara bergantian untuk menjelaskan sebuah proses yang tersembunyi. Akan tetapi kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. *Knowladge Discovery In Database* yang berkaitan dengan proses penemuan pengetahuan yang diterapkan pada database. Hal ini bisa juga didefinisikan sebagai proses *non-trival* untuk mengidentifikasi data yang valid, baru, berpotensi, bermanfaat, dan memiliki pola yang dapat dimengerti.. Dan salah satu tahap dalam proses KDD adalah Data Mining (Arta et al., 2019).

Ada beberapa tahapan pada proses KDD menurut (Antika et al., 2024) yang ditunjukkan pada Gambar 2.1 dibawah ini :



Gambar 2. 1 Tahapan KDD

Pola gambar 2.1 diatas merupakan tahapan dari proses KDD yang bisa disebut dengan interpretation. Lima tahapan pada proses *Knowladge Discovery in Database* (KDD) adalah :

1. *Selection*

   Pada tahapan ini yaitu persiapan dalam pemilihan data yang sudah didapatkan dari wawancara dengan memberiskan data yang memiliki *noise*. Data dari hasil seleksi yang akan digunakan untuk proses data mining tersebut disimpan dalam suatu berkas terpisah dari basis data operasional.

2. *Preprocessing*

   Tahap Processing ini bertujuan untuk mengubah data mentah diolah dengan membuang publikasi data, memeriksa data yang inkosisten dan juga memperbaiki kesalahan pada data yang *noise* atau eror.

3. *Transformation*

   Pada tahapan ini dimana data dirubah disesuaikan dengan format *ekstensi* yang sesuai untuk proses data. Hal ini dilakukan karena beberapa metode dalam data mining memerlukan format tertentu sebelum .

4. Data Mining

   Tahap Data Mining merupakan penemuan pola atau informasi menarik di dalam data melalui penggunaan teknik tertentu sesuai dengan teknik tertentu. Pada tahapan ini menggunakan banyak teknik atau algoritma untuk memperoleh sebuah pola dari data.

5. *Evaluation*

   Pola yang sudah dihasilkan dari proses data mining harus ditampilkan pada tahapan ini agar pihak yang berkepentingan dapat memahaminya. Mempresentasikan hasil model yang sudah diperoleh dan juga menguji akurasi dan kesesuaian terhadap data-data dan memastikan apakah pola atau informasi yang dikumpulkan bertentangan dengan fakta atau hipotesis sebelumnya.

2.1.1 Klasifikasi Data Mining

Klasifikasi data merupakan suatu teknik yang digunakan untuk menilai suatu objek data untuk dikategorikan ke kelas-kelas tertentu dari banyaknya

jumlah kelas yang ada. Klasifikasi adalah salah satu yang melibatkan pembangunan model yang juga bisa memprediksi kelas atau label suatu objek berdasarkan dengan atribut yang dimiliki. Yang dihasilkan oleh proses klasifikasi yaitu dapat digunakan untuk mengklasifikasikan data baru yang belum dikenal dalam suatu kelas yang telah ditentukan sebelumnya (Novianti et al., 2023). Dengan menggunakan teknik ini, data yang telah dikelompokkan dan dapat diolah, berdasarkan hasil analisisnya juga dapat dibuat aturan untuk mengklasifikasikan data baru ke kelompok yang sesuai (Mahesa Putra et al., 2024). Metode klasifikasi berkaitan pada pembentukan kelompok data dengan cara menerapkan algoritma ke gudang data yang di bawah pemeriksaan. Dalam konteks ini, metode Decission Tree digunakan untuk membuat keputusan klasifikasi berdasarkan dengan aturan yang dibangun dari data pelatihan sebelumnya.

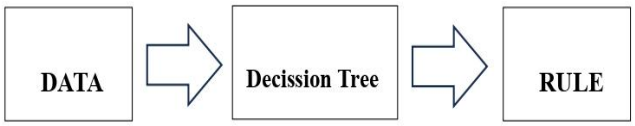
2.2 Pohon Keputusan (Decision Tree)

Pohon keputusan atau Decision tree merupakan metode klasifikasi yang kuat dan terkenal. Metode ini mengubah fakta yang besar menjadi pohon keputusan yang mempresentasikan aturan. Decisiion tree juga dapat disebut sebagai sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil lagi dan menerapkan aturan keputusan. Decision tree merupakan struktur flowcart yang memiliki *tree* (pohon). Teknik ini menggunakan teknik membagi ruang pencarian menjadi himpunan masalah. Pada Decision Tree setiap sampul daun merupakan label kelas. Simpul yang bukan simpul akhir terdiri dari akar dan simpul internal. Simpul akar dan simpul internal ditandai dengan bentuk oval dan simpul daun berbentuk segi empat. Alur Decision tree yaitu dari simpul akar ke simpul daun yang memegang prediksi kelas (Adhi Guna et al., 2023).

Ada beberapa keuntungan menggunakan algoritma decission tree ini, yaitu kemampuan dalam interpretasi, kemampuan menangani data kategori maupun numerik, dan juga kecenderungan untuk tidak menggunakan normalisasi atau standarisasi data. Tetapi, dapat rentan terhadap *overfitting* yaitu model terlalu sesuai dengan data pelatihan dan mungkin juga tidak dapat melakukan

generalisasi dengan baik di data baru. Pohon keputusan juga memiliki keunggulan yaitu kemampuannya dalam mengeliminasi perhitungan atau data yang tidak diperlukan, dalam pengujian sampel hanya kriteria atau kelas tertentu yang digunakan. Namun juga memiliki kelemahan seperti risiko tumpang tindih, terutama pada kelas dan kriteria yang digunakan terlalu rumit yang dapat memperlambat waktu dalam pengambilan keputusan karena membutuhkan kapasitas memori yang lebih besar.

Decission Tree adalah model prediksi yang sifatnya *supervised* yang berarti membutuhkan training dataset yang berperan menggantikan pengalaman manusia di masa lalu dalam membutat sebuah keputusan. Proses Decision Tree merupakan mengubah bentuk data tabel menjadi sebuah model *tree.* Model *tree* akan menghasilkan *rule* dan disederhanakan. Berikut adalah gambar konsep cara kerja dan struktur *Decision tree* (Bahri & Lubis, 2020).

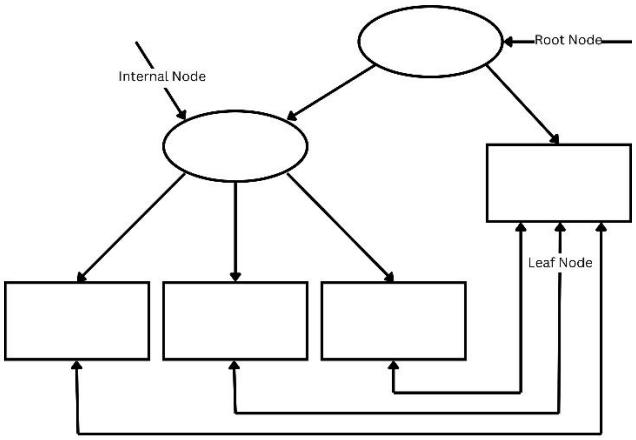


Gambar 2. 2 Konsep Decision Tree

Penjelasan pada gambar 2.2 Konsep Decision Tree yaitu :

1. Data: Proses dimulai dengan data, yang merupakan masukan untuk pohon keputusan.
2. Pohon Keputusan (Decision Tree): Data kemudian dimasukkan ke dalam pohon keputusan, yang merupakan algoritma yang digunakan untuk membuat keputusan berdasarkan data masukan. Pohon keputusan menganalisis data dan membaginya menjadi cabang-cabang berdasarkan kriteria tertentu.
3. Aturan (Rule): Hasil dari proses pohon keputusan adalah sekumpulan aturan. Aturan-aturan ini dapat digunakan untuk membuat prediksi atau mengambil keputusan berdasarkan data baru.

Dibawah ini merupakan contoh gambar struktur Decision Tree menurut (Kandi Sri, 2020) :



Gambar 2. 3 Struktur Decision Tree

Penjelasan pada gambar 2.3 Struktur Decision Tree :

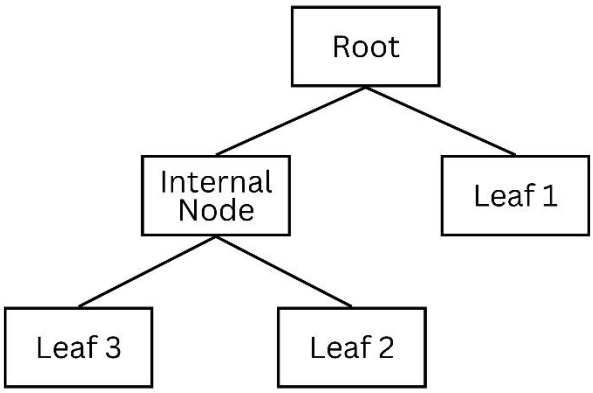
1. Root Node (Simpul Akar): Ini adalah simpul paling atas dalam pohon, dari mana semua simpul lainnya bercabang. Dalam diagram, simpul akar ditandai dengan panah yang mengarah padanya dan berlabel "Root Node."
2. Internal Node (Simpul Internal): Simpul ini memiliki anak atau cabang. Dalam diagram, ada satu simpul internal yang ditandai dengan panah dan berlabel "Internal Node." Simpul internal berfungsi sebagai titik percabangan dalam pohon.
3. Leaf Node (Simpul Daun): Ini adalah simpul yang tidak memiliki anak. Dalam diagram, simpul daun ditandai dengan panah dan berlabel "Leaf Node." Simpul daun merupakan akhir dari setiap cabang pohon.

2.2.1 Model Pohon Keputusan (*Decission Tree*)

Decision Tree merupakan metode klasifikasi yang popular digunakan karena pembangunannya yang cepat dan hasil dari model tersebut dibangun agar mudah untuk dipahami. Cara untuk membuat model *decission Tree* adalah memecah data ke kelompok yang lebih kecil berdasarkan atribut didalam data. Pembagian tersebut dilakukan secara berulang kali sampai seluruh elemen data yang berasal

dari kelas yang sama bisa masuk ke satu kelompok. Ada beberapa algoritma yang biasanya digunakan dalam melatih model Decission Tree, yaitu bernama CART *(Classification and Regression Trees*) yang dipakai dalam *Scikit-learn*. CART ini adalah turunan dari algoritme yang bernama C4.5 yang adalah turunan dari algoritma ID3 (Ramadhon et al., 2024).

Berikut adalah contoh gambar pohon keputusan menurut (Setio et al., 2020) yang ditunjukkan pada gambar dibawah ini :



Gambar 2. 4 Model *Decision Tree*

Penjelasan pada gambar 2.4 Model *Decision Tree* terdapat tiga jenis node menurut (Tukino, 2020), yaitu :

1. *Root Node*, adalah node paling atas, pada node ini tidak ada input dan tidak memiliki output ataupun memiliki output lebih dari satu.
2. *Internal Node*, adalah node percabangan, dalam node ini terdapat satu input dan memiliki output minimal dua.
3. *Leaf Node,* adalah node terakhir, dalam node ini hanya ada satu input dan tidak memiliki output.

   Berikut ini merupakan tahapan dari Decision tree menurut (Ade & Pratomo, 2020).

1. Menyiapakan data training. Dari data histori yang sudah pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon yang diambil dari atribut yang terpilih .

3. menghitung nilai *grain* dari masing-masing atribut. Nilai *grain* yang tertinggi menjadi akar pertama. Sebelum menghitung nilai grain dari atribut, hitung nilai *entropy*.

2.2.2 Algoritma ID3

Menurut (Amrin et al., 2024) ross Quinlan melakukan pengembangan *Iterative Dichotomizer* 3 pada tahun 1986. Algoritme ini ialah salah satu algoritme pohon keputusan yang dapat digunakan untuk prediksi dan klasifikasi. Algoritma pada metode ini menggunakan konsep dari entropy informasi. Algoritma *Iterative Dichotomiser* 3 (ID3) merupakan salah satu metode Data Mining yang menghasilkan pohon keputusan. Data testing akan diuji menggunakan pohon keputusan untuk memperoleh prediksi. Metode ini adalah sebuah *flowchart* yang mirip seperti struktur pohon, setiap titik pohon merupakan atribut yang telah diuji, cabang merupakan hasil uji, dan titik akhir merupakan pembagian kelas yang dihasilkan.

Algoritma ID3 melakukan pemecahan data ke dua kelompok berdasarkan dengan atribut yang ada di dalam data, algoritma ID3 ini menggunakan konsep dari entropy informasi dan pemilihan atribut menggunakan information gain. Proses algoritma ID3 diuraikan sebagai berikut : (Setio et al., 2020).

1. Menentukan nilai dari entropy dengan rumus yang ditulis sebagai (2.1)

2.1

$$Entropy(S) = -\sum_{i=1}^{n} pi \cdot log_2(pi) \qquad (2.1)$$

S adalah data sample yang digunakan untuk latih P+ (jumlah bersolusi positif) atau P- (jumlah bersolusi negatif) dari sejumlah data acak suatu ruang sampel untuk kriteria tertentu. Entropy merupakan jumlah bit yang dibutuhkan untuk menyatukan kelas.

Pada umumnya semakin kecil nilai entropy maka semakin baik untuk mengkestrak kelas. S merupakan himpunan kelas klasifikasi, C merupakan banyaknya kelas klasifikasi dan P merupakan proposi untuk kelas i.

2. Setelah mendapatkan nilai entropy, maka bisa diukur efektivitas atribut dalam melakukan

klasifikasi data yang disebut dengan *information gain*, secara matematis, *information gain* dari suatu atribut A yang dituliskan sebagai berikut : (2.2)

2.2

$$Gain(S, A) = Entropy(S) - \sum_{\text{v}\in\text{Values(A)}} \frac{S\text{v}}{S} Entropy\ (S\text{v}) \qquad (2.2)$$

Dengan, A merupakan atribut

V menyatakan suatu nilai yang mungkin untuk atribut A

Values (A) merupakan himpunan nilai-nilai yang mungkin untuk atribut

Sv merupakan sub-himpunan kelas klasifikasi

Entropy Sv merupakan entropy untuk sampel yang memiliki nilai v

3. Atribut yang memiliki nilai information gain tertinggi dibandingkan dengan atribut yang lain, maka dipilih sebagai pemilah.
4. Membentuk simpul yang berisi atribut yang dimaksudkan di dalam no 3.
5. Proses dalam perhitungan information gain dilakukan secara berulang sampai semua data masuk ke dalam kelas yang sama. Atribut yang sudah dipilih tidak dimasukkan lagi pada perhitungan nilai *information gain*.

2.3 Riview Artikel

Sebagai upaya penguatan topik penelitian, penulis melakukan analisis dari hasil riset penelitian sebelumnya yang berkaitan dengan topik penelitian. Berikut

ini beberapa hasil dari penelitian terdahulu yang relevan ditampilkan pada tabel 2.1

Tabel 2. 1 Hasil Riview Artikel

Judul dan Masalah	Hasil Penelitian	Dataset	Landasan Literatur
Judul: Implementasi Algoritma Decision Tree ID3 untuk Klasifikasi Penyakit Diabetes Mellitus Masalah: Mengklasifikasikan pasien diabetes melitus berdasarkan data rekam medis untuk membantu deteksi dini dan intervensi yang tepat.	Algoritma ID3 menunjukkan akurasi yang baik dalam mengklasifikasikan pasien diabetes, menghasilkan aturan keputusan yang mudah dipahami oleh tenaga medis, yang dapat menjadi alat bantu diagnostik awal.	Data rekam medis pasien diabetes yang mencakup variabel seperti usia, BMI, kadar glukosa, riwayat keluarga, dan tekanan darah, diperoleh dari rumah sakit atau klinik lokal.	Setiawan, A., & Purnomo, H. (2020).
Judul: Prediksi Kelayakan Kredit Menggunakan Algoritma ID3 Masalah: Menentukan kelayakan pemberian kredit kepada calon nasabah secara objektif dan efisien untuk meminimalkan risiko gagal bayar bagi lembaga keuangan.	ID3 mampu membangun model prediksi kelayakan kredit dengan akurasi yang memadai, memberikan rekomendasi yang jelas mengenai persetujuan atau penolakan kredit berdasarkan karakteristik nasabah. Aturan yang dihasilkan mudah diinterpretasikan.	Data historis nasabah kredit yang mencakup informasi seperti penghasilan, riwayat pinjaman, status pekerjaan, dan tanggungan.	Pratama, D. N., & Lestari, S. (2021).
Judul: Analisis Faktor Penentu Calon Penerima Beasiswa Berbasis ID3 Masalah: Mengidentifikasi faktor-faktor utama yang mempengaruhi	ID3 berhasil memetakan kriteria penerima beasiswa, memberikan pemahaman yang lebih baik tentang prioritas dan bobot kriteria seperti IPK, prestasi non-akademik, hasil wawancara, dan kondisi ekonomi.	Data pendaftar beasiswa dari institusi pendidikan, mencakup nilai akademik, prestasi, hasil wawancara, dan informasi	Hidayat, R., & Susanto, A. (2022).

kelulusan atau penerimaan calon mahasiswa dalam seleksi beasiswa untuk meningkatkan transparansi dan objektivitas.		demografi/ekonomi.	
Judul: Klasifikasi Sentimen Ulasan Produk Menggunakan Decision Tree ID3 Masalah: Mengklasifikasikan sentimen (positif, negatif, netral) dari ulasan produk daring untuk membantu perusahaan memahami preferensi dan kepuasan konsumen.	ID3 mampu mengklasifikasikan sentimen dengan akurasi yang cukup baik setelah proses pra-pemrosesan teks yang memadai (misalnya, *tokenisasi*, *stemming*, dan pembobotan TF-IDF). Model yang dihasilkan mudah diinterpretasikan untuk analisis pasar.	Kumpulan ulasan produk dari platform e-commerce atau situs ulasan populer, yang telah diberi label sentimen.	Wicaksono, B. A., & Ramadhani, F. (2023).
Judul: Prediksi Produktivitas Masyarakat di Era Digital Menggunakan Algoritma ID3 Masalah: Mengidentifikasi faktor-faktor yang berkorelasi dengan produktivitas masyarakat, terutama dalam konteks ekonomi digital dan pengembangan Usaha Mikro, Kecil, dan Menengah (UMKM).	ID3 dapat membangun model untuk memprediksi tingkat produktivitas dengan mengidentifikasi variabel kunci seperti akses teknologi, tingkat pendidikan, jenis usaha, dan partisipasi dalam pelatihan digital.	Data survei masyarakat atau data terkait UMKM (misalnya, omzet, penggunaan platform digital, tingkat pendidikan pelaku usaha).	Saputro, T., & Dewi, P. S. (2024).
Judul: Deteksi Penyakit Tiroid dengan Algoritma	ID3 menunjukkan kinerja yang baik dalam memprediksi penyakit tiroid, menghasilkan aturan	Data medis pasien dengan riwayat tes tiroid (misalnya,	Wijaya, S. P., & Nurjanah, S. (2020).

Decision Tree ID3 Masalah: Klasifikasi pasien yang berisiko atau positif mengidap penyakit tiroid berdasarkan hasil tes laboratorium dan gejala klinis, untuk mendukung diagnosis awal.	keputusan yang membantu tenaga medis dalam mengidentifikasi pola-pola gejala dan hasil tes yang mengindikasikan penyakit.	TSH, T3, T4) dan gejala klinis dari rumah sakit atau klinik.	
Judul: Perbandingan ID3 dengan C4.5 untuk Klasifikasi Data Cuaca Masalah: Membandingkan kinerja ID3 dan C4.5 dalam mengklasifikasikan kondisi cuaca (misalnya, hujan/tidak hujan, cerah/berawan) berdasarkan parameter iklim, untuk menentukan algoritma yang lebih optimal.	Kedua algoritma menunjukkan performa yang relatif mirip, namun C4.5 seringkali sedikit lebih unggul dalam akurasi karena kemampuannya menangani data kontinu dan nilai yang hilang. Meskipun demikian, ID3 tetap lebih sederhana dan cepat dalam pembangunan model.	Data historis cuaca dari stasiun meteorologi, mencakup variabel seperti suhu, kelembaban, tekanan udara, dan kecepatan angin.	Rahman, F., & Indah, S. (2021).
Judul: Peningkatan Kinerja Algoritma ID3 dalam Prediksi Gelombang Ekstrim Air Laut Masalah: Meningkatkan akurasi prediksi gelombang ekstrim air laut yang krusial untuk mitigasi bencana maritim, keselamatan pelayaran, dan	Dengan beberapa penyesuaian pada pra-pemrosesan data atau penggunaan teknik feature engineering, ID3 dapat memberikan hasil prediksi gelombang ekstrim dengan tingkat akurasi yang lebih baik dibandingkan implementasi dasar, menunjukkan potensi ID3 dalam domain geofisika.	Data maritim yang meliputi tinggi gelombang, kecepatan angin, arah angin, dan pasang surut dari data historis pelampung laut atau model simulasi.	Utomo, P., & Sari, N. K. (2022).

perencanaan pesisir.			
Judul: Klasifikasi Hasil Belajar Siswa Berdasarkan Minat dan Motivasi Menggunakan ID3 Masalah: Memahami bagaimana minat dan motivasi siswa mempengaruhi hasil belajar mereka, serta mengklasifikasikan siswa ke dalam kategori hasil belajar tertentu (misalnya, sangat baik, baik, cukup) untuk intervensi pendidikan yang personal.	Algoritma ID3 berhasil mengidentifikasi pola antara minat, motivasi, dan hasil belajar, membantu guru dalam merancang strategi pembelajaran yang lebih efektif dan mengidentifikasi siswa yang membutuhkan perhatian lebih.	Data siswa yang mencakup nilai tes, hasil kuesioner minat, tingkat motivasi, dan hasil belajar akhir.	Puspitasari, A., & Nugroho, D. (2023).
Judul: Klasifikasi Risiko Penyakit Jantung Menggunakan ID3 Masalah: Kurangnya sistem pendukung keputusan untuk klasifikasi risiko penyakit jantung.	Penelitian ini menggunakan ID3 untuk mengklasifikasikan tingkat risiko penyakit jantung berdasarkan data medis pasien. Dengan akurasi sebesar 85,71%, ID3 terbukti mampu mengidentifikasi pasien berisiko dengan cukup baik. Fitur yang paling berpengaruh dalam klasifikasi adalah tekanan darah, usia, dan riwayat keluarga. Penelitian ini mendukung penggunaan ID3 dalam sistem pendukung keputusan di bidang kesehatan.	Data Heart Disease UCI (303 entri), atribut seperti tekanan darah, usia, kolesterol.	Fauziah, E. & Zulfikar, A. F., (2023).