BAB II

LANDASI TEORI

2.1. Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) didefinisikan sebagai keseluruhan proses non-trivial untuk mengidentifikasi pola yang sah baru berpotensi berguna, dan pada akhirnya dapat dipahami dari data. Proses KDD terdiri dari beberapa tahapan berurutan, dimulai dari seleksi data, pembersihan data, transformasi, data mining, hingga interpretasi dan evaluasi pola. Pemahaman terhadap setiap tahap ini penting untuk menjamin bahwa hasil analisis data penjualan menggunakan algoritma C4.5 benar-benar akurat dan relevan bagi pengambilan keputusan bisnis.

2.1.1. Defenisi dan Konsep Dasar KDD

Knowledge Discovery in Databases (KDD) merupakan sebuah proses sistematis yang sangat terstruktur yang bertujuan untuk menemukan pengetahuan yang bermanfaat dari data dalam jumlah besar**.** Menurut definisi yang dikemukakan oleh Fayyad, Piatetsky-Shapiro, dan Smyth, KDD adalah proses keseluruhan (the overall process) untuk mengidentifikasi pola yang valid, baru, berpotensi berguna, dan pada akhirnya dapat dipahami dari data. Dengan kata lain, KDD menjawab pertanyaan mendasar: bagaimana cara kita mengubah sekumpulan data mentah yang sangat besar menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan Keputusan (Gonzalo Mariscal, 2022).

Gambar 2. 1. Tahapan KDD

Konsep ini muncul sebagai respons terhadap ledakan data yang terjadi akibat komputerisasi di berbagai sektor bisnis maupun pemerintahan. Volume data yang terus membesar secara eksponensial secara langsung melampaui kapasitas analisis manusia. Dalam situasi seperti ini, KDD menjadi suatu keharusan, bukan sekadar pilihan. KDD menjanjikan peran penting dalam cara orang berinteraksi dengan basis data, terutama di mana analisis dan eksplorasi data menjadi aktivitas yang sangat kompleks.

Salah satu kesalahan konseptual yang paling umum terjadi dalam literatur dan praktik adalah penyamakan antara KDD dan Data Mining**.** Banyak sumber menggunakan kedua istilah ini secara bergantian, padahal terdapat perbedaan hierarkis yang fundamental. KDD adalah proses keseluruhan dari penemuan pengetahuan, sementara data mining adalah salah satu langkah spesifik dalam proses tersebut, yaitu langkah di mana algoritma diterapkan untuk mengekstraksi pola dari data (Rokach, 2023).

2.1.2. Tahapan Proses KDD

Proses KDD yang paling banyak diadopsi dalam literatur dan praktik industri terdiri dari enam hingga tujuh tahapan utama. Penting untuk dipahami bahwa tahapan-tahapan ini tidak selalu berurutan secara linier, melainkan iteratif

yaitu peneliti dapat (dan seharusnya) kembali ke tahap sebelumnya ketika ditemukan kejanggalan atau peluang perbaikan. Berikut adalah penjabaran lengkap dari masing-masing tahapan

Tahap 1, Seleksi data adalah tahap penentuan cakupan data yang akan dianalisis. Dalam praktiknya, organisasi sering kali memiliki banyak sumber data yang berbeda dari basis data transaksi, log sistem, survei pelanggan, hingga data eksternal. Pada tahap ini, peneliti harus memilih subset data yang paling relevan dengan tujuan analisis. Tujuan utama dari tahap ini adalah: (i) mengidentifikasi data mana yang benar-benar diperlukan, (ii) menentukan periode waktu yang relevan, dan (iii) memutuskan atribut-atribut mana yang akan digunakan. Seleksi data yang terlalu luas akan menghasilkan proses komputasi yang lambat dan berisiko memasukkan *noise* yang tidak relevan; seleksi data yang terlalu sempit berisiko kehilangan informasi penting (Sya'roni, 2026).

Tujuan utama dari tahap ini adalah: (i) mengidentifikasi data mana yang benar-benar diperlukan, (ii) menentukan periode waktu yang relevan, dan (iii) memutuskan atribut-atribut mana yang akan digunakan. Seleksi data yang terlalu luas akan menghasilkan proses komputasi yang lambat dan berisiko memasukkan *noise* yang tidak relevan; seleksi data yang terlalu sempit berisiko kehilangan informasi penting.

Tahap 2, Pembersihan data merupakan tahap yang paling memakan waktu (hingga 70-80% dari total waktu proyek data mining), tetapi sekaligus merupakan tahap yang paling kritis. Kualitas hasil analisis sepenuhnya tergantung pada kualitas data yang dimasukkan ke dalam algoritma. Prinsip yang berlaku adalah

data yang buruk akan menghasilkan pengetahuan yang buruk, tidak peduli sehebat apa pun algoritma yang digunakan.

Aktivitas utama dalam tahap ini mencakup empat hal. Pertama, penanganan *missing value*. Data penjualan sering kali memiliki entri yang kosong (misalnya atribut promosi mungkin tidak tercatat dengan baik). Metode penanganannya antara lain membuang baris yang kosong, mengisi dengan nilai rata-rata atau median, atau menggunakan metode estimasi yang lebih canggih. Kedua, penghapusan data duplikat**.** Transaksi yang tercatat dua kali karena kesalahan sistem harus diidentifikasi dan dihapus agar tidak mendistorsi hasil analisis. Ketiga, koreksi inkonsistensi data**.** Perbedaan format seperti Kopi Susu dan kope susu harus distandardisasi. Keempat, deteksi dan penanganan *outlier*. Transaksi dengan nilai yang secara statistik sangat berbeda dari data lain (contohnya pembelian 100 gelas kopi dalam satu transaksi) perlu ditinjau apakah itu kesalahan pencatatan atau suatu kejadian yang valid.

Tahap 3, Setelah data bersih, tahap berikutnya adalah mengubah data ke dalam format yang sesuai untuk algoritma C4.5**.** Algoritma C4.5 memiliki karakteristik dapat menangani data kategorikal (seperti : jenis produk: kopi/non-kopi") maupun data numerik kontinu (seperti harga atau jumlah terjual). Namun, agar pohon keputusan yang dihasilkan mudah diinterpretasikan, data numerik sering kali perlu diubah ke dalam kategori diskrit.

Aktivitas utama dalam transformasi data mencakup: (i) diskretisasi mengubah data numerik kontinu menjadi interval atau kategori (contoh: harga menjadi murah, sedang, mahal), (ii) agregasi menggabungkan beberapa nilai

menjadi satu (contoh: waktu transaksi dikelompokkan menjadi pagi, siang, sore, malam), dan (iii) normalisasi menyamakan skala data jika diperlukan untuk perbandingan (Rachman, 2022).

Tahap 4, Data mining adalah tahap di mana keajaiban terjadi. Pada tahap inilah algoritma spesifik diterapkan pada data untuk mengekstraksi pola-pola yang tersembunyi. Tahap ini sering dianggap sebagai "jantung" dari keseluruhan proses KDD, tetapi penting untuk diingat bahwa jantung hanya akan berfungsi dengan baik jika organ-organ lain (tahap sebelumnya) sehat. Dalam penelitian ini, algoritma data mining yang digunakan adalah C4.5 yang termasuk dalam keluarga algoritma klasifikasi. C4.5 adalah algoritma berbasis pohon keputusan (*decision tree*) yang terkenal karena kemampuannya menghasilkan model yang sangat mudah diinterpretasikan

Tahap 5, interpretasi dan evaluasi adalah jembatan antara dunia teknis dan dunia bisnis. Model yang dihasilkan pada tahap data mining perlu dievaluasi untuk memastikan bahwa model tersebut benar-benar dapat diandalkan (*reliable*) dan valid, serta diinterpretasikan dalam konteks bisnis yang sebenarnya. Tanpa interpretasi yang tepat, pohon keputusan yang dihasilkan C4.5 hanyalah sekumpulan aturan logika yang tidak bermakna secara praktis. Interpretasi pola adalah proses menerjemahkan pohon keputusan ke dalam wawasan bisnis yang dapat ditindaklanjuti. Sebagai contoh, cabang pohon yang menunjukkan bahwa penjualan tinggi terjadi pada sore hari dengan promosi diskon untuk menu kopi diterjemahkan menjadi rekomendasi strategis, Tingkatkan promosi pada sore hari untuk menu kopi untuk memaksimalkan penjualan (Hermawan, 2024).

2.2. Klasifikasi

Model klasifikasi merupakan salah satu pendekatan dalam data mining yang digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang ditemukan dari data historis. Model ini bekerja dengan mempelajari hubungan antar atribut dalam data training dan kemudian membentuk aturan atau pohon keputusan yang dapat digunakan untuk memprediksi kelas dari data baru. Tujuan utama dari klasifikasi adalah untuk membangun model prediktif yang mampu mengklasifikasikan data secara akurat ke dalam kelas-kelas yang telah ditentukan. Berbagai algoritma klasifikasi seperti Decision Tree C4.5, Naïve Bayes, dan K-Nearest Neighbor KNN umum digunakan tergantung pada jenis data dan tujuan analisis (Chandra, E., & Natalia, 2024).

Dalam konteks penelitian ini, klasifikasi digunakan untuk mengelompokkan data transaksi penjualan Coffee Castello ke dalam kategori tingkat penjualan tertentu (misalnya: Penjualan Tinggi, Penjualan Sedang, Penjualan Rendah) berdasarkan atribut-atribut seperti jenis produk, harga, waktu transaksi, dan ada tidaknya promosi. Dengan demikian, model klasifikasi berperan penting dalam mendukung pengambilan keputusan berbasis data dan meningkatkan kualitas layanan secara berkelanjutan (Nazifah, 2023).

Sebuah sistem klasifikasi umumnya memiliki tiga komponen utama:

1. Data latih yaitu kumpulan data yang sudah diketahui label kelasnya. Data latih digunakan oleh algoritma untuk “belajar” pola hubungan antara atribut dan kelas. Kualitas dan representativitas data latih sangat menentukan akurasi model.

2. Aturan atau model klasifikasi yaitu Hasil pembelajaran dari data latih. Model dapat berupa pohon keputusan, himpunan aturan *if-then*, persamaan matematis, atau jaringan syaraf tiruan. Dalam algoritma C4.5, model yang dihasilkan berbentuk pohon keputusan (*decision tree*) yang mudah diinterpretasikan.
3. Data uji, Data baru yang labelnya tidak diketahui (atau diketahui tetapi sengaja disembunyikan untuk evaluasi). Model klasifikasi akan memprediksi kelas data uji berdasarkan pola yang telah dipelajari. Selanjutnya, prediksi tersebut dibandingkan dengan label sebenarnya untuk mengukur kinerja model (Rofiani, 2024).

2.3. Data Mining

Data mining merupakan suatu himpunan teknik dan proses sistematis yang digunakan untuk menemukan pola-pola tersembunyi, valid, baru, berguna dan dapat dimengerti dari kumpulan data dalam jumlah besar. Tujuan utama dari proses ini adalah untuk mengekstraksi informasi yang bermakna dan bernilai dari basis data guna mendukung pengambilan keputusan yang lebih tepat. Tahapan dalam data mining meliputi pemilihan data (data selection), pembersihan dan praproses data (data preprocessing), transformasi data, penerapan teknik data mining itu sendiri, serta interpretasi dan evaluasi hasil temuan pola. Teknik yang umum digunakan dalam data mining antara lain klasifikasi, clustering, asosiasi, dan regresi. Penerapan data mining sudah banyak digunakan dalam berbagai bidang, seperti penentuan penerima program bantuan sosial, peramalan penjualan,

analisis perilaku konsumen, hingga pengelompokan puas mahasiswa dalam dunia pendidikan (Surabaya, 2024).

Dalam pelaksanaannya, data mining memanfaatkan sejumlah algoritma populer, seperti algoritma C4.5 untuk klasifikasi berbasis pohon keputusan, algoritma Apriori untuk analisis asosiasi, algoritma K-Means untuk pengelompokan data, serta algoritina Naive Bayes untuk klasifikasi probabilistik pemilihan metode atau algoritma yang digunakan sangat bergantung pada jenis data dan tujuan analisis yang ingin dicapai. Data mining tidak hanya berperan sebagai alat bantu teknis, melainkan telah menjadi bagian penting dalam strategi pengambilan keputusan di berbagai sektor, termasuk bisnis, pendidikan, pemerintahan dan kesehatan. Melalui pendekatan yang tepat, data mining mampu menghasilkan wawasan yang strategis, meningkatkan efisiensi operasional, serta memberikan solusi konkret terhadap permasalahan yang kompleks (Aji, 2024).

2.3.1. Ciri-Ciri Data Mining

Data mining merupakan proses sistematis untuk memperoleh informasi yang berguna dari kumpulan data berskala besar. Teknik ini berakar dari disiplin ilmu Kecerdasan Buatan (Artificial Intelligence/Al) dan Rekayasa Pengetahuan (Knowledge Engineering/KE), serta memanfaatkan pendekatan statistik dan machine learning dalam pengolahan data. Tujuan utama dari data mining adalah melakukan klasifikasi, prediksi, perkiraan, serta mengekstraksi informasi tersembunyi yang bermanfaat dari data. Proses ini dapat menggunakan berbagai algoritma seperti C4.5 yang menghasilkan pohon keputusan yang mudah dipahami, algoritma Apriori untuk menemukan asosiasi antar item, serta algoritma

Support Vector Machine (SVM) dalam analisis sentimen. Data mining telah diaplikasikan secara luas dalam berbagai bidang seperti kesehatan, keuangan. lingkungan, dan pemerintahan. Seiring kemajuan perangkat keras, pengolahan data besar semakin efisien dan mendukung implementasi data mining. Secara umum, data mining merupakan salah satu tahapan penting dalam proses Knowledge Discovery in Database (KDD) dan melibatkan teknik-teknik utama seperti klasifikasi, regresi, clustering, dan asosiasi. (Sunarti., 2023)

Beberapa karakteristik utama data mining antara lain:

1. Proses Ekstraksi Informasi, yaitu mengambil data penting dari sumber tak terstruktur (teks, gambar, video) ke dalam bentuk yang terstruktur dan siap digunakan. Tahapan ekstraksi meliputi preprocessing, identifikasi entitas, ekstraksi relasi, normalisasi, dan evaluasi hasil.
2. Berbasis Algoritma, di mana algoritma menjadi dasar dalam memecahkan masalah dan membangun sistem otomatis berbasis data.
3. Skala Besar, karena data mining digunakan untuk mengolah data dengan volume besar yang cakupannya luas dan kompleks.
4. Multi-Disiplin, melibatkan kombinasi ilmu komputer, statistik, matematika, manajemen data, dan domain aplikasi lainnya untuk mengkaji permasalahan secara komprehensif.
5. Identifikasi Pola dan Tren, di mana proses ini membantu mengungkap kecenderungan, perubahan, serta struktur tersembunyi dalam data yang pada gilirannya dapat dijadikan dasar dalam pengambilan keputusan strategis.

2.4. Algoritma C4.5

Algoritma C4.5 merupakan salah satu metode klasifikasi yang banyak digunakan dalam data mining dan machine learning. Dikembangkan oleh Ross Quinlan sebagai penyempurnaan dari algoritma C4.5 memiliki kemampuan untuk menghasilkan pohon keputusan yang efisien, akurat, serta mampu menangani berbagai tipe atribut, baik kontinu maupun diskrit Salah satu fitur unggulan dari algoritma ini adalah proses pruning, yaitu pemangkasan cabang-cabang pohon yang tidak memberikan kontribusi signifikan terhadap peningkatan akurasi (Kamber, 2019).

Namun, meskipun C4.5 memiliki banyak keunggulan, sejumlah kelemahan tetap harus diperhatikan. Beberapa studi mencatat bahwa algoritma ini kurang efektif saat menangani data yang tidak seimbang (imbalanced data) atau yang memiliki overlapping antar kelas, sehingga berisiko menghasilkan model yang bias. C4.5 juga sensitif terhadap keberadaan noise dalam data, yang dapat mengakibatkan struktur pohon yang rumit dan kurang optimal. Selain itu, ketika dihadapkan pada dataset yang sangat besar dan mengandung banyak atribut serta kelas, algoritma ini cenderung menghasilkan kompleksitas tinggi dan waktu komputasi yang lama, sehingga efisiensinya menurun dalam skenario real-time. Untuk mengatasi hal tersebut, beberapa peneliti telah mengusulkan pendekatan optimasi seperti penggunaan teknik bagging atau integrasi dengan Particle Swarm Optimization (PSO) untuk meningkatkan performa algoritma. Oleh karena itu, meskipun C4.5 tetap menjadi algoritma yang kuat dan relevan dalam dunia klasifikasi data, pemahaman terhadap kelebihan dan keterbatasannya sangat

penting untuk memastikan implementasi yang efektif dan optimal dalam berbagai aplikasi (Supangat., Wisnu, B., & Rochman, 2024).

2.4.1. Proses Algoritma C4.5

Proses algoritma C4.5 diawali dengan tahap pembentukan pohon keputusan, di mana data dilatih menggunakan atribut dan label yang telah ditentukan untuk menghasilkan struktur pohon yang mampu merepresentasikan pola klasifikasi. Pemilihan atribut terbaik pada setiap node pohon dilakukan dengan menggunakan perhitungan Gain Ratio, yaitu rasio antara Information Gain dan entropi dari atribut tersebut. Pendekatan ini dipilih karena Gain Ratio mampu mengatasi kelemahan

dari Information Gain murni, terutama dalam menghadapi atribut dengan banyak nilai. Setelah pohon keputusan terbentuk, dilakukan tahap pemangkasan (pruning) guna mengurangi kompleksitas model dan mencegah terjadinya overfitting, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga kurang akurat saat digunakan pada data baru.

Tahap akhir dari proses algoritma C4.5 adalah klasifikasi, di mana pohon keputusan yang telah terbentuk digunakan untuk memprediksi atau mengklasifikasikan data baru berdasarkan jalur keputusan yang relevan dari akar hingga daun pohon. Terdapat tahapan yang perlu dilakukan pada Algoritma C4.5 yaitu sebagai berikut (Arifin, 2023).

a. Perhitungan Entropy dan Gain

1. Menghitung nilai entropy dari seluruh dataset untuk mengetahui tingkat ketidakpastian (disorder) dalam data.

2. Kemudian menghitung information gain untuk masing-masing atribut.
3. Setelah itu, menghitung Gain Ratio, yaitu rasio antara information gain dan split information, untuk menentukan atribut terbaik sebagai node keputusan.

b. Pembuatan Pohon Keputusan
   1. Atribut dengan Gain Ratio tertinggi dipilih sebagai node (akar) utama.
   2. Dataset kemudian dibagi berdasarkan nilai dari atribut tersebut, dan proses perhitungan berlanjut secara rekursif untuk setiap cabang hingga seluruh data terklasifikasi atau atribut habis.

c. Pemangkasan (Pruning)
   1. Setelah pohon selesai dibangun, dilakukan proses pruning untuk menyederhanakan struktur pohon.
   2. Tujuannya adalah mengurangi overfitting, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data latih dan tidak general terhadap data baru.

d. Klasifikasi
   1. Pohon yang telah terbentuk digunakan untuk mengklasifikasikan data baru.
   2. Proses klasifikasi dilakukan dengan menelusuri pohon berdasarkan nilai atribut pada data uji hingga mencapai simpul daun yang menyatakan kelas prediksi (Iswandaru, 2026).

2.4.2. Kelebihan Algoritma C4.5

Algoritma C4.5 memiliki sejumlah kelebihan yang menjadikannya salah satu metode klasifikasi yang banyak digunakan dalam data mining. Salah satu keunggulannya adalah kemampuannya untuk menangani data baik yang bersifat kategorikal maupun numerik, sehingga cocok diterapkan pada berbagai jenis atribut dalam dataset. Selain itu, C4.5 juga mampu mengatasi permasalahan data yang tidak lengkap (missing values) tanpa harus menghapus data tersebut, dengan memperkirakan distribusi nilai yang hilang secara proporsional. Dalam pemilihan atribut, algoritma ini menggunakan Gain Ratio yang merupakan pengembangan dari Information Gain pada algoritma sehingga mampu mengurangi bias terhadap atribut dengan banyak nilai. C4.5 juga dilengkapi dengan proses pemangkasan (pruning) untuk menyederhanakan struktur pohon keputusan dan menghindari overfitting, yang berarti model tetap bekerja dengan baik terhadap data baru. Keunggulan lainnya terletak pada hasil model yang berbentuk pohon keputusan yang mudah dipahami dan dijelaskan dalam bentuk aturan logis, sehingga memudahkan dalam interpretasi dan pengambilan keputusan berdasarkan hasil klasifikasi (Nima Rasekh, 2025).

Selain itu, kelebihan lain dari algoritma C4.5 terletak pada fleksibilitasnya dalam menangani dataset dengan banyak kelas target. Tidak seperti beberapa algoritma lain yang hanya bekerja optimal untuk klasifikasi biner, C4.5 mampu membagi data ke dalam lebih dari dua kelas tanpa memerlukan modifikasi khusus. Algoritma ini juga dapat menangani noise atau ketidakteraturan dalam data dengan cukup baik karena prinsip pemangkasan (pruning) yang digunakan

mampu mengurangi kompleksitas pohon dan mempertahankan hanya cabang-cabang yang signifikan. Keunggulan ini menjadikan C4.5 sangat berguna dalam penerapan dunia nyata, termasuk dalam sektor pelayanan publik seperti penelitian kepuasan masyarakat, karena algoritma ini tidak hanya memberikan hasil akurat, tetapi juga menyajikannya dalam bentuk yang mudah dimengerti oleh pihak non-teknis. Dengan demikian, C4.5 tidak hanya kuat secara analisis, tetapi juga unggul dalam hal keterbacaan dan kepraktisan hasil klasifikasinya (Nazifah, 2023).

2.4.3. Kekurangan Algoritma C4.5

Meskipun algoritma C4.5 memiliki banyak kelebihan, terdapat beberapa kekurangan yang perlu diperhatikan. Salah satu kekurangan utama adalah kompleksitas komputasinya yang cukup tinggi, terutama ketika digunakan pada

dataset yang sangat besar atau memiliki banyak atribut. Proses perhitungan gain ratio untuk setiap atribut dapat memakan waktu dan sumber daya yang cukup besar, yang bisa menjadi kendala dalam sistem dengan keterbatasan performa. Selain itu, algoritma ini sensitif terhadap data yang tidak seimbang (imbalanced data), di mana kelas mayoritas bisa mendominasi hasil klasifikasi dan menyebabkan penurunan akurasi terhadap kelas minoritas (Erie, 2022).

Kelemahan lainnya terletak pada cara algoritma C4.5 menangani data numerik, karena perlu dilakukan proses pembagian nilai secara berulang hingga ditemukan titik pemisah (threshold) terbaik, yang bisa meningkatkan risiko overfitting jika tidak disertai dengan proses pruning yang tepat. Selain itu, meskipun hasil klasifikasinya disajikan dalam bentuk pohon keputusan yang visual, struktur pohon dapat menjadi sangat besar dan rumit jika dataset memiliki

banyak atribut, sehingga menyulitkan interpretasi secara menyeluruh. Oleh karena itu, meskipun C4.5 merupakan algoritma yang andal, penggunaannya harus disesuaikan dengan karakteristik data dan tujuan analisis agar tetap memberikan hasil yang optimal.

2.4.4. Integrasi Analisis Kepuasan dengan Algoritma C4.5

Integrasi analisis kepuasan masyarakat dengan algoritma C4.5 dilakukan dengan memanfaatkan keunggulan algoritma ini dalam mengolah data kategori dan menghasilkan model klasifikasi yang mudah dipahami dalam bentuk pohon keputusan. Dalam konteks pelayanan publik, seperti di Cafe Castello. Algoritma C4.5 kemudian menganalisis setiap atribut tersebut untuk menentukan pola dan hubungan antara variabel yang memengaruhi keputusan masyarakat merasa puas atau tidak puas.

Proses ini diawali dengan tahap pelatihan model menggunakan data yang telah diklasifikasikan berdasarkan tingkat kepuasan responden. Algoritma C4.5 menghitung nilai gain dan gain ratio dari setiap atribut untuk menentukan atribut mana yang paling signifikan dalam membedakan kelas (puas atau tidak puas). Selanjutnya, model pohon keputusan dibentuk berdasarkan urutan atribut yang memiliki pengaruh terbesar, sehingga setiap cabang dalam pohon menunjukkan jalur keputusan yang logis berdasarkan data yang tersedia. Dengan cara ini, analisis tidak hanya menghasilkan prediksi, tetapi juga memberikan gambaran faktor-faktor dominan yang perlu diperbaiki oleh pihak Caffe Castello.

Hasil dari integrasi ini memungkinkan instansi pemerintah untuk mengambil keputusan yang lebih tepat sasaran berdasarkan data lapangan.

Melalui visualisasi pohon keputusan, pihak pengelola pelayanan dapat lebih mudah memahami faktor mana yang paling berpengaruh terhadap kepuasan masyarakat, dan tindakan apa yang perlu diambil untuk meningkatkan kualitas layanan. Selain itu, model klasifikasi yang dihasilkan juga dapat digunakan untuk memantau dan mengevaluasi kepuasan secara berkala, serta sebagai dasar pengambilan kebijakan berbasis data dalam upaya mewujudkan pelayanan publik yang lebih responsif dan efisien di masa depan (Fauzi, 2022).

2.4.5. Langkah-Langkah Algoritma C4.5

Algoritma C4.5 merupakan salah satu metode klasifikasi berbasis pohon keputusan yang populer dalam data mining karena kemampuannya menangani data kategorikal maupun numerik, serta menghasilkan model yang mudah diinterpretasikan. C4.5 bekerja dengan membagi dataset ke dalam subset berdasarkan atribut yang memberikan informasi paling besar (information gain) dan memperhitungkan gain ratio untuk menghindari bias terhadap atribut dengan banyak nilai unik. Dengan membangun struktur pohon secara rekursif, algoritma ini mampu mengidentifikasi pola tersembunyi dalam data dan menyederhanakan proses pengambilan keputusan berdasarkan logika yang sistematis. (D. Telaumbanua & I. Kurniawati, 2022)

Berikut ini adalah langkah-langkah dalam algoritma C4.5

a. Menentukan atribut target, tentukan atribut mana yang akan dijadikan sebagai label kelas (target) dalam proses klasifikasi.

b. Menghitung entropy semua atribut, hitung entropy dari setiap atribut dalam dataset untuk mengukur tingkat ketidakpastian atau keragaman nilai pada masing-masing atribut.
c. Menghitung gain dan gain ratio, hitung nilai information gain dari masing-masing atribut, lalu hitung gain ratio untuk menyesuaikan pengaruh atribut yang memiliki banyak nilai unik.
d. Menentukan Atribut Terbaik Sebagai Node, Pilih atribut dengan gain ratio tertinggi sebagai node (akar atau cabang) dari pohon keputusan.
e. Membagi Data Menjadi Subset, Bagi data berdasarkan nilai dari atribut terbaik yang telah dipilih, dan buat cabang pohon keputusan berdasarkan nilai-nilai tersebut.
f. Rekursi untuk Setiap Subset, Ulangi proses dari langkah pertama untuk setiap subset hingga seluruh data pada cabang tersebut memiliki kelas yang sama atau tidak ada lagi atribut yang bisa dipilih.
g. Pemangkasan (Pruning), Setelah pohon selesai dibentuk, lakukan pemangkasan untuk menghilangkan cabang yang tidak signifikan atau menyebabkan overfitting, agar pohon lebih sederhana dan generalisasi model menjadi lebih baik.

Langkah-langkah ini menjadikan algoritma C4.5 sebagai alat yang efektif untuk mengklasifikasikan data serta memberikan pemahaman yang jelas terhadap pola pola yang tersembunyi dalam kumpulan data besar.

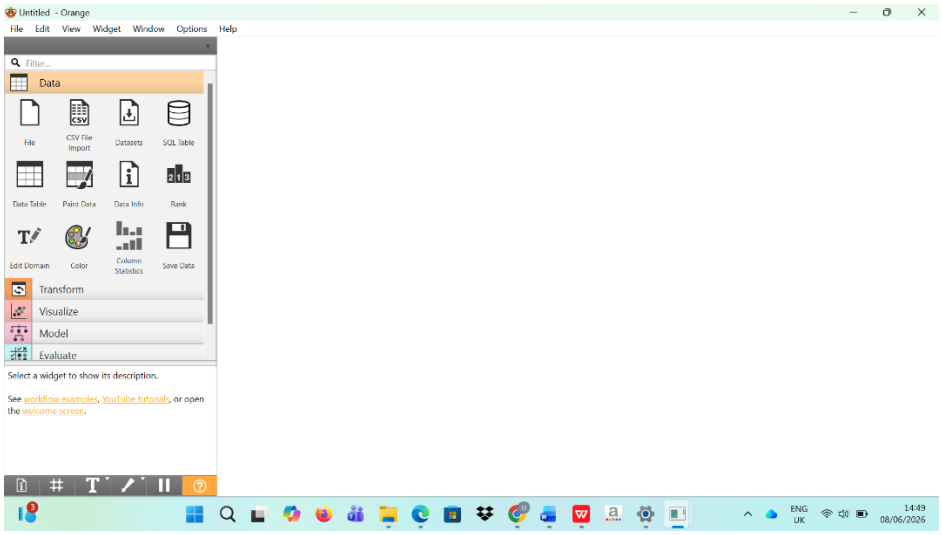
2.4.6. Evaluasi Model Algoritma C4.5

Evaluasi model pada algoritma C4.5 merupakan proses penting untuk menilai sejauh mana model yang dibangun mampu melakukan klasifikasi data secara akurat dan andal. Evaluasi ini biasanya dilakukan dengan menggunakan metrik seperti akurasi, presisi, dan recall yang dihasilkan dari confusion matrix. Akurasi menunjukkan seberapa besar persentase prediksi yang benar dari keseluruhan data, presisi menggambarkan seberapa tepat model dalam memprediksi kelas positif, sedangkan recall mengukur kemampuan model dalam menemukan seluruh data yang termasuk dalam kelas positif. Selain itu, evaluasi dapat menggunakan metode validasi silang (cross-validation) atau membagi data ke dalam set pelatihan dan pengujian, untuk memastikan bahwa model tidak hanya bekerja baik pada data pelatihan, tetapi juga dapat digeneralisasi pada data baru (Chotimah, 2024).

Dalam penelitian mengenai tingkat kepuasan masyarakat terhadap pelayanan di Cafe Castello, evaluasi model C4.5 dilakukan terhadap data yang telah diklasifikasikan berdasarkan atribut-atribut seperti kecepatan pelayanan, kemudahan prosedur, sikap petugas, kebersihan, dan ketersediaan fasilitas. Hasil evaluasi menunjukkan bahwa model memiliki akurasi yang tinggi, menunjukkan bahwa pohon keputusan yang dibentuk mampu mengidentifikasi pola dari data dengan baik. Hal ini membuktikan bahwa C4.5 efektif digunakan dalam menganalisis data survei masyarakat, serta memberikan kontribusi yang signifikan dalam mendukung proses pengambilan keputusan guna meningkatkan kualitas pelayanan publik di wilayah tersebut (Sulaeman, 2025).

2.5. Alat Bantu Program Aplikasi Orange

Aplikasi Orange merupakan salah satu alat bantu analisis data berbasis grafis yang sangat populer dalam dunia data mining dan machine learning. Program ini bersifat open-source dan menyediakan antarmuka visual yang memudahkan pengguna, terutama yang tidak memiliki latar belakang pemrograman, untuk membangun dan menguji berbagai model analisis data. (Musu, A., Parodi, 2025). Orange dilengkapi dengan berbagai widget yang dapat digunakan untuk melakukan proses seperti preprocessing data, visualisasi, klasifikasi, regresi, clustering, dan evaluasi model. Keunggulan utama Orange adalah kemudahan drag-and-drop dalam menyusun alur kerja analisis data, serta tersedianya berbagai algoritma populer seperti Decision Tree, Naive Bayes, KNN, dan lainnya yang dapat digunakan secara instan (Goren, 2025).



Gambar 2. 2. Alat Bantu Program Aplikasi Orange

Dalam penelitian mengenai analisis tingkat kepuasan masyarakat terhadap pelayanan di Cafe Castello, aplikasi Orange digunakan sebagai alat bantu utama dalam membangun dan mengevaluasi model klasifikasi berbasis algoritma C4.5

Decision Tree. Penggunaan Orange memungkinkan peneliti untuk menyusun alur kerja analisis secara sistematis, mulai dari input data, proses preprocessing, pembuatan model klasifikasi, hingga evaluasi hasil dengan lebih efisien dan akurat. Dengan tampilan visual yang intuitif, Orange membantu peneliti untuk memahami pola-pola dalam data masyarakat yang memengaruhi tingkat kepuasan, serta mempermudah dalam penyajian hasil analisis dalam bentuk pohon keputusan yang mudah dipahami oleh pihak terkait (Bouhrine, 2025).

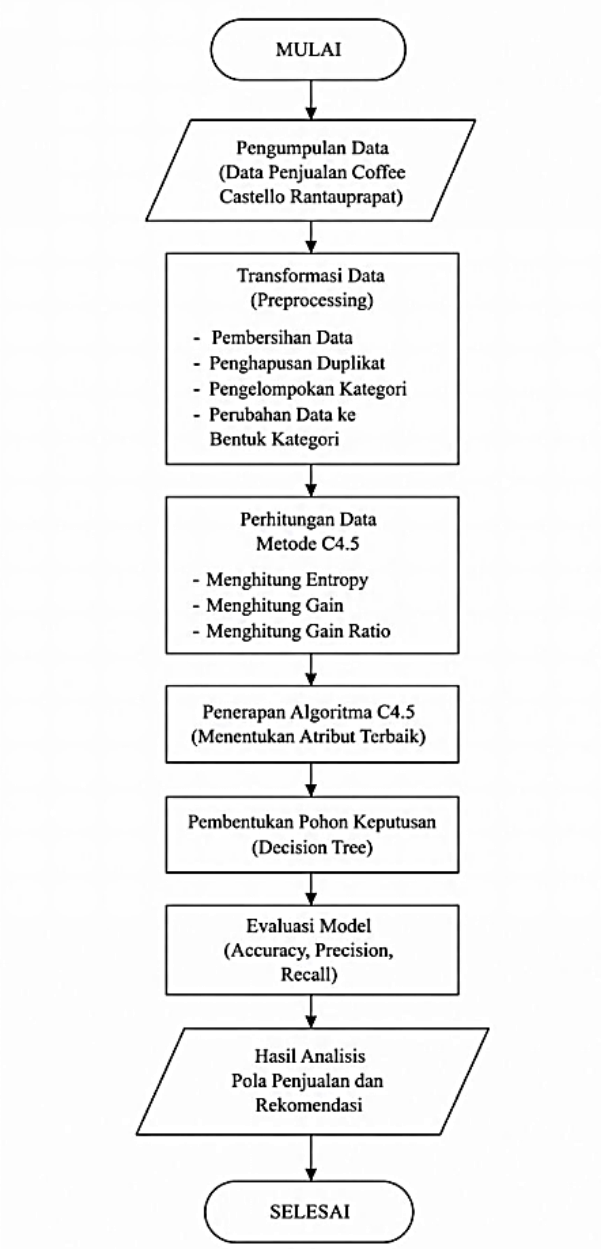
2.6. Penelitian Terdahulu

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian	Perbedaan dengan Penelitian Anda
1	Bartolomeus Wisnu Setyo Wibowo Sunadi (2024)	Analisis Tingkat Keberhasilan Penjualan dengan Metode C4.5 Studi Kasus Pengambilan Keputusan Bisnis	Menghasilkan 100 aturan dari pohon keputusan dan mengidentifikasi harga sebagai faktor utama yang mempengaruhi negosiasi penjualan B2B, dengan akurasi model sebesar 74%.	Objek penelitian adalah perusahaan B2B, fokus pada negosiasi, dan jumlah data yang digunakan sangat besar (1604 record). Penelitian Anda berfokus pada bisnis ritel (kedai kopi) dan menggunakan data yang lebih kecil.
2	Penelitian pada Perusahaan Ritel (2024)	Penerapan Algoritma C4.5 dan Naïve Bayes untuk Prediksi Status Penjualan	Penelitian ini menemukan bahwa harga adalah prediktor terkuat (Gain Ratio 0,93) dan model mencapai akurasi sempurna (100%) menggunakan C4.5.	Perbedaan utama terletak pada temuan atribut utama. Penelitian Anda menemukan bahwa Total Penjualan adalah atribut paling dominan

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian	Perbedaan dengan Penelitian Anda
				berdasarkan perhitungan gain pada skripsi Anda, berbeda dengan temuan di penelitian ini yang menempatkan harga sebagai faktor penentu.
3	Penelitian pada Perusahaan Ritel (2024)	Classification of Product Predicates Based on Sales Rate Using the C4.5 Decision Tree Algorithm in Retail Companies	Algoritma Decision Tree C4.5 efektif mengidentifikasi predikat penjualan produk dengan akurasi model 96,20%.	Penelitian ini dilakukan di perusahaan ritel secara umum dan menggunakan platform RapidMiner. Penelitian Anda menggunakan aplikasi Orange dan berfokus pada satu kedai kopi spesifik.
4	Penelitian pada HERA Promotion (2025)	Klasifikasi Data Penjualan Produk Hera Promotion Menggunakan Algoritma C4.5	C4.5 efektif mengklasifikasikan produk laris dan tidak laris dengan akurasi data uji 99,48%. Faktor harga, jenis produk, dan waktu adalah yang paling berpengaruh.	Atribut yang dianggap paling berpengaruh berbeda. Dalam penelitian Anda, Jumlah Terjual dan Total Penjualan adalah faktor utama setelah Total Penjualan, dan harga memiliki pengaruh yang lebih rendah (Gain 0,038).

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian	Perbedaan dengan Penelitian Anda
5	Penelitian pada UD. Mr Zen Agro (2024)	Analisis Algoritma C4.5 dalam Memprediksi Penjualan Pupuk dan Pestisida di UD. Mr Zen Agro	Algoritma C4.5 digunakan untuk memprediksi penjualan dan mengidentifikasi produk yang diminati atau tidak, yang diimplementasikan dalam sebuah sistem aplikasi.	Objek penelitian bergerak di bidang pertanian (pupuk dan pestisida) dan hasilnya diimplementasikan ke dalam sistem berbasis PHP dan MySQL.
6	Liza Indarwati (2025)	Prediksi Penjualan Pestisida Pada PT Sagri Menggunakan Algoritma C4.5	Mampu mengklasifikasikan data penjualan pestisida ke dalam kategori laris dan tidak laris dengan tingkat akurasi yang memadai, menggunakan atribut seperti penggunaan, nama, harga, musim, dan volume.	Meskipun sama-sama menggunakan algoritma C4.5, objek penelitian adalah perusahaan pestisida (sektor pertanian) yang memiliki karakteristik atribut dan data yang sangat berbeda dengan bisnis kedai kopi.

2.7. Kerangka Kerja Penelitian



Gambar 2. 3. Flowchart Penelitian

Berikut adalah penjelasan dari masing masing poin yang ada pada kerangka kerja penelitian sebagai berikut.

1. Pengumpulan data adalah tahap awal penelitian yang dilakukan dengan mengumpulkan data penjualan pada Coffee Castello Rantauprapat. Data yang digunakan berupa data transaksi penjualan, seperti nama menu, jumlah pembelian, kategori produk, dan waktu transaksi. Data tersebut diperoleh dari arsip atau laporan penjualan yang nantinya digunakan sebagai bahan utama dalam proses analisis menggunakan metode C4.5.
2. Transformasi data (preprocessing) adalah tahap pengolahan awal data sebelum dilakukan proses perhitungan. Pada tahap ini dilakukan pembersihan data, penghapusan data yang duplikat, pengelompokan data ke dalam kategori tertentu, serta perubahan data ke bentuk yang lebih sesuai agar dapat diproses menggunakan algoritma C4.5.
3. Perhitungan data metode C4.5 adalah tahap perhitungan nilai entropy, gain, dan gain ratio dari setiap atribut data penjualan. Perhitungan ini dilakukan untuk menentukan atribut yang paling berpengaruh dalam proses klasifikasi sehingga dapat digunakan dalam pembentukan pohon keputusan.
4. Penerapan algoritma C4.5 adalah tahap penggunaan metode C4.5 untuk melakukan klasifikasi data penjualan berdasarkan hasil perhitungan sebelumnya. Pada tahap ini atribut dengan nilai gain tertinggi akan dipilih sebagai node utama dalam pembentukan aturan keputusan.

5. Pembentukan pohon keputusan (decision tree) adalah tahap penyusunan struktur keputusan berdasarkan hasil penerapan algoritma C4.5. Pohon keputusan digunakan untuk menggambarkan hubungan antar atribut data sehingga pola penjualan pada Coffee Castello Rantauprapat dapat diketahui dengan lebih jelas.
6. Evaluasi model adalah tahap pengujian hasil klasifikasi yang diperoleh dari metode C4.5. Evaluasi dilakukan menggunakan pengukuran seperti accuracy, precision, dan recall untuk mengetahui tingkat kinerja model dalam melakukan klasifikasi data penjualan.
7. Hasil analisis pola penjualan dan rekomendasi adalah tahap akhir penelitian yang berisi hasil analisis terhadap pola penjualan yang diperoleh dari proses klasifikasi. Hasil tersebut kemudian digunakan sebagai dasar dalam memberikan rekomendasi guna membantu Coffee Castello Rantauprapat dalam meningkatkan strategi penjualan dan pengambilan keputusa