BAB II

LANDASAN TEORI

2.1. Data Science

Data Science adalah bidang yang berkembang pesat yang menggabungkan elemen-elemen dari statistik, ilmu komputer, dan pengetahuan domain untuk mengekstrak wawasan dan pengetahuan dari data. Tujuan utama dari Data Science adalah untuk mengolah, menganalisis, dan menginterpretasikan data untuk membuat keputusan yang lebih baik dalam konteks bisnis atau penelitian.

Salah satu langkah awal dalam proses Data Science adalah pengumpulan data, yang dapat berasal dari berbagai sumber seperti sensor, platform media sosial, atau basis data perusahaan. Setelah data dikumpulkan, langkah selanjutnya adalah membersihkan dan mempersiapkannya untuk analisis. Proses ini melibatkan identifikasi dan penanganan nilai-nilai yang hilang, outlier, dan masalah lainnya yang dapat mempengaruhi hasil analisis.

Analisis data merupakan inti dari Data Science, dan metode yang digunakan dapat mencakup statistik deskriptif, pengujian hipotesis, machine learning, dan teknik lainnya. Dengan meLaki-Lakifaatkan algoritma dan model matematis, Data Science memungkinkan identifikasi pola, tren, dan hubungan yang mungkin sulit atau bahkan tidak mungkin ditemukan dengan pendekatan tradisional.

Visualisasi data juga menjadi aspek penting dalam Data Science. Grafik dan visualisasi membantu menyajikan informasi dengan cara yang mudah dimengerti dan membantu para peLaki-Lakigku kepentingan untuk mengambil keputusan yang terinformasi. Selain itu, Data Science juga melibatkan pengembangan solusi

berbasis teknologi untuk memproses dan menganalisis data secara efisien. Ini dapat mencakup penggunaan teknologi cloud, pengolahan paralel, dan infrastruktur lainnya.

Data Science memiliki peran yang krusial dalam transformasi digital dan telah menjadi landasan bagi inovasi di berbagai sektor, termasuk bisnis, kesehatan, keuangan, dan banyak lagi. Dengan meningkatnya jumlah data yang dihasilkan setiap hari, Data Science menjadi semakin relevan dalam membantu organisasi membuat keputusan yang lebih cerdas dan responsif terhadap perubahan di lingkungan mereka.

2.1.1. Data Mining

Data Mining adalah proses ekstraksi pola yang berharga atau pengetahuan yang tersembunyi dari sejumlah besar data [1]. Tujuannya adalah untuk mengidentifikasi hubungan yang belum diketahui atau pola tersembunyi yang dapat memberikan wawasan bisnis atau ilmiah yang berharga [2] [3]. Data Mining melibatkan penggunaan teknik dan algoritma dari bidang statistik, pembelajaran mesin, kecerdasan buatan, dan database untuk menggali informasi dari dataset yang luas dan kompleks [4].

Salah satu aspek penting dari Data Mining adalah penggunaan algoritma untuk mengenali pola dalam data [5]. Ini dapat mencakup pengelompokan (clustering) data, klasifikasi, regresi, dan asosiasi. Clustering memungkinkan pengelompokan data ke dalam kelompok-kelompok yang serupa, sementara klasifikasi melibatkan pengenalan pola untuk mengategorikan data ke dalam kelas atau kelompok tertentu. Regresi digunakan untuk memodelkan hubungan antara

variabel, dan asosiasi digunakan untuk mengidentifikasi hubungan yang mungkin terjadi secara bersamaan dalam dataset.

Data Mining sering kali digunakan dalam kombinasi dengan teknik lain seperti preprocessing data, penggalian informasi, dan visualisasi data untuk memastikan hasil yang akurat dan bermakna. Preprocessing data melibatkan pembersihan dan transformasi data untuk memastikan keakuratannya sebelum proses penggalian dimulai. Penggalian informasi melibatkan identifikasi pola yang berkaitan dengan tujuan tertentu, dan visualisasi data membantu menyajikan temuan secara intuitif.

Penerapan Data Mining dapat ditemukan di berbagai industri, seperti perbankan, ritel, kesehatan, dan lainnya. Contoh penggunaannya termasuk analisis perilaku konsumen, deteksi penipuan keuangan, prediksi tren pasar, dan diagnosis penyakit berdasarkan data medis. Data Mining memberikan nilai tambah dengan memungkinkan organisasi mengambil keputusan berdasarkan pemahaLaki-Laki yang lebih dalam terhadap data mereka, membantu mengidentifikasi peluang, dan meningkatkan efisiensi operasional.

2.1.2. Database dan Data Processing

Database adalah kumpulan data yang terorganisir secara sistematis dan disimpan dalam suatu sistem komputer. Tujuan utama dari database adalah menyediakan cara efisien untuk menyimpan, mengelola, dan mengakses data. Databases dapat digunakan untuk berbagai tujuan, mulai dari mendukung operasi bisnis sehari-hari hingga mendukung aplikasi perangkat lunak yang kompleks. Mereka menyediakan struktur yang terdefinisi untuk menyimpan data,

memungkinkan pengguna untuk melakukan operasi seperti penambahan, penghapusan, dan pencarian data dengan mudah. Jenis database yang umum meliputi database relasional, database NoSQL, dan berbagai model database lainnya.

Dalam konteks bisnis, database sangat penting karena mereka menyediakan fondasi untuk sistem informasi yang mendukung pengelolaan data perusahaan. Database relasional, misalnya, menggunakan tabel untuk menyimpan data dan memungkinkan pengguna untuk membuat hubungan antara data di berbagai tabel. Hal ini membantu memastikan konsistensi, integritas, dan keaLaki-Lakian data. Sistem Laki-Lakiajemen basis data (DBMS) adalah perangkat lunak yang digunakan untuk mengelola database, menyediakan antarmuka untuk pengguna dan aplikasi untuk berinteraksi dengan data.

Data Processing merujuk pada serangkaian tindakan atau proses untuk mentransformasikan data mentah menjadi informasi yang dapat dimengerti dan berLaki-Lakifaat. Proses ini melibatkan pengumpulan, validasi, penyimpanan, pemrosesan, dan distribusi data. Ada dua jenis utama dari Data Processing: pemrosesan batch dan pemrosesan real-time. Pemrosesan batch melibatkan pengumpulan data dalam jumlah besar dan pemrosesan mereka secara bersamaan, sementara pemrosesan real-time memproses data secara segera setelah diterima.

Teknologi modern telah mengubah cara data diproses dengan memperkenalkan komputasi awan, big data, dan analisis data tingkat tinggi. Big data processing, misalnya, memungkinkan organisasi untuk mengelola, menyimpan, dan menganalisis volume data yang sangat besar dengan kecepatan

yang tinggi. Analisis data juga menjadi bagian integral dari pemrosesan data, memungkinkan organisasi untuk mendapatkan wawasan berharga dari data mereka. Data Processing juga mencakup konsep ETL (Extract, Transform, Load), yang digunakan untuk mentransfer data dari satu lokasi atau format ke lokasi atau format lainnya. ETL adalah bagian penting dari proses pengolahan data, terutama dalam konteks data warehouse dan analisis bisnis.

2.1.3. Visualisation

Visualisasi adalah suatu proses representasi data dan informasi dalam bentuk grafis atau visual untuk memudahkan pemahaLaki-Laki dan analisis. Tujuan utama dari visualisasi adalah menyajikan informasi kompleks dengan cara yang mudah dimengerti, memungkinkan pengguna untuk melihat pola, tren, dan hubungan yang mungkin sulit dilihat dalam bentuk data mentah. Dengan meLaki-Lakifaatkan elemen visual seperti grafik, diagram, peta, dan visualisasi lainnya, visualisasi data membantu mendukung pengambilan keputusan yang lebih baik dan komunikasi efektif.

Visualisasi sangat penting dalam dunia data science dan analisis data. Melalui visualisasi, data yang kompleks dapat diubah menjadi narasi yang lebih sederhana, memungkinkan peLaki-Lakigku kepentingan untuk membuat keputusan informasi berdasarkan pemahaLaki-Laki yang lebih baik. Grafik dan visualisasi memainkan peran penting dalam mengungkapkan pola dalam data, memberikan konteks terhadap perubahan seiring waktu, dan membantu dalam identifikasi anomali atau tren yang signifikan.

Berbagai jenis visualisasi dapat digunakan tergantung pada jenis data dan informasi yang disajikan. Beberapa contoh termasuk grafik garis untuk melihat tren sepanjang waktu, diagram batang untuk membandingkan kategori, dan peta panas untuk menunjukkan sebaran intensitas data di suatu wilayah. Teknologi visualisasi data terus berkembang, termasuk penggunaan augmented reality (AR) dan virtual reality (VR) untuk menciptakan pengalaLaki-Laki visual yang lebih immersif.

Selain itu, visualisasi data juga memainkan peran penting dalam penyampaian informasi kepada audiens yang mungkin tidak memiliki pengetahuan teknis yang mendalam tentang data. Presentasi visual yang efektif dapat mengomunikasikan temuan analisis secara lebih persuasif dan menghasilkan dampak yang lebih besar. Dalam dunia bisnis, visualisasi data digunakan dalam berbagai konteks, mulai dari laporan keuangan hingga dashboard kinerja operasional. PemahaLaki-Laki yang mendalam melalui visualisasi data memungkinkan pemimpin organisasi dan pengambil keputusan untuk merespons secara lebih cepat terhadap perubahan dan membuat keputusan yang lebih cerdas.

2.1.4. Statistik

Statistik adalah cabang ilmu matematika yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan pengorganisasian data. Tujuan utama dari statistik adalah untuk menyajikan dan menganalisis informasi yang dapat memberikan wawasan atau mendukung pengambilan keputusan. Dalam prosesnya, statistik membantu kita memahami variasi, mengidentifikasi pola, dan

mengambil kesimpulan berdasarkan sampel data yang diambil dari populasi yang lebih besar.

Ada dua jenis statistik utama: statistik deskriptif dan statistik inferensial. Statistik deskriptif digunakan untuk merangkum dan menggambarkan karakteristik dasar dari suatu dataset. Ini mencakup penggunaan ukuran tendensi sentral seperti mean (rata-rata), median (nilai tengah), dan modus (nilai yang sering muncul), serta ukuran variasi seperti kisaran, simpangan baku, dan kuartil.

Statistik inferensial, di sisi lain, melibatkan membuat inferensi atau perkiraan tentang suatu populasi berdasarkan data yang diambil dari sampel. Ini melibatkan penggunaan konsep probabilitas dan teknik seperti uji hipotesis dan interval kepercayaan. Statistik inferensial memungkinkan kita untuk membuat generalisasi yang lebih luas atau mengambil kesimpulan tentang populasi berdasarkan data yang terbatas.

Statistik memiliki berbagai aplikasi di berbagai bidang, termasuk ilmu sosial, ekonomi, kedokteran, sains alam, dan bisnis. Dalam dunia bisnis, statistik digunakan untuk analisis pasar, peramalan penjualan, pengendalian kualitas, dan pengambilan keputusan berbasis data. Di bidang ilmiah, statistik digunakan untuk menguji hipotesis, mengukur signifikansi hasil eksperimen, dan menyimpulkan generalisasi tentang fenomena alam.

Dalam era digital dan big data, statistik berperan penting dalam menganalisis dataset besar dan kompleks. Teknik-teknik seperti machine learning dan analisis regresi linear merupakan contoh penggunaan statistik dalam mengidentifikasi pola atau membuat prediksi berdasarkan data besar. Dengan

kemajuan teknologi, statistik terus beradaptasi untuk memberikan wawasan yang lebih dalam dan relevan dari data yang semakin kompleks.

2.1.5. Pattern Recognition

Pattern Recognition, atau Pengenalan Pola, adalah cabang ilmu yang berkaitan dengan identifikasi, interpretasi, dan pengklasifikasian pola dalam data. Tujuannya adalah untuk mengembangkan model atau algoritma yang dapat memahami struktur dalam data dan mengenali pola yang dapat memberikan informasi berharga. Ini dapat diterapkan dalam berbagai konteks, termasuk pengolahan citra, pengenalan suara, pengenalan tulisan tangan, dan bidang-bidang lain yang melibatkan analisis data kompleks.

Dalam pengenalan pola, komputer diajarkan untuk mengenali pola-pola ini melalui proses pelatihan menggunakan dataset yang telah diberi label atau kategori. Algoritma pembelajaran mesin, seperti neural networks dan decision trees, sering digunakan dalam pengenalan pola untuk menghasilkan model yang dapat membedakan antara pola yang berbeda atau mengklasifikasikan data ke dalam kategori tertentu.

Salah satu aplikasi utama dari pengenalan pola adalah dalam pengolahan citra, di Laki-Lakia sistem dapat diajarkan untuk mengenali objek atau wajah dalam gambar. Selain itu, di bidang medis, pengenalan pola digunakan untuk mendiagnosis penyakit atau mengenali pola abnormal dalam data medis seperti citra pencitraan medis.

Keberhasilan pengenalan pola tergantung pada kemampuannya untuk mengekstrak fitur yang relevan dari data. Ini dapat mencakup pengenalan bentuk,

tekstur, warna, dan atribut lain yang membantu dalam identifikasi pola. Meskipun teknik ini sangat berguna, pengenalan pola juga dapat menghadapi tantangan, terutama ketika data sangat kompleks atau ketika pola tersebut sulit untuk dibedakan secara jelas.

Pengenalan pola berperan penting dalam era digital dan big data, di Laki-Lakia jumlah data yang besar memerlukan pendekatan otomatis untuk mengidentifikasi dan memahami pola-pola yang dapat memberikan wawasan berharga. Dalam konteks ini, pengenalan pola terus berkembang untuk mengatasi tantangan baru dan meLaki-Lakifaatkan potensi besar yang terdapat dalam analisis data yang lebih canggih.

2.2. Model Klasifikasi

Model klasifikasi adalah suatu bentuk dari algoritma pembelajaran mesin yang digunakan untuk mengelompokkan atau mengkategorikan suatu data ke dalam kelas atau kategori yang telah ditentukan [6] [7]. Tujuannya adalah untuk membangun suatu fungsi atau model yang dapat memetakan input data ke dalam output kelas dengan tingkat akurasi yang tinggi [8]. Model klasifikasi meLaki-Lakifaatkan data pelatihan yang telah diberi label untuk belajar pola dan membuat prediksi pada data baru yang belum terlihat sebelumnya.

Salah satu jenis model klasifikasi yang umum adalah *logistic regression*, yang digunakan ketika variabel target atau output adalah biner, yaitu hanya memiliki dua kelas. Metode ini menciptakan kurva logistik yang memetakan input ke probabilitas terjadinya suatu kejadian. Model klasifikasi lainnya termasuk decision trees, support vector machines, k-nearest neighbors, dan random forest,

yang memiliki kompleksitas yang bervariasi dan cocok untuk situasi atau jenis data tertentu [9].

Pada dasarnya, proses pembangunan model klasifikasi melibatkan beberapa langkah, termasuk pengumpulan data, pemrosesan data, pemilihan fitur, pelatihan model, evaluasi model, dan fine-tuning. Data pelatihan digunakan untuk mengajarkan model pola dan hubungan antara fitur-fitur input dan kelas output. Setelah model dilatih, kemudian diuji menggunakan data yang tidak pernah dilihat sebelumnya untuk menilai kinerjanya dan memastikan bahwa model tersebut dapat digeneralisasikan dengan baik ke data baru.

Model klasifikasi memiliki banyak aplikasi dalam berbagai bidang. Contohnya termasuk klasifikasi email sebagai spam atau bukan spam, identifikasi kategori produk dalam perdagangan elektronik, deteksi fraud di keuangan, dan diagnosis medis. Keberhasilan model klasifikasi sangat tergantung pada kualitas dan representativitas data pelatihan, serta kebijakan pemilihan fitur yang cerdas dan pemilihan model yang sesuai dengan sifat data yang dihadapi. Dalam konteks pembelajaran mesin, model klasifikasi menjadi alat penting dalam membuat keputusan otomatis berdasarkan data.

2.3. Algoritma C4.5

Algoritma C4.5, yang dikembangkan oleh Ross Quinlan, adalah salah satu algoritma pembelajaran mesin yang digunakan untuk membangun model pohon keputusan. Pohon keputusan adalah struktur hirarkis yang memetakan input ke output dengan cara yang intuitif dan mudah dimengerti. Algoritma C4.5 berfokus pada pembangunan pohon keputusan untuk tugas klasifikasi, di Laki-Lakia

tujuannya adalah untuk memprediksi kelas atau label dari suatu data berdasarkan atribut-atribut yang ada.

Proses algoritma C4.5 dimulai dengan memilih atribut terbaik untuk membagi data menjadi kelompok yang paling homogen. Pengukuran homogenitas ini diukur dengan menggunakan metrik seperti Gain Ratio atau Information Gain yang berkaitan dengan teori informasi. Pemilihan atribut ini dilakukan secara rekursif untuk setiap simpul dalam pohon, menciptakan struktur pohon keputusan yang optimal untuk memaksimalkan akurasi klasifikasi.

Salah satu keunggulan utama dari algoritma C4.5 adalah kemampuannya untuk menangani data yang memiliki nilai yang hilang atau tidak lengkap. Algoritma ini juga mampu menangani data yang memiliki atribut dengan tipe data yang berbeda. Proses pembangunan pohon keputusan C4.5 dilakukan dengan mempertimbangkan faktor-faktor seperti homogenitas kelompok dan efisiensi pemilihan atribut, sehingga menghasilkan model yang mudah diinterpretasi dan mampu bekerja dengan baik pada dataset yang kompleks.

Meskipun algoritma C4.5 memiliki kelebihan, seperti kemampuannya dalam menangani data yang tidak terstruktur, namun, seperti algoritma pembelajaran mesin lainnya, ia juga memiliki keterbatasan. Algoritma ini rentan terhadap overfitting, terutama pada dataset yang kecil dan kompleks. Oleh karena itu, teknik pruning (peLaki-Lakigkasan) sering digunakan untuk mengurangi kompleksitas pohon dan meningkatkan generalisasi model.

C4.5 telah menjadi dasar bagi banyak pengembangan algoritma pembelajaran mesin lainnya dan pohon keputusan dalam beberapa bentuk, seperti

CART (Classification and Regression Trees). Meskipun telah ada perkembangan lebih lanjut dalam algoritma pembelajaran mesin, C4.5 tetap menjadi salah satu pendekatan yang penting dan memberikan fondasi yang kuat untuk pemahaLaki-Laki pohon keputusan dan konsep pembelajaran mesin yang lebih luas.

2.4. Alat Bantu Program/Tools Pendukung

2.4.1. Orange

Orange adalah platform open-source yang memfasilitasi analisis data, eksplorasi pola, dan pembelajaran mesin melalui antarmuka visual [10] [11]. Dikembangkan di University of Ljubljana, Slovenia, Orange memberikan kemampuan analisis data tanpa memerlukan keahlian pemrograLaki-Laki yang mendalam. Platform ini menawarkan berbagai widget dan komponen grafis yang memungkinkan pengguna menyusun alur kerja analisis data dengan mudah.

Salah satu keunggulan Orange adalah fokusnya pada visualisasi data yang interaktif [12]. Pengguna dapat membangun alur kerja dengan menyusun dan menghubungkan widget yang mencakup fungsi pemrosesan data, visualisasi, ekstraksi fitur, dan pembangunan model. Orange juga menyediakan beragam algoritma pembelajaran mesin yang dapat diakses melalui antarmuka visualnya, memungkinkan pengguna untuk memilih, mengkonfigurasi, dan menganalisis model tanpa harus menulis kode [13].

Platform ini mendukung sejumlah tugas analisis data, termasuk klasifikasi, regresi, klastering, dan analisis asosiasi. Orange juga memberikan fleksibilitas kepada pengguna untuk mengeksplorasi data, mengidentifikasi pola, dan mendapatkan wawasan mendalam tentang karakteristik dataset mereka.

Keberhasilan Orange sebagian besar didukung oleh komunitas yang aktif dan dukungan dari tim pengembangnya. Adanya fitur penyimpanan dan berbagi workflow memungkinkan kolaborasi antara pengguna, sementara platform ini terus berkembang dengan ditambahkannya fitur-fitur baru.

Dengan menyediakan alat visual yang kuat, Orange telah memainkan peran penting dalam membuat analisis data dan pembelajaran mesin lebih mudah diakses oleh berbagai kalangan, dari peneliti hingga ilmuwan data dan praktisi di berbagai industri.

2.5. Metodologi Penelitian

2.5.1. Penelitian Terdahulu

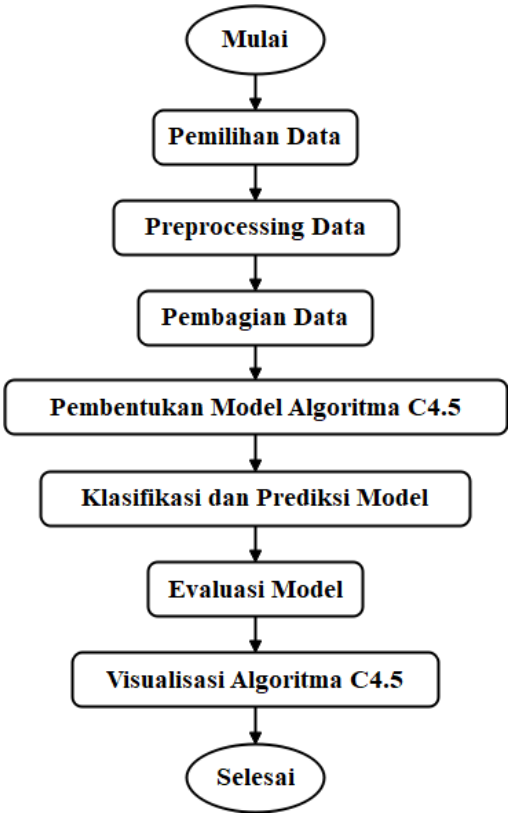
Referensi Penelitian	1
Judul	Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5
Nama	Sanni Ucha Putri1, Eka Irawan2, Fitri Rizky3
Tahun	2021
Hasil	Penelitian ini mencakup penerapan algoritma C4.5 untuk prediksi penyakit diabetes. Dengan meLaki-Lakifaatkan data kesehatan termasuk variabel-variabel seperti riwayat

	keluarga, tingkat aktivitas fisik, dan profil gula darah, algoritma C4.5 diharapkan dapat mengidentifikasi pola-pola yang signifikan yang mempengaruhi risiko seseorang terkena diabetes. Melalui analisis ini, penelitian bertujuan untuk mengembangkan model prediktif yang akurat dan dapat memberikan informasi dini terkait potensi risiko diabetes pada individu. Hasil penelitian ini diharapkan dapat memberikan landasan bagi upaya pencegahan lebih dini dan Laki-Lakiajemen penyakit yang lebih efektif, serta kontribusi terhadap pemahaLaki-Laki mendalam tentang faktor-faktor yang terlibat dalam prediksi penyakit diabetes menggunakan pendekatan algoritma C4.5 [14].
Referensi Penelitian	2

Judul	PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4.5, NAÏVE BAYES, KNN, DAN SVM
Nama	Sri Widaningsih
Tahun	2019
Hasil	Penelitian ini akan fokus pada perbandingan metode data mining untuk prediksi nilai dan waktu kelulusan mahasiswa Program Studi Teknik Informatika dengan menggunakan empat algoritma utama, yaitu C4.5, Naïve Bayes, K-Nearest Neighbors (KNN), dan Support Vector Machine (SVM). Dengan menerapkan algoritma C4.5, penelitian bertujuan untuk mengidentifikasi pola dan hubungan antara berbagai variabel, termasuk hasil akademis dan faktor-

	faktor terkait lainnya, guna memprediksi nilai dan waktu kelulusan mahasiswa. Perbandingan dengan metode lainnya diharapkan dapat memberikan pemahaLaki-Laki yang komprehensif tentang keefektifan dan keunggulan algoritma C4.5 dalam konteks prediksi kelulusan mahasiswa di Program Studi Teknik Informatika. Hasil penelitian ini dapat menjadi panduan berharga bagi institusi pendidikan untuk meningkatkan efisiensi dan efektivitas dalam memahami serta mendukung perkembangan akademis mahasiswa [15].

2.5.2. Kerangka Kerja Penelitian



1. Pemilihan Data

Pemilihan data merupakan tahap kritis dalam proses data mining yang melibatkan pengidentifikasian dan pengumpulan subset data yang relevan untuk analisis. Langkah ini memerlukan pemahaLaki-Laki mendalam terhadap tujuan analisis serta karakteristik data yang ada. Dalam pemilihan data, penting untuk memastikan kualitas data dengan mengidentifikasi dan menangani missing values, outliers, dan noise yang dapat memengaruhi hasil analisis. Selain itu, pemilihan atribut atau fitur yang paling relevan juga merupakan aspek krusial untuk mendapatkan informasi yang berguna. Proses pemilihan data yang cermat dan terarah dapat mengoptimalkan kinerja algoritma data mining dan memastikan hasil analisis yang lebih akurat dan bermakna.

2. *Preprocessing Data*

Preprocessing data adalah fase kritis dalam proses data mining yang melibatkan serangkaian langkah untuk membersihkan, mengorganisir, dan mempersiapkan data sebelum dilakukan analisis lebih lanjut. Tujuan utama dari preprocessing data adalah meningkatkan kualitas data, mengatasi ketidaksempurnaan, dan memastikan bahwa data siap digunakan oleh algoritma data mining. Langkah-langkah umum preprocessing melibatkan penanganan missing values, deteksi dan penanganan outlier, normalisasi atau standarisasi skala, dan konversi atribut kategorikal menjadi bentuk yang dapat diolah oleh algoritma, seperti one-hot encoding. Proses ini juga mencakup pemilihan atribut yang relevan, pengurangan dimensi, dan penghapusan duplikat untuk memastikan bahwa model yang dibangun dari data tersebut dapat memberikan hasil yang lebih akurat dan bermakna. Preprocessing data yang teliti dan efektif dapat secara signifikan meningkatkan kinerja algoritma data mining dan hasil analisis yang dihasilkan.

3. *Pembagian Data*

Pembagian data, atau splitting data, merupakan langkah penting dalam proses data mining untuk mengevaluasi dan menguji kinerja model. Data biasanya dibagi menjadi dua subset utama: data pelatihan (training data) dan data pengujian (testing data). Data pelatihan digunakan untuk melatih model dan mengidentifikasi pola serta hubungan dalam dataset. Sementara itu, data pengujian digunakan untuk menguji sejauh Laki-Lakia model dapat menggeneralisasi pola yang telah dipelajari dari data pelatihan ke data yang belum pernah dilihat sebelumnya. Pembagian data yang baik membantu menghindari

overfitting atau underfitting, memastikan bahwa model memiliki kemampuan yang baik untuk memberikan prediksi yang akurat pada data baru yang tidak digunakan selama proses pelatihan. Selain itu, teknik pembagian data seperti validasi silang (cross-validation) juga dapat diterapkan untuk memaksimalkan penggunaan data dan menghasilkan evaluasi model yang lebih konsisten.

4. Pembentukan Model Algoritma C4.5

Pembentukan model algoritma C4.5 adalah tahap penting dalam proses data mining yang bertujuan untuk membangun pohon keputusan yang dapat digunakan untuk klasifikasi. Algoritma C4.5, dikembangkan oleh Ross Quinlan, menggunakan pendekatan rekursif untuk membagi dataset berdasarkan atribut-atribut yang paling informatif dalam mengklasifikasikan data. Proses dimulai dengan pemilihan atribut terbaik untuk digunakan sebagai node keputusan pada pohon, yang dilakukan dengan menggunakan metrik informasi seperti gain informasi atau rasio gain informasi. Setelah pemilihan atribut, data dibagi menjadi subset yang lebih kecil berdasarkan nilai-nilai atribut tersebut. Proses ini diulang secara rekursif untuk setiap subset hingga terbentuk pohon keputusan lengkap. Selanjutnya, pohon tersebut dapat digunakan untuk mengklasifikasikan data baru dengan mengikuti jalur keputusan dari akar pohon ke daun yang sesuai dengan fitur-fitur data tersebut. Algoritma C4.5 dikenal karena kemampuannya mengatasi data kategorikal dan numerik, serta kemampuannya menghasilkan pohon keputusan yang mudah diinterpretasi.

5. Klasifikasi dan Prediksi Model

Klasifikasi dan prediksi model merupakan tahapan esensial dalam data mining yang bertujuan untuk mengelompokkan atau memprediksi nilai dari suatu

instance berdasarkan pola yang telah diidentifikasi dari data pelatihan. Dalam klasifikasi, model mengkategorikan instance ke dalam kelas atau kelompok tertentu berdasarkan atribut-atribut yang telah didefinisikan. Algoritma klasifikasi, seperti Naive Bayes, Decision Trees, atau Support Vector Machines, digunakan untuk membangun model yang dapat mengenali pola-pola ini dan mengaplikasikannya pada data baru. Sebaliknya, pada prediksi, model memperkirakan nilai dari suatu variabel target berdasarkan atribut-atribut yang ada. Regresi linear, regresi logistik, dan neural networks adalah contoh algoritma yang sering digunakan untuk pembuatan model prediksi. Kesuksesan model dalam klasifikasi dan prediksi sangat tergantung pada kualitas data, pemilihan atribut yang tepat, dan parameter yang disesuaikan dengan baik selama tahap pembentukan model. Model yang baik dapat memberikan hasil prediksi yang akurat dan berguna untuk pengambilan keputusan dalam berbagai domain aplikasi.

6. Evaluasi Model

Evaluasi model adalah langkah kritis dalam proses data mining yang bertujuan untuk menilai kinerja dan kehandalan model yang telah dibangun. Metode evaluasi yang umum digunakan melibatkan penggunaan metrik klasifikasi, seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran tentang kemampuan model dalam mengklasifikasikan data. Selain itu, kurva ROC dan area di bawah kurva ROC (AUC-ROC) sering digunakan untuk mengevaluasi performa model klasifikasi biner. Bagi model prediksi, metrik seperti mean squared error (MSE) atau coefficient of determination (R-squared)

sering digunakan untuk menilai seberapa baik model memprediksi nilai target. Validasi silang (cross-validation) juga dapat digunakan untuk memastikan bahwa model memiliki generalisasi yang baik dan tidak hanya dioptimalkan untuk data pelatihan tertentu. Evaluasi model yang cermat dan holistik membantu memastikan bahwa model dapat diandalkan dalam memecahkan masalah dunia nyata dan memberikan kontribusi yang berarti untuk pengambilan keputusan.

7. Visualisasi Algoritma C4.5

Visualisasi algoritma C4.5 dalam konteks data mining melibatkan representasi grafis dari pohon keputusan yang dihasilkan. Pohon keputusan tersebut memperlihatkan bagaiLaki-Lakia algoritma C4.5 membuat keputusan klasifikasi berdasarkan atribut-atribut dalam dataset. Setiap simpul pada pohon mewakili keputusan berdasarkan suatu atribut, dan cabang-cabangnya mencerminkan nilai-nilai atribut yang membagi data. Visualisasi ini memberikan pandangan yang jelas dan intuitif tentang bagaiLaki-Lakia algoritma mengklasifikasikan instance data ke dalam kelas yang berbeda. Warna atau ketebalan garis pada cabang-cabang pohon dapat digunakan untuk menunjukkan tingkat kepentingan atau informasi dari setiap atribut. Dengan cara ini, visualisasi membantu peneliti, analis, dan peLaki-Lakigku kepentingan dalam memahami logika dan keputusan yang diambil oleh model C4.5, serta memberikan wawasan yang berguna untuk interpretasi dan peningkatan model secara keseluruhan.