BAB II

LANDASAN TEORI

2.1. Data Mining

Data Mining, atau penambangan data, adalah proses ekstraksi pola atau informasi yang berharga dari dataset yang besar dan kompleks [1]. Tujuannya adalah untuk mengungkap wawasan yang mungkin tidak terlihat secara langsung melalui metode analisis konvensional [2]. Data Mining melibatkan penggunaan berbagai teknik, termasuk statistika, machine learning, dan kecerdasan buatan, untuk mengidentifikasi pola, tren, dan hubungan dalam data.

Salah satu aspek utama dari Data Mining adalah identifikasi pola yang dapat digunakan untuk membuat prediksi atau mengambil keputusan. Ini melibatkan penggunaan algoritma untuk mengelompokkan data (clustering), mengklasifikasikan data ke dalam kategori tertentu (classification), mengidentifikasi hubungan antara variabel (association), dan membuat prediksi berdasarkan data historis (regression).

Data Mining sering digunakan dalam berbagai bidang dan industri, termasuk bisnis, kesehatan, keuangan, dan ilmu pengetahuan. Contoh penggunaannya termasuk analisis perilaku konsumen untuk meningkatkan strategi pemasaran, prediksi penyebaran penyakit berdasarkan pola geografis, dan deteksi kecurangan keuangan dengan mengidentifikasi pola anomali.

Proses Data Mining melibatkan langkah-langkah seperti pengumpulan data, pembersihan data, pemilihan atribut, pemilihan model, dan evaluasi hasil. Penggunaan teknologi seperti big data dan sistem manajemen basis data khusus

untuk Data Mining membantu memproses dan menganalisis volume data yang besar. Meskipun Data Mining memberikan wawasan yang berharga, ada juga tantangan dan pertimbangan etika terkait privasi dan keamanan data. Dengan kemajuan teknologi, Data Mining terus berkembang dan menjadi instrumen penting dalam pengambilan keputusan dan inovasi di berbagai sektor.

2.1.1. Database dan Data Processing

Database dan Data Processing merupakan dua aspek kunci dalam pengelolaan dan penggunaan data di dunia teknologi informasi. Database adalah suatu sistem penyimpanan dan pengelolaan data yang terstruktur, memungkinkan pengguna untuk menyimpan, mengambil, dan memanipulasi informasi dengan cara yang terorganisir. Dengan menggunakan database, data dapat diatur dalam tabel dan relasi, memudahkan akses, pengelolaan, dan analisis data.

Sistem manajemen basis data (DBMS) adalah perangkat lunak yang memungkinkan pengguna untuk berinteraksi dengan database. DBMS menyediakan antarmuka untuk memasukkan, mengubah, dan mengambil data dari database, serta melindungi integritas data dan memberikan akses terkontrol kepada pengguna. Contoh DBMS yang populer termasuk MySQL, PostgreSQL, dan Microsoft SQL Server.

Data Processing, di sisi lain, merujuk pada serangkaian langkah atau proses untuk mentransformasikan data mentah menjadi bentuk yang lebih bermanfaat atau dapat dimengerti. Ini melibatkan pengumpulan, pembersihan, transformasi, analisis, dan penyajian data. Data Processing dapat dilakukan secara batch atau real-time, tergantung pada kebutuhan dan sifat data yang dihadapi.

Teknologi modern telah memperkenalkan konsep pemrosesan data skala besar, yang memungkinkan organisasi untuk mengelola, menganalisis, dan mengekstrak wawasan dari volume data yang sangat besar dengan cepat dan efisien. Pemrosesan data berbasis cloud juga semakin umum, memungkinkan akses dan pengelolaan data melalui infrastruktur cloud.

Kombinasi antara database yang efisien dan proses data yang cermat membentuk dasar bagi sistem informasi yang kuat. Dalam era big data, di mana jumlah data yang dihasilkan terus meningkat, kemampuan untuk menyimpan, mengelola, dan memproses data menjadi semakin penting untuk mendukung keputusan bisnis, inovasi, dan pengembangan teknologi.

2.1.2. Visualisation

Visualisasi adalah suatu proses representasi data dan informasi secara visual, yang bertujuan untuk membuat pola, tren, dan relasi dalam data menjadi lebih jelas dan mudah dimengerti. Dengan menggunakan elemen visual seperti grafik, diagram, peta, dan infografis, visualisasi memungkinkan pengguna untuk meresapi informasi kompleks secara cepat dan efektif. Pendekatan ini membantu mendekatkan kesenjangan antara manusia dan mesin, memungkinkan pemahaman intuitif dan pengambilan keputusan yang lebih baik.

Visualisasi memiliki peran kunci dalam berbagai disiplin, termasuk ilmu data, bisnis, ilmu sosial, dan ilmu komputer. Dalam ilmu data, visualisasi membantu data scientist dan analis dalam menganalisis dan menggali wawasan dari dataset yang besar dan kompleks. Dalam bisnis, visualisasi digunakan untuk menyajikan hasil analisis pasar, kinerja perusahaan, dan tren penjualan secara

lebih mudah dipahami oleh pemangku kepentingan. Jenis visualisasi bervariasi tergantung pada tujuan dan jenis data yang dihadapi. Contoh visualisasi meliputi grafik garis untuk melacak perubahan sepanjang waktu, diagram batang untuk membandingkan kuantitas, peta panas untuk menunjukkan distribusi intensitas data, dan banyak lagi. Penggunaan warna, ukuran, dan bentuk membantu menyampaikan informasi dengan cara yang lebih menarik dan mudah diingat.

Selain dari kegunaannya sebagai alat analisis dan komunikasi, visualisasi juga mendorong kreativitas dan inovasi. Penerapan teknologi seperti augmented reality (AR) dan virtual reality (VR) semakin memperkaya pengalaman visualisasi, memberikan dimensi tambahan dalam memahami data. Dalam era big data, di mana jumlah data terus berkembang, visualisasi menjadi semakin penting untuk menggambarkan informasi secara efektif dan mengidentifikasi pola yang relevan. Dengan memadukan kekuatan analisis dan daya tarik estetika, visualisasi membantu membawa data kehidupan dan menjadi alat yang tak ternilai dalam mendukung pemahaman, pengambilan keputusan, dan eksplorasi data.

2.1.3. Statistik

Statistik adalah cabang ilmu matematika yang berkaitan dengan pengumpulan, analisis, interpretasi, presentasi, dan pengorganisasian data. Tujuannya adalah untuk menyajikan 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 utama statistik: statistik deskriptif dan statistik inferensial. Statistik deskriptif berkaitan dengan penggambaran dan ringkasan data, termasuk penggunaan ukuran tendensi sentral seperti mean (rata-rata), median (nilai tengah), dan modus (nilai yang sering muncul), serta ukuran variasi seperti kisaran dan simpangan baku. Sementara itu, statistik inferensial digunakan untuk membuat inferensi atau perkiraan tentang suatu populasi berdasarkan data yang diambil dari sampel. Ini melibatkan konsep probabilitas dan teknik seperti uji hipotesis dan interval kepercayaan.

Statistik memiliki aplikasi luas di berbagai bidang, termasuk ilmu sosial, ekonomi, kedokteran, sains alam, dan bisnis. Dalam bisnis, statistik digunakan untuk analisis pasar, prediksi tren penjualan, dan evaluasi kinerja bisnis. Di bidang ilmiah, statistik membantu menguji hipotesis, mengukur signifikansi hasil eksperimen, dan menyimpulkan generalisasi tentang fenomena alam.

Dengan kemajuan teknologi, statistik terus berkembang, dan menjadi dasar penting untuk analisis data yang lebih canggih, seperti machine learning dan analisis big data. Kemampuan untuk memahami dan menerapkan prinsip statistik menjadi keterampilan kritis dalam membantu kita membuat keputusan yang informasional dan cerdas dalam berbagai aspek kehidupan.

2.1.4. Pattern Recognition

Pengenalan Pola, atau Pattern Recognition, adalah disiplin ilmu yang berfokus pada pengembangan metode dan algoritma untuk mengidentifikasi dan mengekstraksi pola yang tersembunyi atau bermakna dalam data. Tujuan utama dari pengenalan pola adalah memungkinkan komputer atau sistem cerdas untuk

memahami dan merespons terhadap data dengan cara yang mirip dengan cara manusia mengenali pola. Ini melibatkan penggunaan teknik pembelajaran mesin, statistika, dan kecerdasan buatan untuk mengembangkan model yang dapat mengenali pola dalam data, baik itu gambar, suara, teks, atau data lainnya. Dalam pengenalan pola, komputer diajarkan untuk mengenali dan membedakan pola berdasarkan fitur-fitur tertentu yang diidentifikasi dalam data. Ini bisa mencakup pengenalan wajah dalam gambar, identifikasi sidik jari, atau pengklasifikasian teks berdasarkan isi dan konteksnya. Algoritma pembelajaran mesin, seperti neural networks, decision trees, dan support vector machines, sering digunakan dalam pengenalan pola untuk melatih model dengan data yang diberi label.

Aplikasi pengenalan pola sangat luas dan mencakup berbagai bidang. Dalam pengolahan citra, pengenalan pola digunakan untuk pengenalan objek atau orang dalam gambar. Di bidang medis, teknik pengenalan pola membantu dalam mendiagnosis penyakit atau mendeteksi anomali dalam gambar medis. Penggunaan pengenalan pola juga terlihat dalam sistem keamanan, otomasi industri, dan pengenalan suara. Pengenalan Pola juga dihadapkan pada beberapa tantangan, seperti variabilitas data, kompleksitas, dan kemampuan adaptasi terhadap perubahan pola. Oleh karena itu, penelitian terus dilakukan untuk meningkatkan keakuratan dan ketangguhan algoritma pengenalan pola dalam berbagai konteks dan aplikasi. Seiring dengan kemajuan teknologi, pengenalan pola terus berkembang dan menjadi bagian penting dalam pengembangan sistem cerdas dan teknologi yang dapat beradaptasi dengan lingkungan yang semakin kompleks.

2.2. Model Klasifikasi

Model klasifikasi adalah salah satu jenis model dalam pembelajaran mesin yang digunakan untuk memprediksi kategori atau label dari suatu data [3] [4]. Tujuan utama dari model klasifikasi adalah memahami pola dan hubungan dalam data pelatihan yang terlabel dan kemudian mengeneralisasikannya untuk memprediksi kelas atau label dari data yang belum terlihat sebelumnya [5]. Model klasifikasi memerlukan proses pelatihan di mana algoritma belajar dari contoh-contoh data yang sudah diketahui kelasnya.

Proses pelatihan model klasifikasi melibatkan penggunaan berbagai algoritma pembelajaran mesin, seperti decision trees, logistic regression, support vector machines, dan neural networks. Setiap algoritma memiliki kelebihan dan kekurangan sendiri, dan pilihan algoritma dapat bergantung pada karakteristik data, ukuran dataset, dan kebutuhan spesifik dari tugas klasifikasi yang dihadapi [6].

Hasil dari model klasifikasi adalah kemampuannya untuk memprediksi kelas atau label yang benar untuk data baru. Evaluasi kinerja model dapat dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score . Akurasi mengukur sejauh mana model memprediksi dengan benar, sedangkan presisi dan recall memberikan wawasan tentang kemampuan model dalam mengidentifikasi kelas tertentu dan menghindari kesalahan prediksi palsu positif atau palsu negatif.

Model klasifikasi memiliki berbagai aplikasi luas di berbagai industri. Contohnya termasuk klasifikasi email sebagai spam atau bukan spam, identifikasi

kategori produk dalam perdagangan elektronik, diagnosa medis, dan banyak lagi. Kemampuan model klasifikasi untuk memproses dan memahami pola dari data dengan cepat menjadikannya alat yang sangat berharga dalam pengambilan keputusan berbasis data. Namun, penting untuk memahami bahwa pemilihan dan pengembangan model klasifikasi memerlukan perhatian terhadap pemrosesan data yang tepat, pemilihan fitur yang cerdas, serta penanganan overfitting atau underfitting agar hasil prediksi dapat diandalkan dan memiliki generalisasi yang baik.

2.3. Algoritma ID3

Algoritma ID3 (Iterative Dichotomiser 3) adalah suatu algoritma pembelajaran mesin yang dikembangkan oleh Ross Quinlan, dan merupakan salah satu algoritma pohon keputusan yang paling awal dan terkenal [7]. ID3 dirancang untuk membangun model pohon keputusan yang efisien untuk tugas klasifikasi [8]. Pohon keputusan adalah struktur pohon yang memetakan fitur-fitur dari data ke output berupa kelas atau label.

Proses utama dalam algoritma ID3 adalah pemilihan atribut terbaik untuk membagi dataset pada setiap langkah dalam pembangunan pohon [9] [10]. Algoritma ini menggunakan konsep Information Gain sebagai metrik untuk mengukur seberapa baik suatu atribut memisahkan dataset menjadi kelas-kelas yang homogen. Information Gain mengukur seberapa banyak informasi yang diperoleh dari membagi dataset dengan atribut tertentu [11].

ID3 beroperasi secara rekursif, membagi dataset pada setiap simpul pohon keputusan berdasarkan atribut dengan Information Gain tertinggi. Proses ini terus

berlanjut hingga kondisi berhenti tercapai, seperti mencapai tingkat ketidakhomogenan tertentu atau mencapai batasan kedalaman tertentu dalam pohon. Kelebihan dari ID3 termasuk kemudahan interpretasi, karena pohon keputusan yang dihasilkan mudah dimengerti. Selain itu, ID3 dapat menangani atribut dengan tipe data kategorikal. Namun, ID3 juga memiliki kelemahan, seperti kecenderungan untuk overfitting, terutama ketika membangun pohon dengan kedalaman yang besar pada dataset yang relatif kecil.

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

Keterangan:

S = Himpunan kasus

n = Jumlah Partisi S

P_1 = Proporsi S_i terhadap S

$$Log2\ (x) = \frac{\ln(x)}{\ln(2)}$$

$$Log2\ Pi = \frac{\ln(Pi)}{\ln(2)}$$

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

Keterangan:

S = Himpunan kasus

A = Fitur

n = Jumlah Partisi atribut A

|Si| = Proporsi S_i terhadap S

|S| = Jumlah kasus dalam S

Meskipun ID3 telah memberikan dasar bagi banyak pengembangan pohon keputusan lainnya, seperti C4.5, CART, dan Random Forest, kebanyakan algoritma modern menggunakan variasi atau penyempurnaan dari konsep yang diperkenalkan oleh ID3. Meskipun ID3 bukan lagi algoritma utama yang digunakan secara luas, kontribusinya dalam mengenalkan konsep pohon keputusan tetap menjadi bagian integral dalam sejarah pembelajaran mesin.

2.4. Uji Performa

Uji performa algoritma ID3 dengan menggunakan Confusion Matrix adalah metode yang efektif untuk mengevaluasi kinerja model dalam klasifikasi data. Confusion Matrix memungkinkan kita untuk menghitung akurasi, presisi, recall, dan nilai F-1 dari model yang dihasilkan. Akurasi mengukur seberapa tepat model dalam mengklasifikasikan keseluruhan instance, presisi mengukur seberapa banyak instance yang diprediksi sebagai positif adalah benar-benar positif, recall mengukur seberapa banyak instance positif yang berhasil diprediksi oleh model, dan nilai F-1 adalah rata-rata harmonis antara presisi dan recall, memberikan gambaran holistik tentang kinerja model.

		Kelas Prediksi	
	Kelas	**Belar**	**Salah**
Kelas Atribut	Benar	True Positive (TP)	False Positive (FP)
	Salah	False Negative (FN)	True Negative (TN)

$$\boldsymbol{Acuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FN + FP}} \times \mathbf{100\%}$$

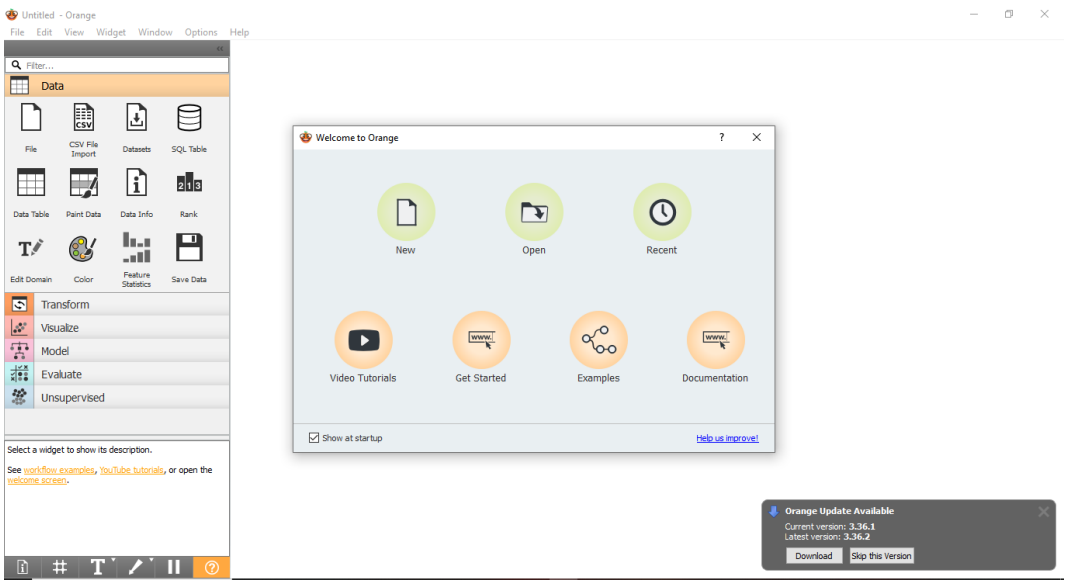
$$Presisi = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

2.5. Alat Bantu Program/Tools Pendukung

2.5.1. Orange

Orange adalah suatu platform perangkat lunak open-source yang memungkinkan analisis data visual dan pembelajaran mesin tanpa memerlukan pengetahuan pemrograman yang mendalam. Dirancang untuk membuka akses kepada berbagai kalangan, termasuk peneliti, ilmuwan data, dan praktisi yang mungkin tidak memiliki latar belakang teknis yang kuat, Orange menawarkan antarmuka grafis yang intuitif untuk menyusun dan memahami alur kerja analisis data.



Gambar 2.5. 1. Tampilan Awal Aplikasi Orange

Salah satu fitur utama dari Orange adalah beragam widget atau komponen visual yang dapat digunakan pengguna untuk menyusun dan mengonfigurasi analisis data mereka. Widget tersebut mencakup berbagai fungsi, seperti pemrosesan data, ekstraksi fitur, visualisasi, dan pembelajaran mesin, yang dapat dihubungkan dan disusun sesuai dengan kebutuhan analisis.

Orange menyediakan berbagai algoritma pembelajaran mesin yang dapat diakses melalui antarmuka visualnya, memungkinkan pengguna untuk memilih, mengonfigurasi, dan menganalisis model tanpa perlu menulis kode. Hal ini membuat Orange menjadi alat yang sangat berguna bagi mereka yang ingin menjelajahi analisis data dan konsep pembelajaran mesin tanpa harus menghadapi kompleksitas kode. Dengan dukungan dari komunitas pengguna yang aktif dan pengembang yang berdedikasi, Orange terus berkembang dan mengalami pembaruan reguler. Platform ini juga mendukung berbagai format data dan integrasi dengan bahasa pemrograman seperti Python, memberikan fleksibilitas tambahan untuk para penggunanya.

Dengan memberikan akses yang lebih mudah dan intuitif ke analisis data dan pembelajaran mesin, Orange membantu mendemokratisasi dunia kecerdasan buatan dan membuat teknologi tersebut dapat diakses oleh berbagai kalangan. Platform ini memiliki peran penting dalam mendukung kegiatan penelitian, analisis data, dan pengembangan model di berbagai industri dan disiplin ilmu.

2.6. Metodologi Penelitian

2.6.1. Penelitian Terdahulu

Referensi Penelitian	1

Judul	Penerapan Data Mining dalam Menganalisa Pola Kelayakan Siswa Pada Kelas Unggulan Menggunakan Algoritma Iterative Dichotomiser 3 (ID3) pada SMP N. 2 Rantau Selatan
Nama	Masyuni Hutasuhut , Dina Octavina, Jufri Halim
Tahun	2019
Hasil	Penelitian ini bertujuan untuk menganalisis pola kelayakan siswa pada kelas unggulan di SMP N. 2 Rantau Selatan menggunakan Algoritma Iterative Dichotomiser 3 (ID3). Dengan merinci variabel-variabel seperti prestasi akademis, partisipasi dalam kegiatan ekstrakurikuler, dan faktor-faktor lain yang mungkin memengaruhi keberhasilan siswa, penelitian ini mengumpulkan data yang diperlukan untuk mengaplikasikan algoritma ID3. Melalui analisis ini, diharapkan

	penelitian dapat memberikan pemahaman mendalam tentang pola kelayakan siswa dalam program kelas unggulan, memfasilitasi identifikasi faktor-faktor kunci yang berkontribusi pada keberhasilan siswa dalam lingkungan pendidikan khusus ini di SMP N. 2 Rantau Selatan [12].
Referensi Penelitian	2
Judul	Determinan Pemberian ASI Eksklusif oleh Ibu Menyusui yang Bekerja dengan Algoritma ID3
Nama	Fadhiyah Noor Anisa1□, Laurensia Yunita2, Ahmad Hidayat3
Tahun	2022
Hasil	Dalam penelitian ini, metode Iterative Dichotomiser 3 (ID3) diterapkan untuk menganalisis determinan pemberian ASI eksklusif oleh ibu menyusui. Penelitian ini memfokuskan pada variabel-variabel seperti

	pengetahuan ibu tentang manfaat ASI eksklusif, dukungan sosial yang diterima, dan faktor-faktor lain yang mungkin memengaruhi keputusan ibu menyusui dalam memberikan ASI eksklusif. Dengan menggunakan ID3, penelitian ini bertujuan untuk mengidentifikasi pola dan hierarki faktor-faktor yang paling signifikan dalam pengambilan keputusan ibu menyusui terkait pemberian ASI eksklusif. Diharapkan hasil penelitian ini dapat memberikan wawasan yang lebih mendalam tentang faktor-faktor yang memengaruhi praktik pemberian ASI eksklusif oleh ibu menyusui, sehingga dapat memberikan dasar untuk pengembangan program intervensi yang lebih efektif dalam meningkatkan praktik pemberian ASI eksklusif [13].
Referensi Penelitian	3

Judul	Analisis Algoritma ID3 Pada Kunjungan Akseptor KB di Kota Banjarmasin
Nama	Laurensia Yunita□, Fadhiyah Noor Anisa2, Rina Saputri3
Tahun	2023
Hasil	Dalam penelitian ini, metode Iterative Dichotomiser 3 (ID3) diterapkan untuk menganalisis algoritma ID3 pada kunjungan akseptor KB di Kota Banjarmasin. Penelitian ini memusatkan perhatian pada variabel-variabel seperti umur, tingkat pendidikan, informasi mengenai metode kontrasepsi, dan faktor-faktor lain yang mungkin mempengaruhi keputusan akseptor KB dalam kunjungan mereka. Dengan menerapkan ID3, penelitian ini bertujuan untuk mengidentifikasi pola keterkaitan dan hierarki faktor-faktor yang paling signifikan dalam

	keputusan akseptor KB terkait dengan kunjungan mereka. Hasil penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi penerimaan metode kontrasepsi di Kota Banjarmasin, memberikan landasan untuk perbaikan layanan kesehatan reproduksi, dan mendukung upaya dalam meningkatkan kesadaran dan partisipasi masyarakat terhadap program keluarga berencana [14].
Referensi Penelitian	4
Judul	Penerapan Algoritma ID3 dalam Prediksi Kebutuhan Pupuk
Nama	Milyani Aritonang
Tahun	2021
Hasil	Dalam penelitian ini, algoritma Iterative Dichotomiser 3 (ID3) diterapkan untuk prediksi kebutuhan

	pupuk. Penelitian ini memusatkan perhatian pada variabel-variabel seperti jenis tanaman, jenis tanah, dan tingkat kelembaban yang dapat mempengaruhi kebutuhan pupuk pada suatu area pertanian. Dengan menerapkan algoritma ID3, penelitian bertujuan untuk mengidentifikasi pola keterkaitan antara variabel-variabel tersebut dan mengembangkan model prediktif yang dapat memberikan estimasi yang akurat terkait kebutuhan pupuk di berbagai kondisi pertanian. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan efisiensi penggunaan pupuk, mengoptimalkan produksi pertanian, dan mendukung praktik pertanian berkelanjutan dengan meminimalkan dampak lingkungan [15].