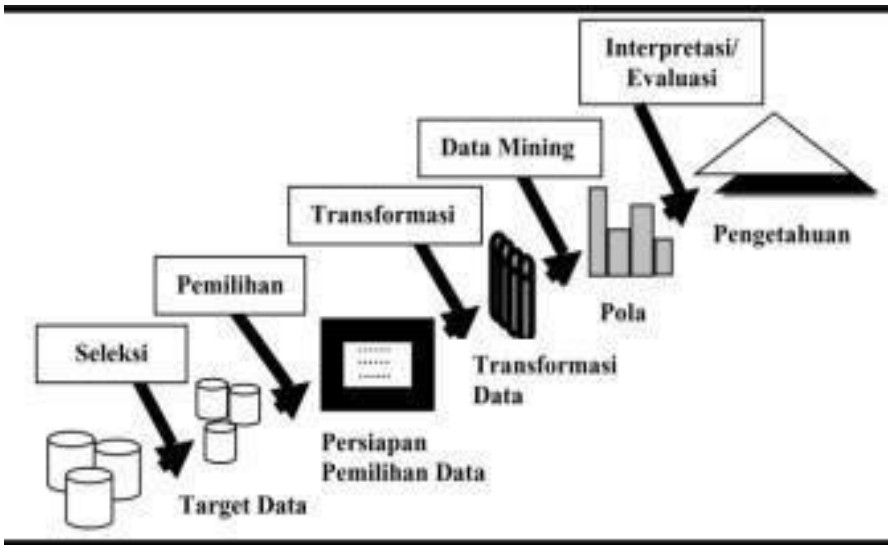
BAB II

LANDASAN TEORI

2.1 Knowledge Discovery In Database

Knowledge Discovery in Database (KDD) adalah proses pengambilan informasi yang tersembunyi, di mana informasi tersebut sebelumnya tidak dikenal. *KDD* merupakan metode yang digunakan untuk mendapat pengetahuan yang berasal dari database yang ada. Hasil pengetahuan yang diterima dapat dimanfaatkan untuk basis pengetahuan *(knowledge base)* yang dipergunakan dalam keperluan mengambil keputusan [4].



Gambar 2. 1 Tahapan Dalam KDD

(Sumber : https://dhindaawanys.blogspot.com)

Salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Seperti yang ditujukan pada proses berikut:

1. Selection (Seleksi Data)

Pada bagian ini Selection/seleksi ataupun pemilihan data-data dari beberapa sekumpulan data – data dari operasional yang perlu dilakukan sebelum tahap – tahapan penggalian sebuah informasi yang ada di dalam Knowledge discovery

database (KDD) dimulai dari awal. Adapun Data dari hasil sebuah seleksi yang akan dipakai atau digunakan untuk sebuah proses dari data mining tersebut biasanya disimpan di dalam sesuatu berkas yang besar dan biasanya dapat terpisah dari beberapa basis data operasional yang dilakukan.

2. Preprocessing/Cleaning (Pemilihan Data)

Pada bagian ini Proses Preprocessing hanya mencakup di antara lain adalah dengan cara membuang duplikasi data yang ada, lalu memeriksa data yang lebih inkonsisten yang ada, dan tentunya dengan memperbaikikesalahan yang ada pada data tersbut, contohnya seperti kesalahan pada cetak atau di sebut tipografi.

3. Transformation (Transformasi)

Pada bagian fase saat ini yang dapat dilakukan adalah dengan mentransformasi dari bentuk – bentuk data yang tentunya belum memiliki sebuah entitas yang sudah sangat jelas ke dalam bentuk – bentuk data yang sangat valid atau bisa di bilang siap untuk pengerjaan yang dilakukan pada proses Data mining yang sedang berjalan.

4. Data Mining

Pada bagian fase saat ini hal yang utama yang dapat dilakukan adalah dengan cara menerapkan sebuah algoritma ataupun metode pencarian pengetahuan yang akan dilakukan pada proses pencarian data.

5. Interpratation/Evaluation (Interpretasi/Evaluasi)

Pada bagian fase yang terakhir ini yang dapatdilakukan adalah dengan cara proses pembentukan dari keluaran yang sangat mudah dimengerti dan biasanya yang bersumber daripada proses–proses data mining pada pola informasi yang dicari.

2.2 Data Mining

Data mining merupakan proses menemukan informasi dari suatu data yang tersimpan dalam suatu database atau datasheet. Pembuatan model dilakukan dengan proses menggunakan algoritma atau rumus tertentu. Proses data mining menggunakan berbagai teknik seperti teknik dalam proses statistik, matematika, dan machine learning yang digunakan dalam melakukan identifikasi dan mengolah berbagai data menjadi informasi yang bermanfaat [5].

Data mining adalah kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran/berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan. Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar [6].

2.3 Algoritma C4.5

2.3.1 Penjelasan Algoritma C4.5 dan Rumusnya

Algoritma C4.5 adalah metode dalam data mining yang digunakan untuk membangun pohon keputusan berdasarkan data. Pohon keputusan ini membantu dalam melakukan klasifikasi atau prediksi dengan mengidentifikasi atribut yang paling signifikan. Algoritma ini merupakan pengembangan dari algoritma ID3 dengan perbaikan, seperti kemampuan menangani data yang tidak lengkap, atribut numerik, dan pembobotan atribut [7].

2.3.2 Tahapan Kerja Algoritma C4.5

1. Pemilihan Atribut Root (Akar): Algoritma memilih atribut dengan informasi paling tinggi menggunakan *Gain Ratio* sebagai metrik seleksi.

2. Pembuatan Node: Data dibagi berdasarkan nilai atribut yang dipilih.
3. Rekursi: Proses ini berlanjut pada setiap subset hingga memenuhi kriteria penghentian, seperti semua data dalam subset termasuk dalam satu kelas.
4. Pruning: Menghapus cabang pohon yang tidak relevan untuk mencegah *overfitting*.

2.3.3 Rumus Utama:

A. Entropy (E)

Entropy adalah ukuran dari ketidakpastian dalam dataset [8]. Rumusnya sebagai berikut:

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

Keterangan:

1. $\boldsymbol{S}$: Dataset yang sedang dievaluasi.
2. $\boldsymbol{n}$: Jumlah kelas yang ada dalam dataset.
3. $\boldsymbol{pi}$: Proporsi data pada kelas $\boldsymbol{i}$, dihitung sebagai:

   $\boldsymbol{pi}$: $\frac{\textbf{Jumlah data pada kelas } \boldsymbol{i}}{\textbf{Total Data di } \boldsymbol{S}}$

B. Information Gain (IG)

Information Gain mengukur pengurangan ketidakpastian (entropy) setelah data dibagi berdasarkan atribut tertentu [9]. Rumusnya:

$$IG\ (S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times E(S_v)$$

Keterangan:

1. $\boldsymbol{A}$: Atribut yang sedang diuji.
2. $\boldsymbol{Values(A)}$: Semua nilai unik yang dimiliki oleh atribut A
3. $\boldsymbol{S_v}$: Subset dari $\boldsymbol{S}$ dimana atribut $\boldsymbol{A}$ memiliki nilai $\boldsymbol{v}$.
4. $\frac{|S_v|}{|S|}$: Proporsi data pada subset $\boldsymbol{S_v}$.

C. Split Information (SplitInfo)

Split Information mengukur distribusi data pada subset yang dihasilkan oleh pembagian atribut [10]. Rumusnya:

$$SplitInfo(A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|}$$

Keterangan:

1. Konsep ini digunakan untuk menormalkan *Information Gain* agar tidak bias terhadap atribut dengan banyak nilai unik.
2. Mirip dengan Entropy, namun menghitung distribusi data berdasarkan pembagian atribut ***A***\

D. Gain Ratio

Gain Ratio adalah metrik utama untuk memilih atribut dalam algoritma C4.5 [11]. Rumusnya:

$$GainRatio(A) = \frac{IG(S, A)}{SplitInfo(A)}$$

Keterangan:

a. Nilai ini memperhitungkan baik pengurangan ketidakpastian (IG) maupun distribusi data setelah pembagian (*SplitInfo*).
b. Atribut dengan *Gain Ratio* tertinggi dipilih sebagai *root* atau *node* berikutnya.

2.4 Regresi Linear

Regresi linear adalah salah satu metode analisis statistik yang digunakan untuk memahami hubungan antara satu variabel dependen (variabel yang ingin diprediksi atau dijelaskan) dengan satu atau lebih variabel independen (variabel yang digunakan untuk melakukan prediksi atau penjelasan). Metode ini mengasumsikan bahwa hubungan antara variabel-variabel tersebut bersifat linear, artinya perubahan pada variabel independen akan menghasilkan perubahan yang proporsional pada variabel dependen. Regresi linier merupakan alat statistik yang

sangat ampuh dan fleksibel untuk mempelajari hubungan antara variabel-variabel dalam suatu fenomena [12].

Regresi linear dikelompokkan menjadi 2, yaitu:

1. Regresi Linear Sederhana

   Regresi linear sederhana digunakan ketika hanya terdapat satu variabel independen. Model ini dirumuskan dalam bentuk persamaan:

$$Y = a + bX + \in$$

   Keterangan:

   a. $\boldsymbol{Y}$: Variabel independen.
   b. $\boldsymbol{X}$: Variabel independen.
   c. $\boldsymbol{a}$: Konstanta atau intersep, yaitu nilai $\boldsymbol{Y}$ ketika nilai $\boldsymbol{X = 0}$.
   d. $\boldsymbol{b}$: Koefisien regresi, yang menunjukkan tingkat perubahan pada $\boldsymbol{Y}$ untuk setiap perubahan satu unit pada $\boldsymbol{X}$.
   e. $\boldsymbol{e}$: Error atau gangguan yang mengakomodasi variabel lain yang tidak dimasukkan dalam model.

2. Regresi Linear Berganda

   Regresi linear berganda digunakan ketika terdapat lebih dari satu variabel independen. Model ini dirumuskan sebagai:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n + \in$$

   Keterangan:

   a. $\boldsymbol{Y}$: adalah variabel dependen.
   b. $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}$: adalah variabel independen.
   c. $\boldsymbol{B_1}, \boldsymbol{B_2}, \ldots, \boldsymbol{B_n}$: adalah koefisien regresi untuk masing-masing variabel independen.
   d. $\in$: adalah error atau gangguan.

Regresi linear dapat digunakan untuk menentukan hubungan antara variabel seperti curah hujan, jumlah pupuk yang digunakan, dan luas lahan dengan hasil panen. Model ini membantu perusahaan untuk memprediksi hasil panen di masa depan berdasarkan faktor-faktor tersebut.

2.5 Prediksi Kelapa Sawit

Prediksi produksi kelapa sawit adalah kegiatan memperkirakan ataupun meramalkan hasil produksi yang akan terjadi pada masa mendatang melalui data historis masa lalu yang dijadikan sebagai acuan untuk tahun selanjutnya agar perusahaan dapat mengambil keputusan yang tepat apabila produksi tanaman kelapa sawit mengalami penurunan serta perusahaan dapat terus meningkatkan hasil produksi kelapa sawit. Prediksi kelapa sawit bertujuan untuk meningkatkan produktivitas, efisiensi, dan keinginan industri kelapa sawit di Indonesia [13].

2.6 RapidMiner

RapidMiner adalah platform perangkat lunak yang kuat untuk ilmu data dan pembelajaran mesin. Ini menyediakan beragam alat untuk persiapan data, pemodelan, evaluasi, dan implementasi. RapidMiner dirancang untuk mudah digunakan dan memungkinkan pengguna untuk dengan mudah membangun dan menguji berbagai model, bahkan tanpa pengalaman pemrograman [14].

RapidMiner menawarkan antarmuka drag-and- drop yang memungkinkan pengguna untuk membangun alur kerja untuk memproses dan menganalisis data. Ini mendukung beragam sumber data, termasuk file datar, basis data, dan platform big data seperti Hadoop dan Spark. Perangkat lunak ini juga mencakup beragam operator yang sudah dibangun, yang merupakan blok bangunan dari alur kerja, yang mencakup semua tahap proses data mining, seperti pembersihan data, pemilihan fitur, dan pemodelan.

Gambar 2. 2 Rapidminer

(sumber images.search.yahoo.com)

2.7 Metodologi Penelitian

Dalam Kamus Besar Bahasa Indonesia (KBBI), metodologi adalah ilmu tentang metode; uraian tentang metode. Sedangkan pengertian penelitian dalam Kamus Besar Bahasa Indonesia (KBBI), penelitian adalah kegiatan pengumpulan, pengolahan, analisis, dan penyajian data yang dilakukan secara sistematis dan objektif untuk memecahkan suatu persoalan atau menguji suatu hipotesis untuk mengembangkan prinsip-prinsip umum. Metodologi penelitian merupakan cara untuk mengetahui hasil dari sebuah permasalahan yang spesifik dari suatu penelitian. Metodologi penelitian ini dilakukan dengan cara sistematis yang akan digunakan sebagai acuan dalam penelitian.

2.8 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

Referensi Penelitian 1	
Judul	Analisis Algoritma C4.5 dan Regresi Linear untuk Memprediksi Hasil Panen Kelapa Sawit
Nama Penulis	Nurahman, Nindi Ernawati
Tahun	2024
Hasil	Berdasarkan hasil penelitian prediksi dengan membandingkan algoritma C45 dan Regresi Linear yang telah dilakukan dalam meneliti hasil panen kelapa sawit PT Surya Inti Sawit

	Kahuripan, memperoleh hasil sebagai berikut. Perbandingan analisis performance yang dihasilkan menggunakan algoritma C45 dan Regresi Linear menghasilkan nilai Absolute Error C45 sebesar 0,085 dengan deviasi sebesar 0,189 dan nilai Absolute Error Regresi Linear sebesar 1,020 dengan deviasi sebesar 0,687. Nilai Root Mean Squared Error C45 sebesar 0,207 dengan deviasi sebesar 0,000 dan nilai Root Mean Squared Error Regresi Linear sebesar 1,230 dengan deviasi sebesar 0,000. Nilai Squared Error C4 sebesar 0,043 dengan deviasi sebesar 0,173 dan nilai Squared Error Regresi Linear sebesar 1,512 dengan deviasi sebesar 2,099. Dari hasil perbandingan nilai performane tersebut, algoritma dengan nilai RMSE yang lebih kecil adalah algoritma terbaik dan memiliki tingkat keakuratan yang lebih baik. Jadi, metode C45 lebih disarankan dalam melakukan prediksi hasil panen kelapa sawit pada PT Surya Inti Sawit Kahuripan karena memiliki angka error yang lebih kecil dibandingkan algoritma Regresi Linear. Kesimpulan dari faktor-faktor yang dipertimbangkan saat memilih algoritma terbaik bergantung pada analisis dataset dan kebutuhan prediksi. Menguji kedua algoritma dan membandingkan hasilnya dalam hal akurasi, interpretabilitas, dan kecepatan akan membantu dalam memahami model yang lebih sesuai untuk prediksi hasil panen kelapa sawit. C4.5 lebih baik dalam melakukan analisis hasil panen kelapa sawit karena memiliki tingkat nilai error lebih rendah dan lebih akurat dibandingkan dengan regresi linear.

Sumber [1]

Tabel 2. 2 Penelitian Terdahulu

Referensi Penelitian 2	
Judul	Prediksi Produksi Kelapa Sawit Menggunakan Metode Regresi Linear Berganda
Nama Penulis	Adji Prasetyo, Salahuddin, Amirullah
Tahun	2021
Hasil	Berdasarkan hasil penelitian yang dilakukan di PT. Perkebunan Nusantara I (PTPN I) dan hasil pengembangan sistem prediksi produksi kelapa sawit dengan teknik Machine learning. Sistem tersebut dapat memprediksi hasil produksi kelapa sawit dalam periode bulan dengan persentase ketepatan akurasi sistem yang diperoleh dengan menggunakan metode Mean Absolute Percentage Error sebesar 14.28 %.

Sumber [15]

Tabel 2. 3 Penelitian Terdahulu

Referensi Penelitian 3	
Judul	Implementasi Algoritma C4.5 Prediksi Produksi Komoditas Tanaman Perkebunan Berdasarkan Luas Lahan
Nama Penulis	Desi Damayanti
Tahun	2022
Hasil	Dapat disimpulkan bahwa dengan menggunakan algoritma C4.5 untuk melakukan perhitungan, ternyata mampu menyelesaikan permasalahan perusahaan dalam memprediksi produksi komodity tanaman perkebunan berdasarkan luas lahan pada Unit Pembinaan Perlindungan Tanaman (UPPT) Biru-biru. Penerapan algoritma C4.5 dapat memberikan informasi berupa rule prediksi untuk menggambarkan proses yang terkait dengan produksi komodity tanaman perkebunan.

Sumber [16]