

**SKRIPSI**

**IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM  
MENGKLASIFIKASI DATA CURAH HUJAN  
UPT. BMKG KOTA MEDAN**

**DISUSUN OLEH**

**NABILA AMELIA SIREGAR**

**NPM: 2009010091**



**UMSU**

Unggul | Cerdas | Terpercaya

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA  
MEDAN  
2024**

**IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM  
MENGKLASIFIKASI DATA CURAH HUJAN  
UPT. BMKG KOTA MEDAN**

**SKRIPSI**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer  
(S.Kom) dalam Program Studi Sistem Informasi pada Fakultas Ilmu Komputer dan  
Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara**

**NABILA AMELIA SIREGAR**

**NPM. 2009010091**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA**

**MEDAN**

**2024**

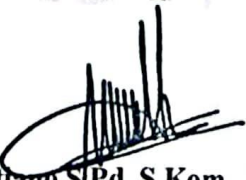
## LEMBAR PENGESAHAN

Judul Skripsi : IMPLEMENTASI ALGORITMA DECISION TREE C4.5  
DALAM MENGGASIFIKASI DATA CURAH HUJAN  
UPT. BMKG KOTA MEDAN  
Nama Mahasiswa : NABILA AMELIA SIREGAR  
NPM : 2009010104  
Program Studi : SISTEM INFORMASI


Menyetujui  
Dosen Pembimbing

  
(Dr. Zainal Azis, M.M., M.Si)  
NIDN. 0113126301

Ketua Program Studi

  
(Martiano SIPd, S.Kom., M.Kom)  
NIDN. 0128029302

Dekan

  
(Dr. Al-Khowarizmi, S.Kom., M.Kom.)  
NIDN. 0127099201

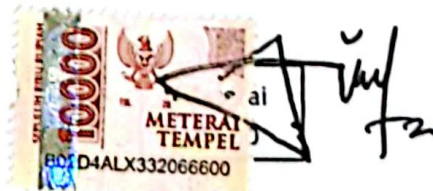
# PERNYATAAN ORISINALITAS

## IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM MENGKLASIFIKASI DATA CURAH HUJAN UPT. BMKG KOTA MEDAN SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 10 Oktober 2024

Yang membuat pernyataan



Nabila Amelia Siregar

NPM. 2009010091

**PERNYATAAN PERSETUJUAN PUBLIKASI**

**KARYA ILMIAH UNTUK KEPENTINGAN  
AKADEMIS**

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama : Nabila Amelia Siregar  
NPM : 2009010091  
Program Studi : Sistem Informasi  
Karya Ilmiah : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif (*Non-Exclusive Royalty free Right*) atas penelitian skripsi saya yang berjudul:

**IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM  
MENGKLASIFIKASI DATA CURAH HUJAN  
UPT. BMKG KOTA MEDAN**

Beserta perangkat yang ada (jika diperlukan) Dengan Hak Bebas Royalti NonEksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya

Medan, 10 Oktober 2024

Yang membuat pernyataan



Nabila Amelia Siregar

NPM. 20090100091

## **RIWAYAT HIDUP**

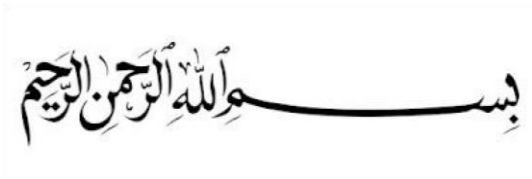
### **DATA PRIBADI**

Nama Lengkap : Nabila Amelia Siregar  
Tempat dan Tanggal Lahir : Medan, 22 Maret 2002  
Alamat Rumah : JL. Rakyat No. 109, Kec, Medan Perjuangan,  
Kel, Tegal Rejo  
Telepon/Faks/HP : 08984019988  
E-mail : nabilaamelia121@gmail.com  
Instansi Tempat Kerja : -  
Alamat Kantor : -

### **DATA PENDIDIKAN**

SD : SD Negeri 060874 TAMAT: 2014  
SMP : SMP Negeri 12 Medan TAMAT: 2017  
SMA : SMA PERTIWI Medan TAMAT: 2020

## KATA PENGANTAR



Puji syukur Alhamdulillah kehadiran Allah Subhanahu Wa Ta'ala yang telah memberikan Nikmat dan Karunia-Nya kepada penulis dalam menyusun skripsi sehingga penulis dapat menyelesaikan skripsi ini tepat pada waktunya, dengan judul “IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM MENGGASIFIKASI DATA PREDIKSI CURAH HUJAN UPT. BMKG KOTA MEDAN” Penulis tentunya berterima kasih kepada berbagai pihak dalam dukungan serta doa dalam penyelesaian skripsi. Penulis juga mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Agussani, M.AP. Rektor Universitas Muhammadiyah Sumatera Utara (UMSU)
2. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
3. Bapak Halim Maulana, S.T., M.Kom. Wakil Dekan I Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU yang telah menjadi dosen pembahas dalam seminar proposal skripsi penulis.
4. Bapak Martiano, S.Pd., S.Kom., M.Kom. Ketua Prodi Sistem Informasi Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
5. Bapak Dr. Zainal Azis M.M., M.Si. Dosen pembimbing penulis yang selalu meluangkan waktu dan dengan sabar memberikan arahan kepada penulis.
6. Cinta pertamaku, Bapak Isroq. Beliau memang tidak sampai merasakan pendidikan sampai bangku perkuliahan, namun beliau mampu mendidik penulis, memotivasi, memberikan

dukungan hingga penulis mampu menyelesaikan studinya sampai sarjanah.

7. Pintu Surgaku, Ibu Fadma Dewi. Beliau sangat berperan penting dalam mendapatkan gelar sarjanah pertama penulis. Beliau juga memang tidak sampai merasakan pendidikan sampai bangku perkuliahan, namun semangat dan doa-doa terbaik yang tiada henti selalu menemani langkah penulis. Terima kasih telah menjadi pendengar dan penasihat meski terkadang pikiran kita tidak sejalan. Mama akan selalu menjadi penyemangat utama dalam hidupku.
8. Adik-adik tercinta, Nadia Aprillia Siregar dan Muhammad Zidane Siregar. Terima kasih telah memberikan semangat, doa, serta dukungan yang tiada henti untuk penulis.
9. Terima kasih kepada diri saya sendiri (Nabila Amelia Siregar) yang telah melalui proses ini dengan penuh kesabaran dan ketekunan, meski banyak tantangan di sepanjang jalan. Ini adalah hasil dari kerja keras, semangat pantang menyerah, dan keyakinan bahwa saya bisa menyelesaikan apa yang telah dimulai. Semoga capaian ini menjadi motivasi untuk terus melangkah dan berkarya lebih jauh di masa depan.
10. Widharta Surya Alam, Huang Renjun, Aulia Shiddiq Asy-syifa, dan orang tersayang dan terkasih lainnya yang terlibat namun tidak bisa disebuti semuanya. Terima kasih telah menjadi tempatku berkeluh kesah selama skripsi ini dan tidak berhenti untuk memberi dukungan juga semangat dalam menyelesaikan skripsi ini tepat waktu.
11. Sahabat-sahabat terbaikku, Hassa Nia Harahap, Nurainada Khalidah, Nur Haini, Nazla Ayu Zafirah Ayumna, Yayang Tri Tami, Meilinda Wafa, Olivia Malinda (adik tingkatku yang beda fakultas), dan sahabat-sahabatku yang lain yang telah



memberikan semangat dan dukungan untuk menyelesaikan skripsi ini tepat waktu.

12. Anessa dan kimmy. Kucing dan kelinci imut dan menggemaskan yang sudah menjadi penyemangat penulis untuk menyelesaikan skripsi ini.
13. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat penulis ucapkan satu-persatu yang telah membantu penyelesaian skripsi ini.

**IMPLEMENTASI ALGORITMA DECISION TREE C4.5 DALAM  
MENGKLASIFIKASI DATA CURAH HUJAN  
UPT. BMKG KOTA MEDAN**

**ABSTRAK**

Curah hujan yang tidak teratur dan sulit diprediksi menjadi tantangan bagi perencanaan kegiatan di berbagai sektor. Penelitian ini bertujuan untuk menerapkan algoritma Decision Tree C4.5 dalam mengklasifikasikan data prediksi curah hujan di Kota Medan dengan memanfaatkan data cuaca dari BMKG. Metode C4.5 dipilih karena kemampuannya dalam menangani atribut data yang bersifat kategorikal maupun numerik, serta menghasilkan model yang mudah dipahami. Data yang digunakan mencakup variabel suhu maksimum, suhu minimum, suhu rata-rata, kelembapan rata-rata, dan curah hujan yang dikategorikan dalam beberapa tingkat intensitas.

Penelitian ini mencakup proses pra-pemrosesan data untuk memastikan data berada dalam format yang sesuai, termasuk penanganan nilai yang hilang dan kategorisasi variabel curah hujan. Setelah pra-pemrosesan, data dibagi menjadi dua bagian: data latih dan data uji. Algoritma C4.5 kemudian diterapkan untuk membangun model prediksi curah hujan. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, dan recall guna memastikan kinerja model dalam mengklasifikasikan curah hujan.

Hasil penelitian menunjukkan bahwa algoritma Decision Tree C4.5 berhasil mengklasifikasikan data curah hujan dengan akurasi yang memuaskan, memberikan informasi yang berguna untuk meningkatkan keakuratan prakiraan cuaca di Kota Medan. Oleh karena itu, model yang dihasilkan diharapkan dapat mendukung pengambilan keputusan yang lebih baik dalam menghadapi dampak perubahan cuaca di area penelitian.

**Kata kunci:** Prediksi curah hujan, Algoritma Decision Tree C4.5, BMKG, Kota Medan, klasifikasi cuaca.

# **IMPLEMENTATION OF THE DECISION TREE C4.5 ALGORITHM IN CLASSIFYING RAINFALL DATA UPT. BMKG KOTA MEDAN**

## **ABSTRACT**

Irregular and unpredictable rainfall poses significant challenges for planning activities across various sectors. This study focuses on applying the Decision Tree C4.5 algorithm to classify rainfall prediction data in Medan City, utilizing weather data sourced from BMKG. The C4.5 method was chosen for its capability to manage both categorical and numerical attributes while producing an interpretable model. The dataset comprises variables including maximum temperature, minimum temperature, average temperature, average humidity, and rainfall, which is categorized into several levels of intensity.

The study involves a data preprocessing phase to ensure that the data is in the appropriate format, addressing missing values and categorizing the rainfall variable. After preprocessing, the data is split into training and testing sets, where the C4.5 algorithm is applied to develop the rainfall prediction model. Model evaluation is conducted using accuracy, precision, and recall metrics to assess the model's performance in classifying rainfall.

The results indicate that the Decision Tree C4.5 algorithm can classify rainfall data with satisfactory accuracy, providing valuable information for improving the precision of weather forecasts in Medan City. The resulting model is expected to support better decision-making in response to the impacts of weather variability in the study area.

**Keywords:** Rainfall prediction, Decision Tree C4.5 algorithm, BMKG, Medan City, weather classification.

## DAFTAR ISI

ABSTRAK.....	ix
ABSTRACT.....	x
DAFTAR ISI.....	xi
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR.....	xiv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	5
1.3. Batasan Masalah.....	5
1.4. Tujuan Penelitian.....	6
1.5. Manfaat Penelitian.....	7
BAB II LANDASAN TEORI.....	9
2.1. Data Mining.....	9
2.2. Machine Learning.....	9
2.3. Metode Decision Tree.....	12
2.4. Algoritma C4.5.....	14
2.5. Evaluasi Model dengan Confusion Matrix.....	16
2.6. Curah Hujan.....	17
2.7. Prediksi Curah Hujan.....	19
2.8. Bahasa Pemrograman Python.....	20
2.9. Perangkat Lunak.....	21
2.9.1 Google Colab.....	21
2.9.2 Microsoft Excel.....	22
2.10. Penelitian Terdahulu.....	23
BAB III METODOLOGI PENELITIAN.....	29
3.1 Jenis Penelitian.....	29
3.2 Tempat dan Waktu Penelitian.....	29
3.3 Data dan Sumber Data.....	29
3.4 Teknik Pengumpulan Data.....	31
3.4.1 Observasi.....	32
3.4.2 Studi Literatur.....	33
3.5 Alat dan Bahan.....	34
3.6 Algoritma Decision Tree C4.5.....	34
3.6.1 Prinsip Kerja Algoritma Decision Tree C4.5.....	35

3.6.2 Konsep Entropi dan Information Gain .....	36
3.6.3 Pemangkasan Pohon (Pruning) .....	37
3.7 Proses Sistem Prediksi .....	37
3.7.1 Pelatihan Model .....	38
3.7.2 Pengujian Model .....	40
3.7.3 Evaluasi Hasil .....	41
<b>BAB IV HASIL DAN PEMBAHASAN.....</b>	<b>45</b>
4.1 Hasil Perhitungan Manual C4.5 .....	45
4.1.1 Handling Missing Values (Mengatasi Nilai yang Hilang) .....	45
4.1.2 Menghitung Probabilitas .....	49
4.1.3 Menghitung Entropy .....	50
4.1.4 Total Entropy .....	50
4.1.5 Interpretasi .....	51
4.2 Implementasi Algoritma.....	53
4.2.1. Preprocessing Data .....	54
4.2.2. Pembagian Data (Data Splitting) .....	58
4.2.3. Pelatihan Algoritma Decision Tree C4.5 .....	59
4.2.4. Prediksi Menggunakan Data Uji .....	60
4.2.5. Evaluasi Model .....	60
4.2.6 Visualisasi Pohon Keputusan .....	64
<b>BAB V SARAN DAN KESIMPULAN .....</b>	<b>69</b>
<b>DAFTAR PUSTAKA .....</b>	<b>73</b>

## DAFTAR TABEL

Tabel 1.1 Korban Dampak Banjir Kota Medan .....	2
Tabel 1.2 Korban yang Berhasil Dievakuasi .....	2
Tabel 2.1 Elemen <i>Confusion Matrix</i> . ....	16
Tabel 2.2 Penelitian Terdahulu. ....	24
Tabel 2.3 Flow Direction Symbol .....	27
Tabel 2.4 Processing Simbol.....	28
Tabel 2.5 Simbol DFD (Data Flow Diagram).....	30
Tabel 4.1 Tabel data curah hujan .....	48
Tabel 4.2 Tabel data curah hujan yang sudah dikategorikan .....	51
Tabel 4.3 Kategori Curah Hujan .....	52

## DAFTAR GAMBAR

Gambar 1. 1 Peta Administrasi Kota Medan. ....	3
Gambar 2. 1 Alur <i>Machine Learning</i> .....	10
Gambar 2. 2 Tampilan Antarmuka Google <i>Colab</i> .....	22
Gambar 2. 3 Tampilan <i>Micorsoft Excel</i> . ....	23
Gambar 3. 2 Halaman <i>Dashboard</i> Database BMKG .....	31
Gambar 2. 1 Tampilan Antarmuka Google <i>Colab</i> . ....	21
Gambar 2. 2 Tampilan <i>Micorsoft Excel</i> . ....	22
Gambar 3. 3 Tampilan Dataset <i>Excel</i> BMKG. ....	31
Gambar 3. 4 Flowchart Implementasi Algoritma .....	38
Gambar 4.1 Dataset yang masih missing value .....	49
Gambar 4-2 Data yang sudah di cleaning .....	50
Gambar 4.3 Flowchart Implementasi Algoritma .....	56
Gambar 4.4 Kode memanggil dataset .....	57
Gambar 4.5 Kode data cleaning .....	57
Gambar 4.6 Kode menghitung jumlah nilai kosong pada setiap kolom .....	59
Gambar 4.7 Kode memisahkan fitur (x) dan label (y) .....	60
Gambar 4.8 Kode memastikan data dalam format numerik yang benar .....	60
Gambar 4.9 Kode membagi data Train dan Test .....	60
Gambar 4.10 Kode data latih dan uji agar mudah diidentifikasi .....	60
Gambar 4.11 Kode membuat model algoritma .....	61
Gambar 4.12 Kode prediksi terhadap data testing .....	61
Gambar 4.13 Kode akurasi model .....	62
Gambar 4.14 Model yang menghasilkan akurasi sebesar 1.0 atau 100% .....	62
Gambar 4.15 Matriks Kebingungan ( <i>Confusion Matrix</i> ) .....	63
Gambar 4.16 Output dari Matriks kebingungan ( <i>Confusion Matrix</i> ) .....	63
Gambar 4.17 Kode evaluasi menggunakan <i>classification report</i> .....	64

Gambar 4.18 Output dari evaluasi menggunakan <i>classification report</i> .....	64
Gambar 4.19 output dari evaluasi menggunakan <i>classification report</i> .....	66
Gambar 4.20 Visualisasi Pohon Keputusan ( <i>Decision Tree</i> ) .....	67



# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Indonesia terletak di daerah iklim tropis, di mana sebagian besar wilayahnya memiliki tingkat kelembapan yang tinggi. Perubahan iklim yang tidak menentu menyebabkan kesulitan dalam memprediksi kondisi cuaca (Suma, B., 2020). Faktor-faktor yang mempengaruhi cuaca meliputi suhu, tekanan udara, kecepatan angin, tingkat kelembapan, dan curah hujan.

Lembaga pemerintah Indonesia yang dikenal sebagai Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) bertanggung jawab utama untuk melakukan pengamatan dan pengukuran terhadap berbagai gejala alam yang terjadi serta faktor-faktor yang memengaruhi cuaca, khususnya terkait dengan fenomena hujan. Proses pembuatan prakiraan cuaca menghadapi beberapa tantangan signifikan yang perlu diatasi. Pertama, untuk menghasilkan prakiraan cuaca yang akurat, diperlukan berbagai sumber data yang luas dan beragam, termasuk pengamatan langsung terhadap kondisi cuaca, citra satelit yang menunjukkan keadaan awan, serta hasil pemindaian radar. Kedua, keakuratan prakiraan cuaca sering kali sangat tergantung pada pengetahuan dan keterampilan para ahli dalam melakukan prakiraan, yang dapat menyebabkan variasi dalam hasil prakiraan antara satu waktu dengan waktu lainnya. Karena setiap prakiraan mungkin memiliki interpretasi data yang berbeda. Perbedaan pandangan ini dapat menciptakan kebingungan dalam pengambilan keputusan yang berkaitan dengan prakiraan curah, dan pada akhirnya bisa mengurangi kualitas informasi prakiraan curah hujan yang disampaikan kepada masyarakat luas. Hal ini dapat berdampak

negatif pada penerimaan informasi yang akurat oleh publik dan pengambilan keputusan yang didasarkan pada informasi tersebut (Rofiq, et al., 2020).

Curah hujan yang tidak terduga dan berubah-ubah menyebabkan kondisi cuaca di Indonesia kini semakin sulit, termasuk juga di Kota Medan yang berada pada koordinat 3° 30' – 3° 43' Lintang Utara dan 98° 35'–98° 44' Bujur Timur. Adanya hujan dengan intensitas tinggi yang menyebabkan meluapnya air dari sungai sehingga mengakibatkan banjir. Dampak terparah pernah terjadi di Kota Medan pada tahun 2020 yang menyebabkan 5.965 jiwa merasakan dampak dari banjir di kota medan. Dampaknya sebanyak 6 (enam) orang meninggal dunia, 181 jiwa yang berhasil dievakuasi. Korban yang berhasil dievakuasi antara lain, 50 jiwa anak-anak, 38 jiwa balita, 67 jiwa orang dewasa, dan 26 jiwa orang lanjut usia (BNPB.GO.ID, 2020)

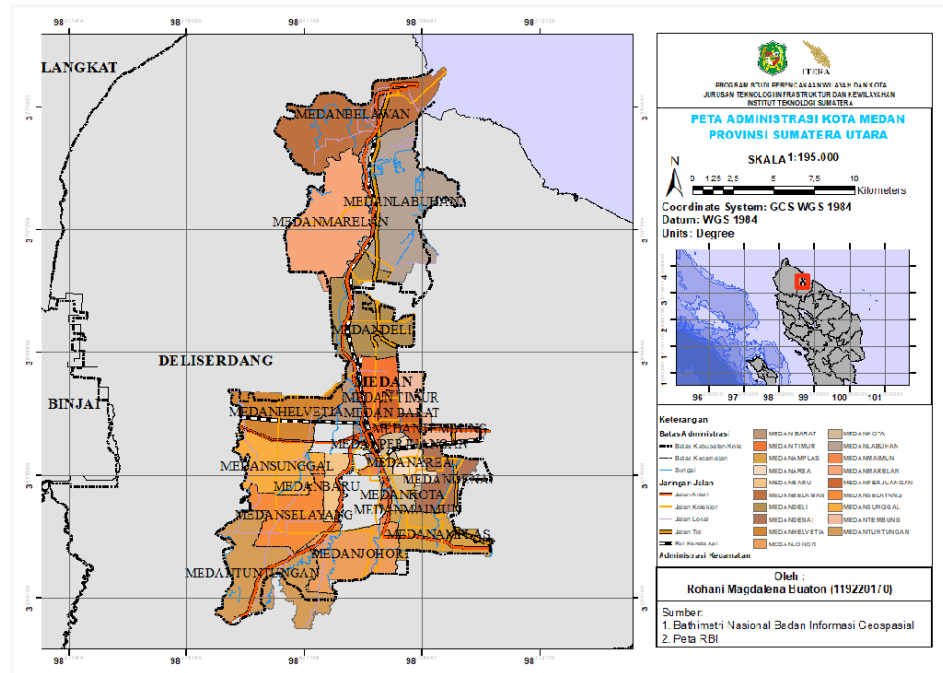
Tabel 1. 1 Korban Dampak Banjir Kota Medan

Sumber : BNPB, 2020

<b>Kondisi</b>	<b>Jumlah Korban (Jiwa)</b>
Meninggal	6
Evakuasi	181

Tabel 1. 2 Korban yang Berhasil Dievakuasi

<b>Usia</b>	<b>Jumlah Korban (Jiwa)</b>
Balita	38
Anak-Anak	50
Dewasa	67
Lanjut Usia	26



Gambar 1. 1 Peta Administrasi Kota Medan.

Perkiraan curah hujan sangat penting untuk mendukung berbagai aktivitas di banyak bidang, sehingga dapat menjadi pedoman dalam menjalankan kegiatan tanpa terpengaruh oleh cuaca buruk. Hal ini disebabkan oleh adanya faktor-faktor yang memengaruhi fluktuasi curah hujan. Oleh karena itu, prediksi curah hujan menjadi sangat krusial, terutama bagi mereka yang beraktivitas di luar ruangan (Nurdin et al., 2021). Selain itu, kondisi iklim yang tidak menentu saat ini mengakibatkan curah hujan berfluktuasi antara tingkat yang tinggi dan rendah, sehingga menyulitkan proses pemrediksian curah hujan (Muflih et al., 2019).

Curah hujan adalah ukuran ketinggian air hujan yang terkumpul di permukaan datar tanpa mengalami penguapan, penyerapan, atau aliran. Sebuah curah hujan sebesar 1 (satu) milimeter berarti bahwa pada area seluas satu meter persegi, terdapat genangan air setinggi satu milimeter atau setara dengan satu liter air. Sementara itu, intensitas curah hujan merujuk pada jumlah curah hujan yang terjadi dalam periode waktu tertentu. Jika intensitas curah hujan tergolong tinggi,

ini menunjukkan hujan yang sangat lebat, yang dapat berpotensi menyebabkan masalah serius seperti banjir, tanah longsor, dan dampak negatif bagi pertanian. Anomali iklim menyebabkan ketinggian yang tidak biasa, sehingga perubahan cuaca menjadi pemicu meningkatnya curah hujan yang sangat sulit untuk dipahami. Beberapa faktor memiliki dampak pada keragaman curah hujan yang dipengaruhi oleh faktor geografis. Kepulauan tersebut memiliki fitur geografis yang unik, termasuk orografis, topografi, orientasi, dan strukturnya yang sangat beragam (Arizona Saragih, I. J., 2020). Berbagai hal yang mempengaruhi sesuatu dampak dari pola distribusi hujan yang menjadi tidak seragam adalah hubungan antar wilayah yang berbeda (Fadholi. A., 2020).

Oleh karena itu diperlukan inovatif untuk menangani perkiraan tentang hujan. Salah satu cara yang dapat digunakan adalah dengan menggunakan pendekatan *Machine Learning* supervised learning pendekatan *Machine Learning* dapat digunakan untuk memprediksi curah hujan dengan menggunakan *Decision Tree*. Pohon keputusan (*Decision Tree*) ini dihasilkan dari sebuah aplikasi yang dibangun dengan menerapkan algoritma C4.5. Algoritma C4.5 adalah metode probabilistik untuk pengklasifikasian *Decision Tree* yang berguna dalam mengeksplorasi data dan mengidentifikasi hubungan tersembunyi antara berbagai variabel input dan variabel target (Kusrini et al., 2009).

Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma *Decision Tree* C4.5 dalam mengklasifikasikan data cuaca dari BMKG Kota Medan untuk memprediksi curah hujan. Dengan pendekatan ini, diharapkan dapat meningkatkan kualitas prediksi curah hujan dan memberikan kontribusi positif

terhadap pemanfaatan data cuaca dalam perencanaan dan pengelolaan sumber daya.

Berdasarkan pemaparan masalah di atas, dilakukan penelitian dengan judul **“IMPLEMENTASI ALGORITMA *DECISION TREE* C4.5 DALAM MENGGASIFIKASI DATA PREDIKSI CURAH HUJAN UPT. BMKG KOTA MEDAN”**. Penelitian ini diharapkan dapat memberikan pemahaman yang berharga bagi Balai Besar Meteorologi, Klimatologi, dan Geofisika Wilayah 01 Kota Medan dalam meningkatkan akurasi sistem prediksi curah hujan.

### **1.2. Rumusan Masalah**

Berdasarkan penjelasan latar belakang yang telah disampaikan sebelumnya, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana cara mengimplementasikan algoritma Decision Tree C4.5 dalam mengklasifikasikan data cuaca yang diperoleh secara online dari BMKG Kota Medan untuk memprediksi curah hujan?
2. Seberapa akurat hasil klasifikasi curah hujan menggunakan algoritma Decision Tree C4.5 berdasarkan lima kategori intensitas?
3. Apa saja faktor yang memengaruhi kinerja algoritma Decision Tree C4.5 dalam klasifikasi data curah hujan harian?

### **1.3. Batasan Masalah**

Adapun batasan-batasan yang diberlakukan dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan terbatas pada data harian curah hujan yang diperoleh secara daring dari situs resmi BMKG.

2. Fokus penelitian ini terbatas pada atribut yang digunakan hanya mencakup intensitas curah hujan yang dikategorikan menjadi lima tingkat, yaitu: Tidak Hujan, Hujan Ringan, Hujan Sedang, Hujan Lebat, dan Hujan Sangat Lebat.
3. Penelitian ini mengaplikasikan algoritma *Decision Tree C4.5* dengan fokus pada perhitungan entropi, tanpa melibatkan perhitungan gain dalam implementasinya.
4. Algoritma yang digunakan dalam penelitian ini adalah Decision Tree C4.5. Metode klasifikasi atau regresi lainnya, seperti Random Forest, SVM, atau Neural Networks, tidak akan dibandingkan atau diterapkan dalam penelitian ini dan Evaluasi model hanya akan dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score. Metode evaluasi lainnya, seperti validasi silang atau teknik statistik lanjutan, tidak akan diterapkan.
5. Analisis dan implementasi penelitian akan dilakukan menggunakan Google Colab dengan bahasa pemrograman Python. Platform atau bahasa pemrograman lain tidak akan dipertimbangkan atau diuji dalam penelitian ini.
6. Penelitian ini akan menggunakan data yang tersedia pada periode waktu tertentu dari BMKG saat penelitian dilakukan. Data historis jangka panjang atau data terbaru yang mungkin tersedia setelah penelitian dimulai tidak akan dianalisis.

#### **1.4. Tujuan Penelitian**

Merujuk pada rumusan masalah yang telah dijelaskan, tujuan dari penulisan penelitian ini adalah sebagai berikut:

1. Mengidentifikasi dan menganalisis dalam memodelkan data curah hujan secara online, termasuk faktor-faktor yang mempengaruhi kinerja model Decision Tree C4.5.
2. Menilai tingkat akurasi model klasifikasi curah hujan yang dibangun menggunakan data dari website BMKG.
3. Mengimplementasikan algoritma *Decision Tree C4.5* untuk mengklasifikasikan data curah hujan harian berdasarkan lima kategori intensitas yang diperoleh dari situs BMKG.
4. Mengevaluasi efektivitas model Decision Tree C4.5 dalam memprediksi curah hujan dengan menerapkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

### **1.5. Manfaat Penelitian**

Diharapkan penelitian ini dapat memberikan manfaat sebagai berikut:

1. Memberikan solusi yang lebih akurat dan andal untuk prediksi curah hujan, yang dapat membantu dalam membuat prakiraan cuaca yang lebih tepat.
2. Penelitian ini akan memperkaya literatur ilmiah mengenai penerapan algoritma Decision Tree C4.5 dalam prediksi curah hujan dengan data online, serta berkontribusi pada pengembangan metodologi baru dalam analisis cuaca.
3. Hasil dari penelitian ini, terutama dalam hal klasifikasi curah hujan yang dikategorikan, dapat dijadikan acuan atau model awal bagi penelitian selanjutnya yang bertujuan untuk mengembangkan model prediksi dengan lebih banyak variabel atau algoritma yang berbeda.





## **BAB II**

### **LANDASAN TEORI**

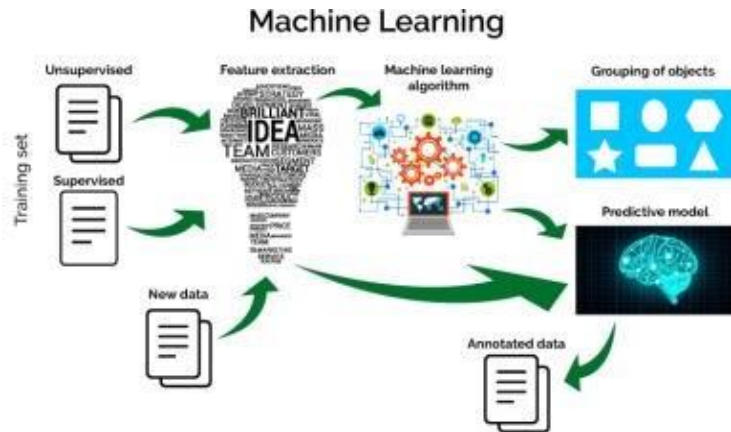
#### **2.1. Data Mining**

*Data Mining* atau DM adalah teknik yang menggunakan teknologi informasi dan metode statistika untuk mencari potensi informasi yang layak dari basis data yang besar yang dapat digunakan untuk mendukung keputusan administrative. Pentingnya penggunaan DM adalah data dapat diubah menjadi informasi yang berguna dan pengetahuan secara otomatis dan cerdas (Ba'abbad, I et al., 2021).

Menurut Kaur teknik *Data Mining* yang paling terkenal adalah Asosiasi atau *Assosiation*, kemudian *clustering* adalah teknologi *Data Mining* yang menciptakan sebuah kelompok objek yang berguna dengan memiliki fitur komparatif menggunakan strategi. *Decision Tree* adalah salah satu teknik *Data Mining* yang paling umum (Harkiran, K., 2017).

#### **2.2. Machine Learning**

*Machine Learning* (ML) merupakan salah satu jenis sistem kecerdasan buatan yang memungkinkan komputer untuk belajar sendiri tanpa diprogram sebelumnya. Secara umum, pekerjaan *Machine Learning* (ML) sering digunakan untuk mengelompokkan suatu permasalahan ke dalam beberapa kategori. Dalam kehidupan sehari-hari, manusia dapat mengenali objek dengan mudah, meskipun tidak selalu dapat menjelaskannya secara detail. Oleh karena itu, diperlukan *Machine Learning* untuk mengenali, mengidentifikasi, atau memprediksi data tertentu dengan mempelajari data historis (Nurhayati et al., 2019).



Gambar 2. 3 Alur *Machine Learning* \

Sumber : (Pantech, 2018).

Pada era 1940-an, ketika mesin komputer pertama kali dibuat, para ahli hanya berfokus pada cara melakukan komputasi dengan kecepatan maksimal. Berbagai teknik dan metode komputasi pun dikembangkan. Namun, belum ada gagasan untuk menciptakan program komputer yang dapat belajar seperti manusia. Saat itu, belum terpikirkan program yang memungkinkan komputer menjadi semakin pintar melalui pengalaman yang diperolehnya secara otomatis (Suyanto, 2018).

Menurut Suyanto (2018), pada awalnya komputer dirancang hanya untuk melakukan perhitungan. Namun, sejak tahun 1960-an, para ahli mulai mempertimbangkan fungsi lain yang lebih luas, yaitu kemampuan Belajar. Mereka berupaya mengembangkan konsep yang memungkinkan komputer belajar dari pengalaman sehingga menjadi lebih pintar.

Tom M. Mitchell dalam bukunya (Mitchell, 1997) menyatakan bahwa sebuah program komputer dikatakan belajar dari pengalaman  $E$  terkait dengan tugas-tugas tertentu  $T$  dan ukuran performa  $P$ , jika performanya pada tugas-tugas  $T$ , yang diukur menggunakan  $P$ , meningkat seiring dengan bertambahnya pengalaman  $E$  (Mitchell, 1997).

Pada era 1940-an, ketika mesin komputer pertama kali diciptakan, para ahli berfokus pada cara mempercepat proses komputasi. Beragam teknik dan metode komputasi pun dikembangkan. Namun, saat itu belum ada gagasan untuk merancang program komputer yang dapat belajar seperti manusia, yakni program yang secara otomatis dapat menjadi lebih pintar melalui pengalaman yang diperolehnya (Suyanto, 2018).

Tugas pembelajaran mesin biasanya diklasifikasikan ke dalam tiga kategori besar, tergantung pada sifat "sinyal" pembelajaran atau "umpan balik" yang tersedia untuk sistem pembelajaran. Ini adalah: (Goel et al., 2019)

### 1. *Supervised Learning*

Sebuah model atau algoritma diberikan contoh input dan output yang diinginkan untuk menemukan pola serta hubungan antara keduanya. Tujuannya adalah mempelajari aturan umum yang memetakan input ke output. Proses pelatihan ini berlanjut hingga model mencapai tingkat akurasi yang diinginkan pada data pelatihan. Beberapa contoh kehidupan nyata adalah:

- **Klasifikasi Gambar:** berlatih dengan gambar atau label sebagai contoh, model dilatih agar nantinya, ketika diberikan gambar baru, komputer diharapkan dapat mengenali objek baru tersebut.
- **Prediksi/Regresi Pasar:** melatih komputer menggunakan data pasar historis dan memintanya untuk memprediksi harga di masa depan.

### 2. *Unsupervised Learning*

Pada Unsupervised Learning tidak diberikan label pada algoritma pembelajaran, membiarkan algoritma tertentu menemukan pola sebagai masukannya. Dapat dimanfaatkan dalam mengelompokkan populasi yang memiliki perbedaan. Unsupervised Learning dapat mengelompokkan data tersendiri dan menemukan pola tersembunyi yang berada didalam data. 4 Pengelompokan: Anda meminta komputer untuk mengelompokkan data yang serupa ke dalam kategori, yang sangat penting dalam penelitian dan ilmu pengetahuan. Visualisasi Dimensi Tinggi: Gunakan komputer untuk membantu memvisualisasikan data yang memiliki banyak dimensi. Model Generatif: Setelah model memahami distribusi probabilitas dari data input, model tersebut dapat menghasilkan lebih banyak data, yang sangat berguna untuk memperkuat classifier Anda.

### 3. Reinforcement Learning

Sebuah agent yang dapat beradaptasi dengan lingkungan sekitar yang sangat dinamis di mana agent tersebut dapat memiliki tujuan tertentu (seperti berkendara), tanpa seorang guru secara eksplisit mengatakan apakah ia mendekati tujuannya atau 5 tidak. Contoh lain adalah belajar bermain game dengan bermain melawan lawan. (Bharadwaj et al., 2021)

#### 2.3. Metode Decision Tree

*Decision tree* merupakan model prediksi untuk suatu keputusan yang menggunakan struktur hirarkis atau pohon. Setiap pohon memiliki cabang yang mewakili atribut yang harus dipenuhi untuk melanjutkan ke cabang berikutnya, hingga mencapai daun yang menunjukkan akhir dari cabang tersebut. Dalam decision tree, data dinyatakan dalam bentuk tabel yang terdiri dari atribut dan

rekaman. Atribut berfungsi sebagai parameter yang ditetapkan sebagai kriteria dalam pembuatan pohon (Sartika et al., 2017).

Komponen-komponen dalam pohon keputusan pada umumnya adalah sebagai berikut :

1. Simpul (*Nodes*)

Dalam pohon keputusan terdapat tiga jenis simpul :

- a. Simpul Akar

Simpul ini mewakili pilihan yang akan membagi semua data menjadi dua atau lebih kelompok.

- b. Simpul Internal

Simpul ini adalah pilihan yang tersedia pada suatu titik dalam pohon dan terhubung dengan simpul induk dan simpul anak.

- c. Simpul Daun

Simpul hasil akhir dari kombinasi keputusan atau peristiwa.

2. Cabang (*Branches*)

Cabang menghubungkan satu simpul pohon ke simpul pohon lainnya dalam ruang lingkup pohon keputusan, dan cabang akan merepresentasikan aliran keputusan dari satu simpul pohon ke dalam satu simpul lainnya.

Dengan kata lain, pohon keputusan bekerja dengan membentuk pohon keputusan yang dapat disimpulkan aturan-aturan klasifikasi tertentu, salah satu algoritma yang menerapkan pohon keputusan adalah algoritma C4.5 (N, Azwant, 2018)

## 2.4. Algoritma C4.5

Salah satu metode klasifikasi yang digunakan dalam penelitian ini adalah klasifikasi pohon keputusan dengan Algoritma C4.5. Algoritma pohon keputusan merupakan algoritma yang kuat, populer, berbasis logika, dan mudah diinterpretasikan. Algoritma ini dapat diterapkan secara luas dalam berbagai kasus di bidang data mining, serta merupakan algoritma siap pakai yang memberikan kinerja baik dan cukup mudah dipahami (Lakshmi dkk, 2016). Algoritma C4.5 dikembangkan oleh Ross Quinlan sebagai pengembangan dari Algoritma ID3.

Beberapa hal yang membedakan algoritma C4.5 dari ID3 adalah:

1. Tahan terhadap data dengan noise.
2. Mampu menangani variabel dengan tipe diskret dan kontinu.
3. Mampu menangani variabel yang memiliki nilai yang hilang.
4. Dapat memangkas cabang-cabang dari pohon keputusan.

Secara umum, langkah-langkah algoritma C4.5 untuk membangun model pohon keputusan adalah sebagai berikut:

1. Memilih variabel sebagai akar.
2. Membuat cabang untuk setiap nilai.
3. Membagi kasus ke dalam cabang-cabang.
4. Mengulangi proses untuk setiap cabang hingga semua kasus pada cabang memiliki kelas yang sama.

Proses pemilihan atribut akar sendiri, didasarkan pada suatu nilai yang dinyatakan sebagai nilai *information gain* tertinggi dari atribut-atribut yang sudah ada. *Information Gain* adalah nilai yang digunakan untuk dapat mengukur efektivitas suatu atribut dalam mengklasifikasikan data, untuk mengetahui nilai

dari *Information Gain*, kita diharuskan menghitung *Entropy* terlebih dahulu. Dijelaskan juga bahwa, kita biasanya memakai *Entropy* sebagai parameter atau acuan untuk mengetahui keberagaman (heterogenitas) dari data. semakin beragam sampel pada data yang digunakan, maka nilai *Entropy* akan besar juga :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2.1)$$

Dimana :

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

|S<sub>i</sub>| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Untuk perhitungan *entropy* sendiri, dapat dilihat pada persamaan dibawah ini

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (2.2)$$

Dimana :

S : Kumpulan kasus

N : Jumlah partisi S

P<sub>i</sub> : Proporsi dari S<sub>i</sub> terhadap S [11].

## 2.5. Evaluasi Model dengan Confusion Matrix

Confusion matrix adalah alat evaluasi yang menunjukkan jumlah prediksi benar dan salah yang dihasilkan oleh model untuk setiap kelas. Komponen utama dari confusion matrix mencakup True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dari confusion matrix ini, kita dapat menghitung berbagai metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score, yang memberikan gambaran menyeluruh mengenai kinerja model.

### 1. Elemen Confusion Matrix:

Elemen-elemen utama dari confusion matrix meliputi:

- **True Positive (TP):** Jumlah instance yang benar-benar positif dan diprediksi sebagai positif.
- **True Negative (TN):** Jumlah instance yang benar-benar negatif dan diprediksi sebagai negatif.
- **False Positive (FP):** Jumlah instance yang sebenarnya negatif tetapi diprediksi sebagai positif (kesalahan tipe I).
- **False Negative (FN):** Jumlah instance yang sebenarnya positif tetapi diprediksi sebagai negatif (kesalahan tipe II).

Tabel 2.1 Elemen Confusion Matrix

Tabel 2. 1 Elemen *Confusion* Matrix.

	<b>Prediksi Positif</b>	<b>Prediksi Negatif</b>
<b>Aktual Positif</b>	TP	FN
<b>Aktual Negatif</b>	FP	TN

### 2. Metrik Evaluasi dari Confusion Matrix



Beberapa metrik evaluasi umum yang dapat dihitung dari confusion matrix adalah:

- **Akurasi:** Persentase prediksi yang benar dari seluruh prediksi yang dibuat oleh model.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots(2. 1)$$

- **Presisi:** Ketepatan dari prediksi positif yang dilakukan model.

$$Presisi = \frac{TP}{TP+FP} \dots\dots\dots(2. 2)$$

- **Recall:** Kemampuan model dalam menemukan semua contoh positif dalam data.

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots(2. 3)$$

- **F1-Score:** Rata-rata harmonis dari presisi dan recall.

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \dots\dots\dots(2. 4)$$

Confusion matrix dan metrik evaluasi yang dihasilkan memberikan wawasan mendalam tentang kekuatan dan kelemahan model dalam berbagai aspek klasifikasi. Evaluasi model dengan menggunakan confusion matrix memungkinkan identifikasi kesalahan spesifik dan peningkatan kinerja model melalui penyesuaian parameter atau pemilihan algoritma yang lebih tepat (Sammut & Webb, 2017).

## 2.6. Curah Hujan

Curah hujan adalah tinggi air hujan yang jatuh di area datar dan tidak mengalir (tenang). Jika air hujan tersebut sampai ke permukaan bumi, maka itu disebut hujan (Latifah et al., 2023). Namun, jika air keluar dari dasar awan tetapi

tidak jatuh hingga ke permukaan, itu disebut virga. Di Indonesia, curah hujan biasanya diukur dalam satuan milimeter. Ketika hujan mencapai satu milimeter atau lebih, hal ini menunjukkan bahwa satu meter persegi tanah datar tertutup air dengan kedalaman satu milimeter atau setara dengan satu liter atau lebih air yang terkumpul.

Curah hujan diukur dalam satuan milimeter. Curah hujan 1 (satu) milimeter berarti air hujan yang jatuh dengan tinggi 1 mm di atas permukaan seluas 1 m<sup>2</sup> tanpa menguap, meresap, atau mengalir. Dengan kata lain, ini setara dengan jumlah air hujan yang jatuh sebanyak 1 liter pada setiap luasan 1 m<sup>2</sup> (BMKG, 2020). Sebagai ilustrasi:

- a. Curah hujan 10 mm pada luasan 100 m<sup>2</sup> = 1000 liter air = 1 meter kubik air
- b. Curah hujan 100 mm pada luasan 1 Km = 100 juta liter air = 100 ribu meter kubik air

Klasifikasi besar curah hujan bulanan menurut BMKG (2010) adalah sebagai berikut:

- a. Rendah → 0-100 mm/bulan
- b. Sedang → 100-300 mm/bulan
- c. Tinggi → 300-500 mm/bulan
- d. Sangat Lebat → > 500 mm/bulan

Banyaknya curah hujan yang mencapai permukaan bumi atau tanah dalam jangka waktu tertentu dapat diukur dengan menghitung tinggi air hujan menggunakan metode tertentu. Hasil pengukuran ini disebut curah hujan, tanpa mempertimbangkan jenis atau bentuknya saat sampai di permukaan bumi, dan

tidak memperhitungkan endapan yang meresap ke dalam tanah, yang hilang karena penguapan, atau yang mengalir (Sideng et al., 2022).

Berdasarkan bentuk dan sifatnya, hujan dapat dibedakan menjadi dua jenis, yaitu shower atau hujan tiba-tiba. Hujan ini ditandai oleh permulaan dan akhir yang mendadak, dengan variasi intensitas yang umumnya cepat, serta tetesan air atau partikel yang lebih besar dibandingkan hujan biasa. Hujan shower biasanya berasal dari awan-awan Cumulus (Cu) atau Cumulonimbus (Cb) yang berkembang secara konvektif. Di sisi lain, hujan kontinyu memiliki permulaan dan akhir yang tidak mendadak, dan tidak terlihat pengurangan perawanan dari awal hingga akhir kejadian hujan. Hujan ini biasanya berasal dari awan-awan yang berbentuk merata, seperti awan Stratus (St), Altostratus (As), atau Nimbostratus (Ns).

## **2.7. Prediksi Curah Hujan**

Prediksi adalah upaya untuk memperkirakan apa yang akan terjadi di masa depan dengan memanfaatkan data relevan dari masa lalu (historis) yang diperoleh melalui metode ilmiah (Panggabean et al., 2020). Tujuan dari prediksi adalah untuk memahami kemungkinan kejadian di masa depan. Metode peramalan dapat diterapkan secara kuantitatif melalui perhitungan matematis atau secara kualitatif.

Beberapa sumber informasi yang dapat digunakan untuk melakukan prediksi bergantung pada apa yang ingin diramalkan. Untuk prediksi cuaca, sumber informasi dapat berupa data mengenai kondisi cuaca saat ini, seperti tekanan udara, kelembapan, dan temperatur. Data ini dapat diperoleh dari stasiun BMKG. Dalam memprediksi curah hujan, BMKG selama ini menggunakan metode subjektif, yaitu prediksi yang dibuat berdasarkan pertimbangan atau penilaian dari

prakiraan (Merdekawati & Ismail, 2019). Sistem prediksi dapat diterapkan dalam berbagai bidang, termasuk cuaca dan iklim, ekonomi dan keuangan, kesehatan, bisnis, dan banyak lagi. Tujuan utama dari sistem prediksi adalah untuk memberikan informasi yang berharga kepada pengguna agar mereka dapat membuat keputusan yang lebih baik dan lebih terinformasi di masa depan.

## **2.8. Bahasa Pemrograman Python**

*Python* adalah bahasa pemrograman tingkat tinggi yang digunakan untuk keperluan umum yang sudah digunakan dengan sangat luas untuk saat ini. *Python* adalah salah satu bahasa pemrograman yang banyak digunakan oleh perusahaan besar maupun pengembang untuk mengembangkan berbagai jenis aplikasi, baik berbasis desktop, web, maupun mobile. Python diciptakan oleh Guido van Rossum di Belanda pada tahun 1990, dan namanya diambil dari acara televisi kesukaan Guido, Monty Python's Flying Circus (Romzi & Kurniawan, 2020). Para pengembang akan menggunakan bahasa pemrograman *Python* sebagai bahasa utama untuk menjalankan algoritma ini, para pengembang juga akan menggunakan beberapa modul untuk pengerjaan algoritma ini, seperti *tkinter*, *pillow*, dan *random*.

Alasan menggunakan bahasa *Python* adalah karena bahasa ini merupakan bahasa pemrograman *high-level*, yang berarti bahasa ini mudah dibaca karena penulisannya sangat mirip dengan bahasa manusia sehari-hari, selain itu, bahasa ini mendukung fungsi-fungsi dan modul-modul yang diperlukan untuk penulisan algoritma ini, seperti *array splicing*, *index finder*, *random*, *class* dan lain-lain, selain itu bahasa ini juga ringan dan tidak perlu menggunakan Resource atau

sumber daya yang sangat besar, sehingga bahasa ini mudah dijalankan di mana saja.

Meskipun *Python* memiliki keuntungan yang banyak serta kelebihan yang melimpah, *Python* tetap memiliki kekurangan-kekurangan yang cukup fatal bagi beberapa *programmer* sehingga tidak semua posisi bisa ditutupi dengan bahasa pemrograman *Python*, salah satu contoh dari kekurangan bahasa pemrograman *Python* adalah dikarenakan bahasa *Python* merupakan bahasa dinamis maka ini membuat *Python* tidak secepat dan seefisien bahasa statis seperti bahasa pemrograman C (Wardana, 2019).

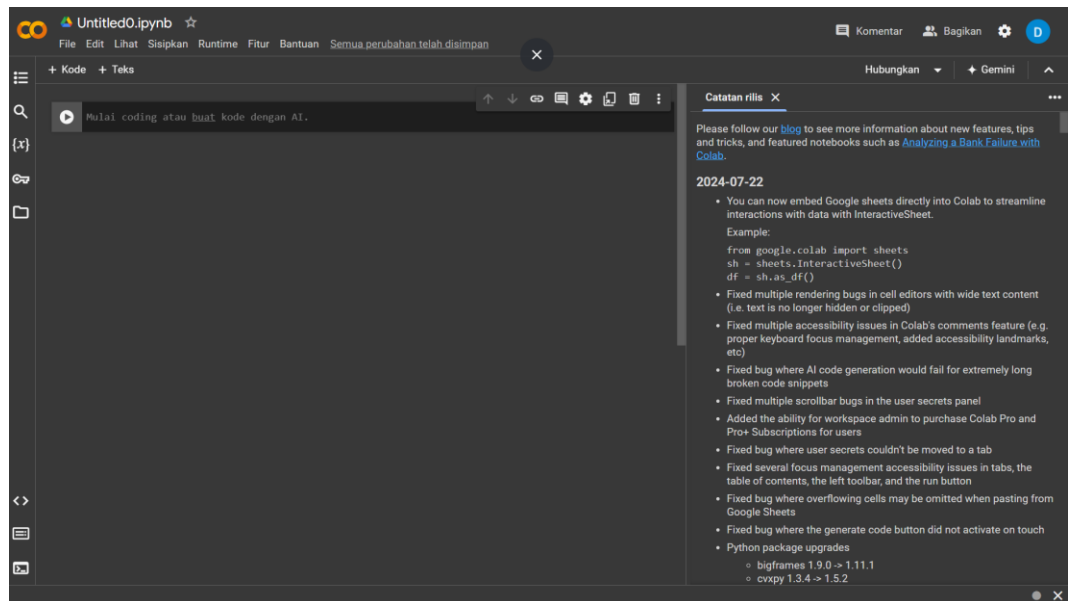
## **2.9. Perangkat Lunak**

Perangkat lunak merupakan serangkaian program komputer yang dirancang untuk menjalankan tugas-tugas tertentu pada perangkat keras komputer. Pada penelitian ini perangkat lunak yang digunakan yaitu, *Google Colab*, dan *Microsoft Excel*.

### **2.9.1 Google Colab**

*Google Colab* adalah platform berbasis cloud gratis yang disediakan oleh Google, memungkinkan pengguna untuk menulis dan menjalankan kode Python dalam lingkungan kolaboratif. Platform ini memberikan akses ke sumber daya GPU dan TPU, yang dapat mendukung keperluan Machine Learning dan data science (scaler, 2022).

Dalam konteks *data science*, *platform* ini menyediakan akses ke berbagai library Python yang sudah terinstal, seperti TensorFlow, PyTorch, dan OpenCV, yang memungkinkan analisis untuk mengembangkan dan menjalankan model Machine Learning dengan lebih cepat dan efisien (RevoU, 2024).

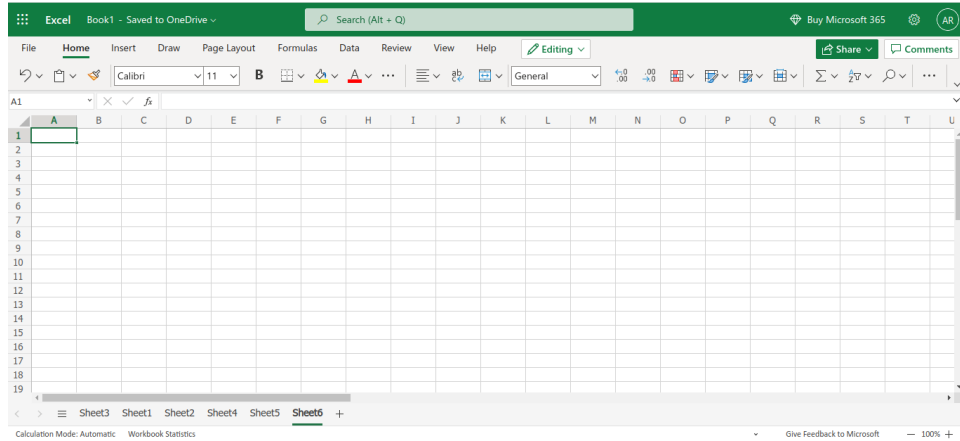


Gambar 2. 4 Tampilan Antarmuka Google Colab.

## 2.9.2 Microsoft Excel

Pada tahun 1982, Microsoft mengembangkan sebuah program spreadsheet dengan fitur kalkulasi yang canggih. Versi Excel untuk Macintosh dirilis pada tahun 1985, diikuti oleh versi untuk Windows pada tahun 1987. Microsoft Excel adalah aplikasi yang digunakan untuk mengolah data secara otomatis melalui berbagai cara, termasuk penggunaan rumus, perhitungan dasar, pengelolaan data, pembuatan tabel, pembuatan grafik, dan manajemen data.

Dalam penelitian ini, Microsoft Excel digunakan sebagai perangkat lunak untuk mengumpulkan data, membuat tabel, dan mengelola atribut.



Gambar 2. 5 Tampilan *Micorsoft Excel*.

## 2.10. Penelitian Terdahulu

Berikut adalah tabel penelitian terdahulu yang terkait dengan Implementasi Metode *Decision Tree* C4.5 pada Klasifikasi Data Prediksi Curah Hujan Menggunakan Data online UPT stasiun Balai Besar Meteorologi Klimatologi, dan Geofisika (BMKG) Wilayah 01 Kota Medan:

Tabel 2. 2 Penelitian Terdahulu.

No	Peneliti & Tahun	Judul Penelitian	Metode & Cara kerja	Kesimpulan
1	Ramdhani dan A. Mutamakin pada tahun 2022	"Implementasi Algoritma Decision Tree C4.5 dan Supports Vector Regression untuk Prediksi Penyakit Stroke"	Metode penelitian ini melibatkan pengumpulan dan analisis data untuk memprediksi stroke dengan menggunakan algoritma Support Vector Regression dan Decision Tree C4.5. Proses dimulai dengan pengumpulan dataset yang terdiri dari 5110 entri dan 12 fitur, dilanjutkan dengan preprocessing yang mencakup pembersihan data dan pembagian data dengan rasio 70:30 untuk pelatihan dan pengujian. Setelah itu, model dilatih menggunakan algoritma yang dipilih dan dievaluasi dengan mengukur performa model berdasarkan nilai error dari kedua algoritma menggunakan data uji.	Penelitian ini menunjukkan bahwa algoritma Decision Tree C4.5 memiliki kinerja yang lebih baik dalam memprediksi stroke dibandingkan dengan algoritma Support Vector Regression, dengan nilai error masing-masing sebesar 0.235 dan 0.399 pada rasio data 70:30. Penggunaan dataset yang terdiri dari 5110 entri dari Kaggle memungkinkan analisis yang lebih mendalam dan akurat. Hasil penelitian ini dapat digunakan sebagai acuan untuk mengembangkan strategi pencegahan stroke dan pengambilan keputusan medis.



No	Peneliti & Tahun	Judul Penelitian	Metode & Cara kerja	Kesimpulan
2	Dimas Imam Baihaqi, Anik Nur Handayani, Utomo Pujiyanto, pada tahun 2019 (Jurnal SIMETRIS, Vol. 10 No. 1 April 2019)	Perbandingan Metode Naïve Bayes dan C4.5 untuk Memprediksi Mortalitas pada Peternakan Ayam Broiler	Penelitian ini menggunakan metode algoritma C4.5 dan Naïve Bayes untuk klasifikasi tingkat kemiskinan. Cara kerjanya dimulai dengan preprocessing data, yang mentransformasi data mentah menjadi data berkualitas melalui pengklasteran. Selanjutnya, metode 10-Fold Cross Validation membagi data menjadi 10 subset, di mana 9 subset digunakan untuk pelatihan dan 1 untuk pengujian, diulang 10 kali untuk mendapatkan rata-rata akurasi, presisi, dan recall, serta evaluasi performa dilakukan menggunakan confusion matrix.	Penelitian ini menunjukkan bahwa metode C4.5 lebih unggul dibandingkan Naïve Bayes dalam memprediksi mortalitas pada peternakan ayam broiler, dengan akurasi, presisi, dan recall yang lebih tinggi. Hasil evaluasi menggunakan 10-Fold Cross Validation mengindikasikan bahwa C4.5 mampu memprediksi kebenaran lebih banyak dibandingkan Naïve Bayes. Oleh karena itu, C4.5 direkomendasikan sebagai metode yang lebih efektif untuk analisis mortalitas ayam broiler..
3	Ahmad Ari Aldino dan Heni Sulistiani, pada tahun 2020	Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat	Penelitian ini menggunakan algoritma C4.5 untuk klasifikasi penerima bantuan pendidikan, dengan langkah-langkah yang mencakup pengumpulan data sekunder dari mahasiswa, pra-pemrosesan data untuk menghapus data redundan dan mengubah format, serta pemilihan fitur penting seperti GPA, kompetisi, dan organisasi. Selanjutnya,	Algoritma C4.5 berhasil diterapkan untuk klasifikasi penerima bantuan pendidikan di Universitas Teknokrat Indonesia, dengan akurasi, presisi, dan recall masing-masing mencapai 87%. Model klasifikasi ini dapat diandalkan untuk merekomendasikan penerimaan beasiswa kepada mahasiswa berdasarkan kriteria yang telah

No	Peneliti & Tahun	Judul Penelitian	Metode & Cara kerja	Kesimpulan
		Indonesia)	klasifikasi dilakukan dengan menghitung nilai entropi dan information gain untuk membangun pohon keputusan hingga semua atribut terklasifikasi. Hasil penelitian menunjukkan bahwa algoritma C4.5 efektif dalam merekomendasikan penerima beasiswa.	ditentukan. Hasil penelitian ini menunjukkan bahwa penggunaan algoritma C4.5 dapat mempercepat dan mempermudah proses pengambilan keputusan dalam pemberian bantuan pendidikan.
4	Alfian Al Arif, Muhammad Firdaus, Rahmaddeni, dan Yustis Maruhawa, pada tahun 2022.	PERBANDINGAN METODE DATA MINING dengan Algoritma C4.5, Naïve Bayes dan KNN	Metode yang digunakan dalam penelitian ini adalah C4.5, Naïve Bayes, dan KNN untuk memprediksi curah hujan. C4.5 terbukti sebagai algoritma terbaik dengan akurasi mencapai 88,03% dan tingkat error sebesar 11,97%. Cara kerja metode ini melibatkan penggunaan data cuaca seperti suhu, kelembaban, lama penyinaran, dan kecepatan angin untuk memprediksi curah hujan. Algoritma C4.5 diterapkan untuk membangun model prediksi yang paling akurat.	Dari penelitian ini, dapat disimpulkan bahwa algoritma C4.5 merupakan yang paling efektif dalam memprediksi curah hujan, dengan akurasi mencapai 88,03% dan tingkat error sebesar 11,97%. Metode data mining ini dapat digunakan untuk memperkirakan tingkat curah hujan yang diinginkan dengan hasil yang memuaskan.
5	Ari Muzakir dan Rika Anisa Wulandari, pada tahun Mei 2016	DECISION TREE PREDIKSI HIPERTENSI KEHAMILAN	Metode yang digunakan dalam penelitian ini adalah Algoritma C4.5 dalam data mining untuk memprediksi penyakit hipertensi pada kehamilan.	Implementasi data mining dengan algoritma C4.5 dapat membantu memprediksi penyakit hipertensi pada kehamilan. Evaluasi menunjukkan

No	Peneliti & Tahun	Judul Penelitian	Metode & Cara kerja	Kesimpulan
			Evaluasi dilakukan dengan tingkat akurasi mencapai 92,6573%. Cara kerja metode ini melibatkan penggunaan data training untuk membangun decision tree yang dapat memprediksi penyakit hipertensi pada kehamilan. Evaluasi dilakukan dengan menggunakan supplied test set untuk mengukur tingkat akurasi prediksi.	tingkat akurasi sebesar 92,6573%. Kesimpulannya, metode ini efektif dalam memprediksi risiko hipertensi pada ibu hamil.

Dalam berbagai penelitian, algoritma C4.5 secara konsisten menunjukkan kinerja yang unggul dibandingkan dengan metode lainnya. Penelitian oleh Ramdhani dan A. Mutamakin (2022) menunjukkan bahwa C4.5 lebih efektif dalam memprediksi stroke dibandingkan dengan Support Vector Regression, dengan nilai error yang lebih rendah. Penelitian oleh Dimas Imam Baihaqi, Anik Nur Handayani, dan Utomo Pujianto (2019) menemukan bahwa C4.5 juga lebih baik daripada Naïve Bayes dalam memprediksi mortalitas ayam broiler, berdasarkan akurasi, presisi, dan recall.

Di Universitas Teknokrat Indonesia, penelitian Ahmad Ari Aldino dan Heni Sulistiani (2020) menunjukkan bahwa C4.5 efektif dalam klasifikasi penerima bantuan pendidikan, dengan akurasi, presisi, dan recall mencapai 87%. Selanjutnya, penelitian oleh Alfian Al Arif, Muhammad Firdaus, Rahmaddeni, dan Yustis Maruhawa (2022) menunjukkan bahwa C4.5 adalah algoritma terbaik dalam memprediksi curah hujan, dengan akurasi 88,03% dan error 11,97%. Terakhir, penelitian Ari Muzakir dan Rika Anisa Wulandari (2016) membuktikan bahwa C4.5 efektif dalam memprediksi risiko hipertensi pada ibu hamil, dengan tingkat akurasi sebesar 92,6573%.

Secara keseluruhan, algoritma C4.5 terbukti unggul dalam berbagai aplikasi, termasuk dalam prediksi kesehatan, mortalitas hewan, klasifikasi pendidikan, dan cuaca. Keunggulan C4.5 terletak pada akurasi dan kemampuannya dalam mengolah data, yang dapat mempercepat dan mempermudah proses pengambilan keputusan di berbagai bidang.

## **BAB III**

### **METODOLOGI PENELITIAN**

#### **3.1. Jenis Penelitian**

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen Tujuan penelitian ini adalah untuk menggunakan algoritma Decision Tree C4.5, yang dipilih karena kemampuannya untuk menangani data numerik dan kategorikal serta menghasilkan pohon keputusan yang dapat diinterpretasikan secara visual. Algoritma ini digunakan untuk mengklasifikasikan prediksi curah hujan berdasarkan data cuaca online BMKG, khususnya di Kota Medan.

Metode eksperimen digunakan untuk menguji model pada data cuaca historis; Pendekatan ini memungkinkan peneliti untuk menyelidiki bagaimana variabel cuaca seperti suhu dan kelembapan meredam jumlah hujan yang terjadi . Dengan demikian, penelitian ini akan menghasilkan model yang dapat memprediksi jumlah hujan yang mungkin berdasarkan sejumlah parameter cuaca.

#### **3.2. Tempat dan Waktu Penelitian**

Studi tersebut dilakukan di kantor Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).yang terletak di Medan, Sumatera Utara., Pada tanggal 31 Januari 2024.

#### **3.3 Data dan Sumber Data**

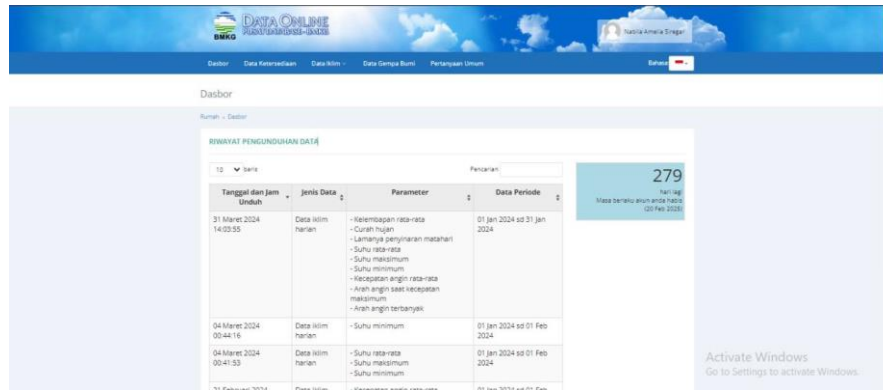
Penelitian ini memanfaatkan data sekunder yang diambil secara daring(online) dari situs resmi Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Data yang digunakan merupakan data harian mengenai curah hujan di Kota Medan dari tahun-tahun sebelumnya hingga tahun 2023. Data ini mencakup informasi tentang intensitas curah hujan yang Data curah hujan dikategorikan

menjadi lima kelas untuk memudahkan analisis dan prediksi. Kategori-kategori ini meliputi Tidak Hujan, Hujan Ringan, Hujan Sedang, Hujan Lebat, dan Hujan Sangat Lebat. Dengan pengelompokan ini, diharapkan mampu memberikan pemahaman yang lebih mendalam tentang intensitas curah hujan yang terjadi, sehingga informasi ini dapat dimanfaatkan untuk perencanaan dan pengambilan keputusan yang lebih efektif di berbagai sektor.


Kategorisasi ini dilakukan untuk memudahkan proses klasifikasi dengan memanfaatkan algoritma Decision Tree C4.5 untuk meramalkan kemungkinan terjadinya curah hujan dengan intensitas tertentu. Kategori tersebut ditetapkan sebagai berikut:

- **Tidak Hujan:** 0 - 4 mm/hari
- **Hujan Ringan:** 5 - 20 mm/hari
- **Hujan Sedang:** 21 - 50 mm/hari
- **Hujan Lebat:** 51 - 100 mm/hari
- **Hujan Sangat Lebat:** >100 mm/hari

Data curah hujan ini diambil dalam rentang waktu yang cukup panjang untuk mencakup berbagai variasi iklim yang berlangsung sepanjang tahun, termasuk musim penghujan dan musim kering, sehingga mampu menggambarkan pola curah hujan di Kota Medan. Pemilihan atribut curah hujan saja dimaksudkan untuk menyederhanakan model prediksi, dengan fokus pada intensitas hujan sebagai variabel utama.



Gambar 3. 1 Halaman *Dashboard* Database BMKG.

	A	B	C	D	E	F	G	H	I	J	K
1		ID WMO	: 96041								
2		Station Name	: Balai Besar Meteorologi Klimatologi dan Geofisika Wilayah I								
3		Latitude	: 3.53970								
4		Longitude	: 98.64000								
5		Elevation	: 0								
6											
7											
8											
9		Date	Tn	Tx	Tavg	RH_avg	RR		Notes :		
10		01-01-2023	21,7	31	28,5	71	8888		8888: Unmeasured Data		
11		02-01-2023	23,2	32,7	29,1	71	0		9999: No Data		
12		03-01-2023	23,4	32,2	28,8	78	0		Tn: Minimum temperature (°C)		
13		04-01-2023	23,2	31,8	28,6	78	23,3		Tx: Maximum temperature (°C)		
14		05-01-2023	23,2	30,7	28,2	78	0,2		Tavg: Average temperature (°C)		
15		06-01-2023	23,8	31,2	28,8	75	8888		RH_avg: Average humidity (%)		
16		07-01-2023	23,6	30,8	28	78	0,7		RR: Rainfall (mm)		
17		08-01-2023	23,8	30,4	27,7	77	8888				
18		09-01-2023	22,4	32,5	29,3	69	0,3				
19		10-01-2023	24	31,2	28,8	76					
20		11-01-2023	23,6	30,2	27,7	83	11,3				
21		12-01-2023	22,6	32,2	27,9	81	1,5				
22		13-01-2023	23,1	31,6	28,5	79	8,1				
23		14-01-2023	23,4	32	28,5	80					
24		15-01-2023	23,4	32	28,5	80					

Gambar 3.2 Tampilan Dataset *Excel* BMKG.

Dapat dilihat pada Gambar 3.3, di mana terdapat contoh data yang digunakan sebagai referensi informasi dalam proses pelatihan (data training) dan pengujian (data testing) dengan memanfaatkan metode yang telah ditentukan. Data tersebut berfungsi sebagai dasar untuk mengembangkan model prediksi yang lebih akurat. *Decision Tree*..

### 3.4 Teknik Pengumpulan Data

Perlu diketahui bahwa hasil akhir suatu penelitian sangat dipengaruhi oleh kualitas data yang dikumpulkan. Oleh karena itu, dalam pengumpulan data perlu

diperhatikan berbagai teknik yang ada agar pengumpulan data dapat dilakukan secara optimal. Dalam kerangka ini, dua teknik pengumpulan data berbeda digunakan dalam penelitian ini: observasi dan studi literatur. Dengan menggunakan kombinasi teknik-teknik tersebut, diharapkan data yang dikumpulkan dapat memberikan wawasan yang lebih dalam mengenai fenomena yang diteliti dan memungkinkan analisis yang lebih komprehensif dan akurat untuk mencapai tujuan penelitian.

### **3.4.1 Observasi**

Dalam penelitian, observasi adalah memusatkan perhatian pada suatu objek dengan menggunakan seluruh indera untuk mengumpulkan data. Ini adalah pengamatan langsung melalui indera seperti penglihatan, persepsi, pendengaran, perabaan, atau, jika diperlukan, pengecap. Dengan menggunakan observasi, peneliti dapat melihat dan mengamati objek penelitian secara langsung, yang memungkinkan mereka untuk mencatat dan mengumpulkan data yang diperlukan untuk mengungkap hasil penelitian mereka. (Thalha et al., 2019).

Dalam konteks ini, penulis melakukan observasi atau pengamatan langsung di Kantor Stasiun Balai Besar Meteorologi, Klimatologi, dan Geofisika Wilayah 01 Kota Medan untuk mendapatkan informasi yang berguna sebagai dasar untuk memulai penelitian. Informasi yang didapatkan dari proses observasi ini adalah tercatat variasi cuaca yang signifikan di wilayah Kota Medan, dengan pola curah hujan yang beragam tergantung pada faktor-faktor geografis dan topografi. Observasi visual terhadap awan juga menunjukkan adanya perubahan yang cepat dalam formasi awan, yang dapat menjadi indikator potensi curah hujan di wilayah studi.



Peneliti melakukan observasi partisipan dan non-partisipan secara berulang dalam konteks permasalahan yang sedang diteliti. Tujuan dari observasi yang berulang ini adalah agar responden menjadi terbiasa dengan kehadiran peneliti, sehingga mereka dapat berperilaku secara alami dan mengungkapkan budaya yang sesungguhnya tanpa dibuat-buat. Observasi ini dilakukan dengan merekam dan memotret kegiatan yang terjadi, yang nantinya akan dijadikan sebagai bahan analisis lebih lanjut dalam penelitian ini.

### **3.4.2 Studi Literatur**

Teknik pengumpulan data penelitian sastra yang disebut juga dengan penelitian kepustakaan adalah pencarian literatur yang memuat teori-teori dalam karya ilmiah terbitan maupun tidak terbitan baik dalam bentuk buku hardcopy maupun artikel dan jurnal di Internet.

Pada penelitian ini peneliti menggunakan studi literatur teknik ini digunakan karena memiliki peran penting dalam penelitian karena dalam literatur terdahulu, telah dilakukan berbagai penelitian yang mengimplementasikan metode regresi linear untuk memprediksi curah hujan dengan menggunakan data cuaca dari berbagai sumber, termasuk data dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Studi-studi tersebut memberikan dasar teoretis dan praktis yang kuat untuk pengembangan model prediksi curah hujan yang lebih akurat dan dapat diandalkan. Namun, sebagian besar penelitian ini cenderung difokuskan pada wilayah tertentu yang berbeda dari wilayah Kota Medan, sehingga memberikan motivasi untuk mengadaptasi metode tersebut dalam konteks geografis yang berbeda.

### 3.5 Alat dan Bahan

Perangkat yang digunakan dalam penelitian ini terdiri dari komponen hardware dan software untuk pengumpulan serta pengolahan data, sebagai berikut:

- **Perangkat Lunak:**
  - **Python:** Bahasa pemrograman utama untuk implementasi algoritma, pengolahan data, dan visualisasi.
  - **Libraries:**
    - **Pandas:** Untuk manipulasi dan analisis data.
    - **NumPy:** Untuk komputasi numerik.
    - **Scikit-learn:** Untuk membangun dan mengevaluasi model machine learning..
    - **Matplotlib:** Untuk visualisasi data.
  - **Google Colab:** Sebagai lingkungan pengembangan interaktif.
- **Hardware:** Komputer dengan spesifikasi minimal: Prosesor Intel Core i3, RAM 4GB, dan penyimpanan SSD 256GB.
- **Data:** Data curah hujan, suhu udara, dan kelembaban harian diperoleh dari situs web BMKG dalam format CSV.

### 3.6 Algoritma Decision Tree C4.5

Algoritma *Decision Tree* C4.5 merupakan perbaikan dari algoritma ID3. (Iterative Dichotomiser 3) yang digunakan untuk tugas klasifikasi. C4.5 adalah salah satu algoritma yang paling populer dalam data mining karena kemampuannya menangani atribut numerik dan menangani data yang memiliki

nilai hilang (missing values). C4.5 membangun sebuah pohon keputusan dengan memecah dataset menjadi beberapa bagian berdasarkan atribut yang memberikan *information gain* tertinggi.

### 3.6.1 Prinsip Kerja Algoritma Decision Tree C4.5

Pohon keputusan dibentuk melalui pembagian data secara berulang, dengan memilih atribut yang paling relevan untuk membedakan kelas target (dalam hal ini, prediksi curah hujan). Algoritma ini bekerja dengan melakukan proses berikut:

1. **Pemilihan Atribut Berdasarkan Entropi:** Pada setiap langkah, C4.5 menghitung nilai entropi untuk setiap atribut. Entropi mengukur ketidakpastian dalam pengelompokan data. Atribut yang memiliki entropi terendah (artinya paling sedikit ketidakpastian) akan dipilih untuk menjadi node pada pohon. Dengan memilih atribut berdasarkan entropi, algoritma dapat memisahkan data ke dalam kategori curah hujan dengan lebih akurat.
2. **Pembentukan Node dan Cabang Pohon:** Setiap node mewakili atribut yang dipilih berdasarkan entropi. Setiap cabang yang keluar dari node tersebut mewakili salah satu nilai atau rentang nilai atribut yang ada. Proses ini berlanjut sampai semua data dapat diklasifikasikan dengan baik, atau sampai tidak ada atribut yang tersisa untuk dipisahkan.
3. **Penerapan Pruning (Pemangkasan Pohon):** Salah satu kelebihan dari algoritma C4.5 adalah kemampuannya untuk melakukan pruning atau pemangkasan pohon. Pemangkasan pohon dilakukan untuk mengurangi

overfitting, yaitu ketika model mengalami overfitting terhadap data pelatihan sehingga kinerjanya buruk pada data uji. Proses pruning menghapus cabang-cabang yang tidak memberikan kontribusi signifikan pada akurasi prediksi.

### 3.6.2 Konsep Entropi dan Information Gain

Untuk memilih atribut terbaik, algoritma C4.5 menggunakan konsep *entropy* dan *information gain*. Entropy mengukur tingkat ketidakpastian atau kekacauan dalam dataset. Atribut yang meminimalkan entropy setelah pemisahan dianggap sebagai atribut yang terbaik.

Rumus entropy adalah sebagai berikut:

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

Di mana  $p_i$  adalah proporsi dari elemen-elemen kelas  $i$  dalam dataset  $S$ .

Setelah menghitung entropy, *information gain* dihitung untuk setiap atribut, yaitu selisih antara entropy awal sebelum pemisahan dan entropy rata-rata setelah pemisahan:

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

Fitur dengan nilai gain tertinggi akan dipilih sebagai node dalam pohon keputusan..

### 3.6.3 Pemangkasan Pohon (Pruning)

Pemangkasan pohon bertujuan untuk menyederhanakan struktur pohon keputusan tanpa mengurangi akurasi prediksi pada data uji. Pruning dilakukan dengan menghapus cabang-cabang pohon yang dianggap tidak signifikan. Ada dua metode pruning yang umum digunakan:

- Pre-pruning

Pohon dipangkas selama proses pembentukan, yaitu dengan menghentikan pemisahan ketika kriteria tertentu terpenuhi (misalnya ketika node sudah memiliki jumlah data yang sedikit).

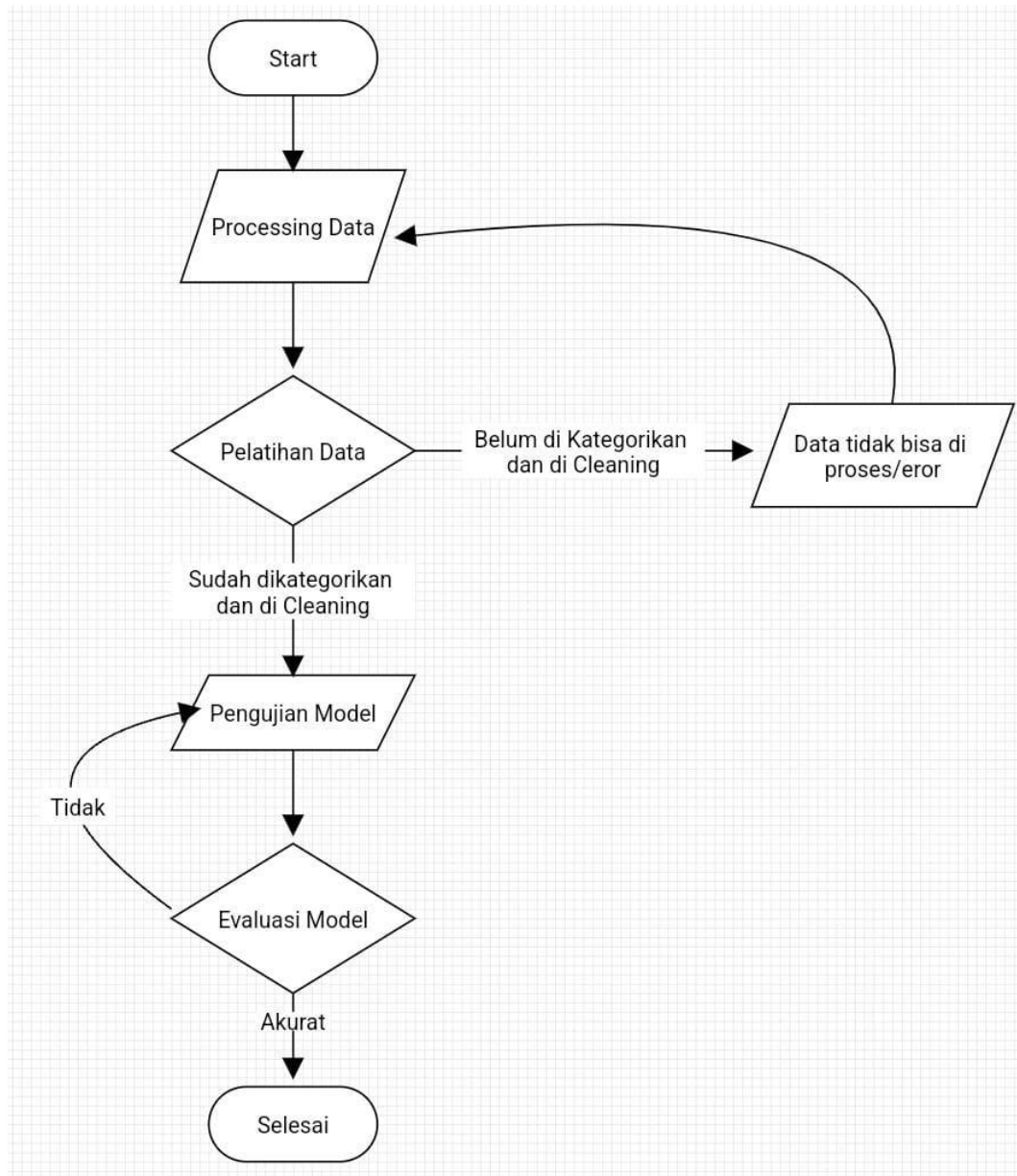
- Post-pruning

Dilakukan setelah pohon terbentuk, di mana cabang-cabang yang memberikan kontribusi kecil terhadap akurasi model dipangkas. Dengan melakukan pruning, pohon keputusan menjadi lebih sederhana dan lebih mudah dipahami.

## 3.7 Proses Sistem Prediksi

Sistem ramalan curah hujan yang dibuat dalam penelitian ini mengandalkan algoritma Decision Tree C4.5, yang merupakan salah satu metode machine learning yang paling populer dan efektif dalam menangani data kategorik. Keunggulan utama dari algoritma ini adalah kemampuannya untuk menghasilkan model yang mudah dipahami dan diinterpretasikan, sehingga memudahkan analisis dan pengambilan keputusan berdasarkan hasil prediksi. Proses sistem prediksi ini melibatkan beberapa tahapan penting yang saling

terkait, yaitu pelatihan model, pengujian model, dan evaluasi hasil, yang masing-masing memiliki peranan krusial dalam memastikan keberhasilan sistem.



Gambar 3.4 Flowchart Proses Sistem Prediksi

### 3.7.1 Pelatihan Model

Pelatihan model merupakan langkah awal dan sangat penting dalam membangun sistem prediksi yang efektif. Dalam tahap ini, data latih yang telah

dibersihkan dan dikategorikan digunakan untuk melatih algoritma Decision Tree C4.5. Proses pelatihan ini meliputi beberapa langkah berikut :

1. Inisialisasi Algoritma C4.5

Pada tahap ini, algoritma C4.5 diinisialisasi dengan menerima input berupa dataset latih yang telah melalui proses pembersihan. Algoritma kemudian mulai membangun struktur pohon keputusan dengan cara menganalisis atribut-atribut yang ada dan menghitung nilai entropy serta information gain untuk masing-masing atribut. Proses ini bertujuan untuk menentukan atribut mana yang paling informatif dalam memprediksi kategori curah hujan.

2. Pembentukan Pohon Keputusan

Setelah inisialisasi, algoritma melanjutkan dengan proses pembentukan pohon keputusan. Pada tahap ini, setiap atribut dalam dataset dievaluasi dengan cara menghitung nilai entropy dan information gain. Atribut yang menghasilkan information gain tertinggi akan diambil sebagai akar dari pohon keputusan. Proses ini berlangsung secara rekursif, di mana untuk setiap node, algoritma terus mencari atribut terbaik hingga semua data dapat dikategorikan. Kriteria penghentian dapat berupa kedalaman maksimum dari pohon, jumlah minimum data dalam node, atau mencapai nilai tertentu dari information gain.

3. Pemangkasan Pohon (Pruning)

Setelah pohon keputusan terbentuk, langkah berikutnya adalah pemangkasan untuk menghindari masalah overfitting, di mana model

terlalu kompleks dan tidak dapat generalisasi dengan baik pada data baru. Pemangkasan dilakukan dengan cara menghapus cabang-cabang dari pohon yang tidak memberikan kontribusi signifikan terhadap akurasi model. Proses ini bertujuan untuk menjaga keseimbangan antara kompleksitas model dan kemampuannya dalam melakukan prediksi yang akurat.

### **3.7.2 Pengujian Model**

Setelah model dilatih, tahap berikutnya adalah menguji kinerja model dengan memanfaatkan dataset yang terpisah, yaitu data uji. Data uji ini sangat penting karena digunakan untuk menilai kemampuan model dalam memprediksi kategori curah hujan berdasarkan pohon keputusan yang telah dibentuk. Proses pengujian meliputi langkah-langkah berikut:

#### **1. Prediksi Kategori Curah Hujan**

Setiap entri dalam dataset uji akan dievaluasi dengan menggunakan pohon keputusan yang dihasilkan selama pelatihan. Model akan menganalisis setiap atribut dari data uji dan memberikan kategori prediksi berdasarkan aturan yang telah dipelajari sebelumnya. Langkah ini penting untuk memastikan bahwa model mampu beradaptasi dengan data yang belum pernah ditemui sebelumnya dan memberikan hasil yang konsisten.

#### **2. Pengumpulan Hasil Prediksi**

Hasil prediksi yang dihasilkan untuk setiap entri dalam data uji dicatat dan dibandingkan dengan kategori yang sebenarnya dari dataset. Data ini



kemudian digunakan untuk menghitung metrik evaluasi, yang akan memberikan gambaran mengenai kinerja model dalam memprediksi kategori curah hujan.

### 3.7.3 Evaluasi Hasil

Evaluasi hasil merupakan tahap akhir yang krusial dalam proses sistem prediksi, di mana kinerja model diukur menggunakan berbagai metrik evaluasi. Proses ini tidak hanya penting untuk menilai akurasi model, tetapi juga memberikan wawasan mendalam mengenai kemampuan model dalam memprediksi kategori curah hujan dengan benar. Dalam penelitian ini, metrik yang digunakan untuk evaluasi mencakup:

#### 1. Akurasi

Akurasi adalah metrik dasar yang dihitung sebagai perbandingan antara jumlah prediksi yang benar dan total jumlah prediksi yang dilakukan oleh model. Metrik ini memberikan gambaran umum mengenai seberapa efektif model dalam melakukan prediksi. kategori curah hujan. Secara matematis, akurasi dapat dinyatakan dengan rumus sebagai berikut:

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Total\ Jumlah\ Prediksi} * 100\%$$

Tingginya nilai akurasi menunjukkan bahwa model dapat dengan efektif membedakan antara kategori-kategori curah hujan yang ada. Namun, perlu diingat bahwa akurasi saja tidak cukup untuk memberikan penilaian yang

komprehensif, terutama dalam situasi di mana terdapat ketidakseimbangan kelas yang signifikan.

## 2. Presisi dan Recall

Dalam konteks klasifikasi multikelas, evaluasi kinerja model tidak dapat hanya bergantung pada akurasi. Oleh karena itu, metrik presisi dan recall juga digunakan. Presisi mengukur proporsi prediksi positif yang benar dibandingkan dengan semua prediksi positif yang dibuat oleh model. Metrik ini penting untuk menilai seberapa akurat model dalam menghasilkan prediksi yang tepat tanpa menghasilkan banyak kesalahan. Rumus presisi dapat dituliskan sebagai:

$$\text{Presisi} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Sementara itu, recall mengukur proporsi aktual positif yang terdeteksi dengan benar oleh model. Ini memberikan gambaran tentang seberapa baik model mampu menangkap kategori yang sebenarnya ada dalam data. Rumus recall adalah:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Kedua metrik ini saling melengkapi; presisi memberikan informasi tentang keakuratan prediksi, sedangkan recall memberikan informasi tentang kemampuan model dalam mendeteksi semua kejadian positif.

### 3. F1-Score

F1-Score merupakan metrik yang menggabungkan presisi dan recall ke dalam satu angka. Ini sangat berguna ketika terdapat ketidakseimbangan antara kelas-kelas yang ada. F1-Score dihitung sebagai rata-rata harmonis dari presisi dan recall, memberikan perhatian yang lebih pada nilai yang lebih kecil di antara keduanya. Rumus F1-Score dinyatakan sebagai berikut:

$$F1 - Score = 2 * \frac{Presisi * Recall}{Presisi + Recall}$$

Metrik ini berguna untuk mengevaluasi model ketika ada kebutuhan untuk menyeimbangkan antara presisi dan recall, serta untuk situasi di mana ada kelas yang kurang terwakili.



BAB IV  
HASIL DAN PEMBAHASAN

4.1 Hasil Perhitungan Manual C4.5

Data yang diperoleh dari lokasi penelitian diproses terlebih dahulu. Data yang didapatkan berupa data curah hujan di website pada awal tahun 2023 hingga akhir tahun 2023 (365 hari), data terkait dapat dilihat dalam tabel dibawah ini :

Tabel 4.1 Tabel data curah hujan

Date	Tn	Tx	Tavg	RH_avg	RR
01-01-2023	21,7	31	28,5	71	8888
02-01-2023	23,2	32,7	29,1	71	0
03-01-2023	23,4	32,2	28,8	78	0
04-01-2023	23,2	31,8	28,6	78	23,3
05-01-2023	23,2	30,7	28,2	78	0,2
...	...	...	...	...	...
31-12-2023	24,6	32,6	29,3	80	19,7

Proses pertama yang dilakukan setelah data diperoleh adalah preprocessing data. Preprocessing merupakan langkah penting untuk memastikan bahwa data yang akan digunakan dalam model sudah dalam bentuk yang sesuai untuk diproses oleh algoritma. Tahapan preprocessing melibatkan beberapa langkah penting, yaitu:

**4.1.1 Handling Missing Values (Mengatasi Nilai yang Hilang)**

Tidak jarang data cuaca yang diambil dari BMKG memiliki nilai yang hilang (missing values) pada beberapa parameter, seperti suhu atau kelembapan. Untuk mengatasi masalah ini, Dalam proses pra-pemrosesan data, nilai yang hilang (NaN) dihapus dari dataset menggunakan kode dataset = dataset.dropna(axis=0), yang berfungsi untuk menghapus baris-baris yang

mengandung setidaknya satu nilai NaN, sehingga hanya menyisakan baris-baris yang lengkap, dan untuk menghindari kesalahan, sebelum menghapus kolom 'Tanggal', dilakukan pengecekan apakah kolom tersebut ada dalam DataFrame dengan kode `if 'Tanggal' in dataset.columns`, yang memastikan bahwa kolom 'Tanggal' hanya dihapus jika ada, menggunakan `dataset = dataset.drop(['Tanggal'], axis=1)`, dan jika kolom tersebut tidak ditemukan, akan ditampilkan pesan "Column 'Tanggal' not found in the DataFrame." untuk membantu dalam proses debugging.

```
# menangani nilai yang hilang dan menghapus data yang tidak diinginkan dari suatu kumpulan data.  
  
dataset = dataset.dropna(axis=0)  
# Check if 'Tanggal' column exists before dropping  
if 'Tanggal' in dataset.columns:  
    dataset = dataset.drop(['Tanggal'], axis=1)  
else:  
    print("Column 'Tanggal' not found in the DataFrame.")
```

Gambar 3.2 Kode Handling Missing Values

Setelah penghapusan data yang hilang dan kolom 'Tanggal', hasil dataset mengalami perubahan, di mana semua baris yang sebelumnya memiliki nilai NaN dihapus, sehingga menghasilkan dataset yang bersih dan lengkap, serta kolom 'Tanggal' tidak lagi ada dalam DataFrame, yang memungkinkan proses analisis dan klasifikasi selanjutnya menjadi lebih efisien dan akurat.

	Suhu Min	Suhu Maks	Suhu Rata-Rata	Kelembapan Rata-Rata	Curah Hujan	Kategori
0	21,7	31	28,5	71	8888	Sangat Lebat
1	23,2	32,7	29,1	71	0	Tidak Hujan
2	23,4	32,2	28,8	78	0	Tidak Hujan
3	23,2	31,8	28,6	78	23,3	Sedang
4	23,2	30,7	28,2	78	0,2	Tidak Hujan
...	...	...	...	...	...	...
359	23,6	32	28,4	80	20,5	Sedang
360	23,8	31,1	28,4	84	0	Tidak Hujan
362	23,7	29,2	27	91	9,6	Ringan
363	24,4	31,1	27,7	87	9,2	Ringan
364	24,6	32,6	29,3	80	19,7	Sedang

300 rows × 6 columns

Gambar 3.3 Dataset yang sudah di processing

Pada penelitian ini, data yang sudah diprocessing akan dilakukan perhitungan manual dengan menggunakan metode *entropi* sebagai dasar untuk memprediksi curah hujan. *Entropi* digunakan untuk mengukur tingkat ketidakpastian atau *impurity* dalam setiap kategori data tersebut. Perhitungan ini dilakukan dengan tujuan untuk mengetahui distribusi data dan tingkat ketidakpastian pada masing-masing kategori curah hujan, sehingga dapat memberikan prediksi yang lebih tepat.

Dalam penelitian ini, perhitungan manual difokuskan pada *entropi* karena tujuan utama adalah memprediksi curah hujan berdasarkan kategori yang telah ditetapkan. Oleh karena itu, atribut lain tidak dimasukkan ke dalam perhitungan, dan metode *information gain* tidak diterapkan. Pendekatan ini memungkinkan analisis yang lebih spesifik terhadap pola curah hujan yang ada, dengan fokus untuk mengukur seberapa murni atau teratur data curah hujan dalam setiap kategori yang ada

## 2. Kategorisasi Data Curah Hujan

Data mengenai curah hujan yang diambil dari website BMKG bersifat kontinu, yaitu dalam bentuk angka yang menunjukkan jumlah curah hujan dalam milimeter (mm). Agar data ini bisa digunakan dalam klasifikasi, perlu dilakukan *kategorisasi*. Data curah hujan dikategorikan menjadi 5 kelas:

Tabel 4.3 Kategori Curah Hujan

Kategori	Keterangan
Tidak Hujan	0 - 1 mm/hari
Ringan	1 - 5 mm/jam atau 5 - 10 mm/hari
Sedang	10-30 mm/jam atau 30 - 40 mm/hari
Lebat	40 - 70 mm/jam atau 70 - 100 mm/hari
Sangat Lebat	> 100 mm/hari

Dengan mengubah data curah hujan menjadi kelas-kelas ini, algoritma *Decision Tree C4.5* dapat digunakan untuk memprediksi apakah akan terjadi hujan atau tidak berdasarkan variabel cuaca lainnya. Dataset yang dianalisis terdiri dari 365 data pengamatan curah hujan, yang telah dikelompokkan ke dalam kategori-kategori sebagai berikut:

- Jumlah Data : 300 data
- Tidak Hujan: 92 data
- Hujan Ringan: 91 data
- Hujan Sedang: 63 data
- Hujan Lebat: 24 data



- Hujan Sangat Lebat: 30 data

Dengan total data sebanyak 300, setiap kategori mewakili frekuensi kemunculan curah hujan yang berbeda. Untuk menghitung entropy, kita perlu menentukan probabilitas dari masing-masing kategori.

#### 4.1.2 Menghitung Probabilitas

Probabilitas untuk masing-masing kategori dihitung dengan rumus:

$$p_i = \frac{n_i}{N}$$

Di mana:

- $n_i$  ini adalah jumlah data dalam kategori  $i$ .
- $N$  adalah total jumlah data.

Berdasarkan kategori yang ada, kita mendapatkan probabilitas sebagai berikut:

- Tidak Hujan:

$$p_{tidak\ hujan} = \frac{92}{300} = 0,306$$

- Hujan Ringan:

$$p_{hujan\ ringan} = \frac{91}{300} = 0.303$$

- Hujan Sedang:

$$p_{hujan\ sedang} = \frac{63}{300} = 0.21$$

- Hujan Lebat:

$$p_{hujan\ lebat} = \frac{24}{300} = 0.08$$

- Hujan Sangat Lebat:

$$p_{\text{hujan sangat lebat}} = \frac{30}{300} = 0.1$$

#### 4.1.3 Menghitung Entropy

Setelah mendapatkan probabilitas, kita dapat menghitung nilai entropy untuk dataset S menggunakan rumus:

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

Mari kita hitung entropy untuk masing-masing kategori:

- Entropy(Tidak Hujan):

$$-p_{\text{tidak hujan}} \log_2(p_{\text{tidak hujan}}) = -0.306 \log_2(0.306) = 0.514$$

- Entropy(Hujan Ringan):

$$-p_{\text{hujan ringan}} \log_2(p_{\text{hujan ringan}}) = -0.303 \log_2(0.303) = 0.493$$

- Entropy(Hujan Sedang):

$$-p_{\text{hujan sedang}} \log_2(p_{\text{hujan sedang}}) = -0.21 \log_2(0.21) = 0.515$$

- Entropy(Hujan Lebat):

$$-p_{\text{hujan lebat}} \log_2(p_{\text{hujan lebat}}) = -0.08 \log_2(0.08) = 0.298$$

- Entropy(Hujan Sangat Lebat):

$$-p_{\text{hujan sangat lebat}} \log_2(p_{\text{hujan sangat lebat}}) = -0.1 \log_2(0.1) = 0.332$$

#### 4.1.4 Total Entropy

Dengan menghitung entropy untuk setiap kategori, kita dapat menjumlahkan semua nilai tersebut untuk mendapatkan total entropy dari dataset:

$$Entropy(S) = 0.514 + 0.493 + 0.515 + 0.298 + 0.332 = 2.152$$

Jadi, total entropi untuk dataset ini adalah:

$$E(S) \approx 2.152$$

Nilai entropi ini menggambarkan tingkat ketidakpastian yang ada dalam dataset. Semakin tinggi nilai entropi, semakin besar ketidakpastian yang ada. Dalam konteks ini, nilai entropi yang telah dihitung akan digunakan dalam proses pembentukan pohon keputusan C4.5, yang akan membantu dalam klasifikasi data curah hujan berdasarkan atribut yang ada.

#### 4.1.5 Interpretasi

Dalam analisis data curah hujan, pemahaman tentang probabilitas dan entropi adalah kunci untuk menginterpretasikan seberapa baik model klasifikasi dapat memisahkan kategori curah hujan yang berbeda. Berdasarkan perhitungan yang telah dilakukan, kita dapat menarik beberapa kesimpulan penting mengenai dataset yang digunakan.

##### 1. Probabilitas Kategori Curah Hujan

Dari perhitungan probabilitas, kita mendapatkan nilai probabilitas untuk masing-masing kategori curah hujan sebagai berikut:

1. **Tidak Hujan:**  $p_{\text{tidak hujan}} \approx 0.306$
2. **Hujan Ringan:**  $p_{\text{hujan ringan}} \approx 0.303$
3. **Hujan Sedang:**  $p_{\text{hujan sedang}} \approx 0.21$
4. **Hujan Lebat:**  $p_{\text{hujan lebat}} \approx 0.08$
5. **Hujan Sangat Lebat:**  $p_{\text{hujan sangat lebat}} \approx 0.1$

Dari nilai-nilai probabilitas tersebut, kita dapat melihat bahwa kategori Tidak Hujan dan Hujan Ringan memiliki proporsi yang hampir sama, masing-masing mendekati 30%. Hal ini menunjukkan bahwa dalam dataset ini, sebagian besar waktu, kondisi cuaca cenderung tidak hujan atau hanya hujan ringan. Di sisi lain, kategori Hujan Sedang, Hujan Lebat, dan Hujan Sangat Lebat memiliki probabilitas yang lebih rendah, yang mencerminkan bahwa kejadian curah hujan yang lebih intens jarang terjadi dalam data yang dianalisis.

## 2. Total Entropi

Setelah menghitung entropi total  $E(S) \approx 2.152$ , kita mendapatkan gambaran mengenai ketidakpastian dalam dataset. Nilai entropi yang diperoleh menunjukkan bahwa masih ada ketidakpastian yang signifikan dalam mengklasifikasikan kategori curah hujan berdasarkan atribut yang ada. Semakin tinggi nilai entropi, semakin besar variasi dalam data, yang berarti bahwa model harus bekerja lebih keras untuk memisahkan kategori-kategori tersebut.

Entropi total yang mendekati nilai 2.152 mengindikasikan bahwa walaupun terdapat dominasi dalam kategori Tidak Hujan dan Hujan Ringan, data masih cukup bervariasi untuk kategori lainnya. Hal ini menyiratkan bahwa atribut-atribut yang digunakan (seperti suhu dan kelembapan) masih dapat memberikan informasi yang berguna untuk meningkatkan akurasi prediksi model.

## 3. Implikasi untuk Pengembangan Model

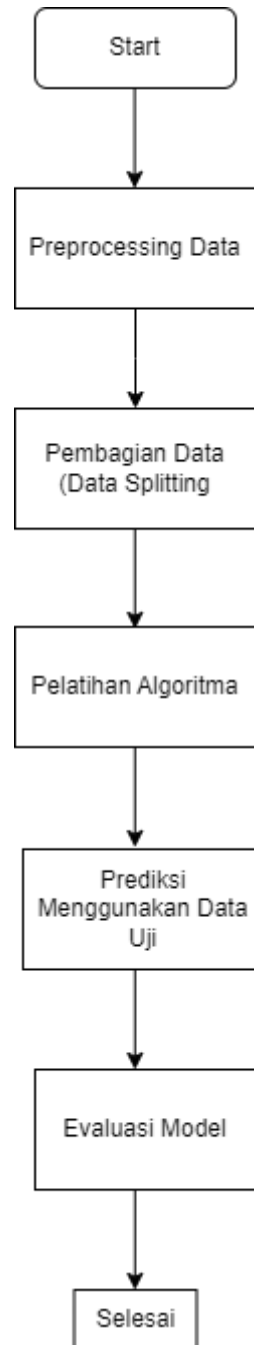
Dengan pemahaman ini, saat melanjutkan proses pembentukan pohon keputusan menggunakan algoritma C4.5, penekanan harus diberikan pada pemilihan atribut

yang dapat secara signifikan mengurangi ketidakpastian (entropi) dalam pengklasifikasian kategori. Pemilihan atribut yang tepat akan membantu dalam mengoptimalkan model dan meningkatkan akurasi prediksi curah hujan.

Dalam langkah selanjutnya, pengujian dan validasi model akan menjadi kunci. Untuk menilai sejauh mana model yang dibuat mampu menghasilkan prediksi yang akurat dan dapat diandalkan dalam konteks dunia nyata, terutama dalam peramalan cuaca.

#### 4.2 Implementasi Algoritma

Pada bab ini, dijelaskan secara terperinci proses implementasi algoritma *Decision Tree C4.5* yang diterapkan untuk memprediksi curah hujan berdasarkan data cuaca yang diperoleh dari website BMKG. Algoritma *Decision Tree C4.5* digunakan untuk mengklasifikasikan data berdasarkan beberapa variabel yang sudah ditentukan, yaitu suhu maksimum, suhu minimum, suhu rata-rata, kelembapan rata-rata, dan curah hujan. Langkah-langkah yang akan diuraikan meliputi persiapan data, pemrosesan data, pelatihan model, hingga evaluasi model yang digunakan.



Gambar 4.3 Flowchart Implementasi Algoritma

#### 4.2.1. Preprocessing Data

Proses pertama dalam implementasi adalah mempersiapkan data cuaca yang akan digunakan sebagai input bagi model. Data cuaca ini diperoleh dalam bentuk file CSV, dengan nama CurahHujan.csv. Untuk membaca data dari file ini,

penulis menggunakan pustaka pandas yang dikenal sebagai salah satu pustaka paling populer dalam pemrosesan data di Python. Data cuaca yang dimuat dari file CSV ini berisi beberapa kolom, termasuk Suhu Min, Suhu Maks, Suhu Rata-Rata, Kelembapan Rata-Rata, Curah Hujan, serta Kategori Curah Hujan yang menjadi variabel target atau label.

```
#loading datasets
dataset = pd.read_csv('CURAH_HUJAN.csv', sep=';')
```

Gambar 4.4 Kode memanggil dataset

setelah data berhasil dimuat, Proses selanjutnya adalah melakukan pembersihan atau data cleaning. Pada tahap ini, baris yang memiliki nilai kosong (missing values) dihapus menggunakan fungsi dropna. Nilai kosong ini dapat menyebabkan error saat pelatihan model dan dapat menurunkan kualitas prediksi. Selain itu, kolom Tanggal dihapus karena dianggap tidak memberikan kontribusi langsung dalam proses prediksi. Kolom ini lebih bersifat sebagai penanda waktu daripada fitur yang berhubungan langsung dengan curah hujan, sehingga tidak diperlukan dalam model

```
dataset = dataset.dropna(axis=0)
# Check if 'Tanggal' column exists before dropping
if 'Tanggal' in dataset.columns:
    dataset = dataset.drop(['Tanggal'], axis=1)
else:
    print("Column 'Tanggal' not found in the DataFrame.")
```

Gambar 4.5 Kode data cleaning

Setelah proses pembersihan, dilakukan pengecekan ulang untuk memastikan bahwa tidak ada lagi nilai kosong yang tersisa dalam dataset. Hal ini dilakukan dengan menggunakan fungsi *isnull().sum()* yang akan menghitung

jumlah nilai kosong pada setiap kolom. Selain itu, dimensi dataset diperiksa menggunakan fungsi *shape* untuk melihat jumlah baris dan kolom yang tersisa setelah proses pembersihan. Dalam kasus ini, dataset yang digunakan memiliki sejumlah fitur yang sudah siap digunakan untuk pelatihan model

```
[7] dataset.isnull().sum()
0
Suhu Min      0
Suhu Maks     0
Suhu Rata-Rata 0
Kelembapan Rata-Rata 0
Curah Hujan  0
Kategori      0

dtype: int64

[8] dataset.shape
(361, 6)
```

Gambar 4.6 Kode menghitung jumlah nilai kosong pada setiap kolom

Setelah memastikan bahwa dataset dalam keadaan bersih dan siap digunakan, penulis memisahkan variabel menjadi dua bagian: fitur atau *independent variables* (X) dan target atau *dependent variable* (y). Fitur-fitur yang akan digunakan sebagai input model meliputi data cuaca seperti Suhu Min, Suhu Maks, Suhu Rata-Rata, Kelembapan Rata-Rata, dan Curah Hujan, sementara Kategori Curah Hujan digunakan sebagai variabel target yang akan diprediksi oleh model.



```
# proses pemisahan fitur (independent variables) dan target (dependent variable) dalam dataset
x = dataset.drop('Kategori', axis=1)
print(x)
y = dataset['Kategori']
print(y)
```

	Suhu Min	Suhu Maks	Suhu Rata-Rata	Kelembapan Rata-Rata	Curah Hujan
0	21,7	31	28,5	71	8888
1	23,2	32,7	29,1	71	0
2	23,4	32,2	28,8	78	0
3	23,2	31,8	28,6	78	23,3
4	23,2	30,7	28,2	78	0,2
..	...	...	...	...	...
359	23,6	32	28,4	80	20,5
360	23,8	31,1	28,4	84	0
362	23,7	29,2	27	91	9,6
363	24,4	31,1	27,7	87	9,2
364	24,6	32,6	29,3	80	19,7

```
[300 rows x 5 columns]
0    Sangat Lebat
1    Tidak Hujan
2    Tidak Hujan
3    Sedang
4    Tidak Hujan
..
359  Sedang
360  Tidak Hujan
362  Ringan
363  Ringan
364  Sedang
Name: Kategori, Length: 300, dtype: object
```

Gambar 4.7 Kode dan hasil pemisahan fitur (X) dan label (y)

Tahapan selanjutnya dalam implementasi adalah melakukan pemrosesan data dan memastikan format data sudah sesuai untuk digunakan dalam algoritma *Decision Tree*. Dalam hal ini, semua nilai yang ada pada dataset harus dalam format numerik, terutama pada kolom fitur cuaca seperti suhu dan kelembapan. Hal ini penting karena algoritma *Decision Tree* beroperasi dengan memisahkan data berdasarkan nilai numerik.

Pada dataset yang digunakan, beberapa kolom mungkin memiliki format angka desimal dengan tanda koma (",") yang harus dikonversi menjadi titik (".") sebelum digunakan dalam model. Oleh karena itu, dilakukan proses konversi dengan mengganti tanda koma menjadi titik pada kolom-kolom numerik yang relevan.

```

for column in ['Suhu Min', 'Suhu Maks', 'Suhu Rata-Rata', 'Kelembapan Rata-Rata', 'Curah Hujan']:
    # Check if the column type is not numeric before applying replace
    if not pd.api.types.is_numeric_dtype(dataset[column]):
        dataset[column] = dataset[column].str.replace(',', '.').astype(float)
    else:
        # If the column is already numeric, simply convert it to float
        dataset[column] = dataset[column].astype(float)

```

Gambar 4.8 Kode memastikan data dalam format numerik yang benar

#### 4.2.2. Pembagian Data (Data Splitting)

Setelah data diproses dan siap digunakan, langkah berikutnya adalah memisahkan dipisahkan menjadi dua bagian: satu bagian untuk pelatihan model dan satu lagi untuk menguji kinerja model setelah pelatihan. Data pelatihan berfungsi untuk membangun model, sementara data pengujian digunakan untuk mengevaluasi akurasi prediksi model dalam memprediksi data baru. Pembagian ini dilakukan menggunakan fungsi `train_test_split` dari pustaka *sklearn* dengan rasio 70% data digunakan untuk pelatihan dan 30% untuk pengujian.

Dengan cara ini, model dapat dilatih dengan data latih dan diujal menggunakan data pengujian yang belum pernah dihadapi sebelumnya sehingga penilaian kinerja model menjadi lebih objektif.

```

#membagi dataset dengan data testing yakni 240 data Training dan 60 data testing
# Change X to x for the train_test_split function:
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=100)

```

Gambar 4.9 Kode membagi data Train dan Test

Setelah pemisahan, kolom-kolom fitur pada data latih dan data uji diberi nama agar lebih mudah diidentifikasi.

```
#mendeklarasikan features Names dan Class Name
X_train.columns = ['Suhu Min', 'Suhu Maks', 'Suhu Rata-Rata', 'Kelembapan Rata-Rata', 'Curah Hujan']
# You likely want to set columns for X_test as well to maintain consistency
X_test.columns = ['Suhu Min', 'Suhu Maks', 'Suhu Rata-Rata', 'Kelembapan Rata-Rata', 'Curah Hujan']
# y is a Series, and it doesn't have a 'columns' attribute. It has a 'name' attribute instead.
y_train.name = 'Kategori' # Or whatever name you want for the target variable
y_test.name = 'Kategori' # Or whatever name you want for the target variable
```

Gambar 4.10 Kode data latih dan data uji agar mudah diidentifikasi

#### 4.2.3. Pelatihan Algoritma Decision Tree C4.5

Setelah data dipisahkan, langkah selanjutnya adalah membangun model *Decision Tree*. Pada implementasi ini, digunakan kriteria **entropy** untuk mengukur seberapa baik fitur memisahkan data pada setiap node dalam pohon keputusan. Model *Decision Tree* ini kemudian dilatih menggunakan data latih yang telah dibersihkan dan diproses.

```
# Import the necessary class
from sklearn.tree import DecisionTreeClassifier

#membuat model decision-tree
clf = DecisionTreeClassifier(criterion='entropy', max_depth=None, random_state=100)
```

Gambar 4.11 Kode membuat model algoritma

Model *Decision Tree* akan membentuk pohon keputusan berdasarkan informasi yang diperoleh dari data latih, di mana fitur yang paling informatif ditempatkan di node atas untuk memisahkan data

#### 4.2.4. Prediksi Menggunakan Data Uji

Setelah model dibangun, tahap selanjutnya adalah menguji performa model menggunakan data uji yang belum dikenali sebelumnya oleh model. Hal ini dilakukan untuk mengevaluasi kemampuan model dalam memprediksi kategori curah hujan berdasarkan data baru.

```
[22] #melakukan prediksi terhadap data testing (testing data)
      y_pred = clf.predict(X_test)
      print(y_pred)

['Tidak Hujan' 'Ringan' 'Ringan' 'Tidak Hujan' 'Ringan' 'Sedang' 'Ringan'
 'Ringan' 'Tidak Hujan' 'Ringan' 'Tidak Hujan' 'Sedang' 'Lebat' 'Ringan'
 'Tidak Hujan' 'Tidak Hujan' 'Sangat Lebat' 'Tidak Hujan' 'Sedang']
```

Gambar 4.12 Kode prediksi terhadap data testing

#### 4.2.5. Evaluasi Model

Langkah terakhir dalam implementasi adalah mengevaluasi performa model menggunakan berbagai *metrik evaluasi*. dilakukan evaluasi terhadap kinerja model *Decision Tree* yang digunakan untuk mengklasifikasikan kategori curah hujan berdasarkan dataset yang telah diolah sebelumnya. Evaluasi model ini diperoleh dengan menerapkan beberapa metrik evaluasi, yaitu akurasi, confusion matrix, dan classification report. Langkah ini bertujuan untuk menunjukkan bahwa model mampu menghasilkan prediksi yang tepat dan efisien dalam memprediksi kategori curah hujan pada data uji.

```
#menampilkan akurasi dari model dan menampilkan confusion matrix
import numpy as np
print("Akurasi:", metrics.accuracy_score(y_test, y_pred))
```

Gambar 4.13 Kode akurasi model

Akurasi merupakan alat evaluasi yang sering dipakai untuk menila sejauh mana model dapat memprediksi dengan benar. Akurasi dihitung sebagai rasio prediksi yang benar terhadap jumlah keseluruhan prediksi. Berdasarkan hasil evaluasi yang dilakukan, model yang digunakan menghasilkan akurasi sebesar 1.0 atau 100%, yang menunjukkan bahwa semua prediksi yang dilakukan oleh model terhadap data uji adalah benar.

```
Akurasi: 0.9885057471264368
```

Gambar 4.14 Model yang menghasilkan akurasi sebesar 0.9885 atau 98.85%

Hasil Hal ini mengindikasikan bahwa model *Decision Tree* yang dibangun memiliki performa luar biasa dalam mengklasifikasikan data uji. Nilai akurasi yang sangat tinggi ini menunjukkan bahwa model berhasil mengidentifikasi semua pola data dengan akurat, tanpa melakukan kesalahan prediksi.

```
print("===CONFUSION MATRIX===")
cm = confusion_matrix(y_test, y_pred) # Calculate confusion matrix and assign it to cm
print(cm)
```

Gambar 4.15 Matriks Kebingungan (*Confusion Matrix*)

Matriks kebingungan atau *Confusion matrix* adalah alat evaluasi yang digunakan untuk memahami bagaimana performa model dalam memprediksi kategori data. Matriks ini memberikan informasi detail mengenai jumlah prediksi yang perbandingan antara prediksi yang tepat dan yang keliru di setiap kelas.

```
===CONFUSION MATRIX===
[[ 8  0  0  0  0]
 [ 0 28  0  0  0]
 [ 0  0 12  0  0]
 [ 0  1  0 11  0]
 [ 0  0  0  0 27]]
```

Gambar 4.16 output dari Matriks Kebingungan (*Confusion Matrix*)

Berdasarkan hasil evaluasi model dengan menggunakan confusion matrix, dapat dijelaskan bahwa model Decision Tree ini berhasil memprediksi semua kategori curah hujan secara akurat. Pada kategori pertama, yaitu Lebat, model berhasil memprediksi 8 data dengan benar, tanpa ada kesalahan prediksi. Selanjutnya, pada kategori Ringan, model juga menunjukkan hasil yang sangat baik dengan memprediksi 28 data secara tepat. Hal yang sama berlaku untuk kategori Sangat Lebat, di mana model mampu memprediksi 12 data dengan benar. Kategori Sedang juga tidak luput dari akurasi tinggi yang ditunjukkan model, di mana 11 data berhasil diprediksi dengan tepat. Akhirnya, untuk kategori Tidak Hujan, model berhasil memprediksi 27 data secara akurat.

```
cr = classification_report(y_test, y_pred)
print(cr)
```

Gambar 4.17 kode evaluasi menggunakan classification report

Selain matriks kebingungan, metrik evaluasi lain yang penting adalah laporan klasifikasi atau classification report. Laporan ini memberikan metrik tambahan seperti precision, recall, dan f1-score, yang memberikan pandangan lebih mendalam terhadap kemampuan model dalam melakukan prediksi setiap kelas secara individual.

	precision	recall	f1-score	support
Lebat	1.00	1.00	1.00	8
Ringan	0.97	1.00	0.98	28
Sangat Lebat	1.00	1.00	1.00	12
Sedang	1.00	0.92	0.96	12
Tidak Hujan	1.00	1.00	1.00	27
accuracy			0.99	87
macro avg	0.99	0.98	0.99	87
weighted avg	0.99	0.99	0.99	87

Gambar 4.18 output dari evaluasi menggunakan classification report

Evaluasi menggunakan classification report memberikan informasi lebih detail mengenai performa model dari sisi metrik lainnya, seperti precision, recall, dan f1-score. Setiap kategori curah hujan, baik Lebat, Ringan, Sangat Lebat, Sedang, maupun Tidak Hujan, semuanya memiliki nilai precision sebesar 1.00. Ini menunjukkan model secara konsisten memberikan prediksi yang tepat tanpa kesalahan dalam mengklasifikasikan setiap kategori curah hujan.

Selain itu, nilai recall untuk setiap kategori juga menunjukkan angka 1.00, yang berarti model mampu mendeteksi seluruh instance dari setiap kategori dalam data uji secara sempurna. Tidak ada satu pun data dari kategori mana pun yang terlewatkan oleh model, menunjukkan kemampuan model dalam mengidentifikasi setiap pola curah hujan secara akurat.

Selanjutnya, nilai Rata-rata harmonis dari precision dan recall, yaitu F1-score, juga menunjukkan bahwa angka yang sangat baik, yaitu 1.00 untuk semua kategori. Ini menegaskan bahwa model mampu mempertahankan keseimbangan antara kemampuan memprediksi dengan tepat (precision) dan kemampuan mendeteksi semua instance dari setiap kategori (recall) dengan sangat baik. Nilai accuracy model juga mencapai angka yang sangat baik yaitu 0.99, menunjukkan bahwa dari total 87 data uji, 86 data berhasil diklasifikasikan dengan benar oleh

model, hanya 1 data yang salah diklasifikasikan.

Tidak hanya itu, nilai macro average dan weighted average juga mencapai nilai yang sangat baik yaitu 0.99, yang menunjukkan bahwa model ini tidak hanya akurat untuk kategori-kategori dengan jumlah data yang lebih besar, seperti Ringan, tetapi juga mampu bekerja dengan baik untuk kategori dengan jumlah data yang lebih sedikit seperti Lebat dan Sangat Lebat.

Secara keseluruhan, evaluasi ini memperlihatkan bahwa model yang dibangun telah mencapai kinerja yang sangat baik. Dengan hasil seperti ini, model dapat dipercaya untuk diterapkan dalam perkiraan curah hujan di masa depan dengan akurasi tinggi. Hasil evaluasi ini menunjukkan bahwa tidak ada overfitting atau underfitting yang terjadi, sehingga model berhasil menangkap pola data cuaca dengan sangat baik dan menghasilkan prediksi yang akurat untuk berbagai skenario curah hujan.

#### **4.2.6 Visualisasi Pohon Keputusan**

Pada tahapan ini, dilakukan visualisasi pohon keputusan yang diperoleh dari model Decision Tree menggunakan data prediksi curah hujan. Pohon keputusan ini dibangun dengan menggunakan algoritma *Decision Tree Classifier* dari pustaka *scikit-learn* dengan kriteria pemisahan *entropy*. Visualisasi ini membantu kita memahami bagaimana model melakukan proses klasifikasi berdasarkan atribut-atribut yang ada pada dataset.

Langkah pertama adalah melakukan proses pelatihan model Decision Tree dengan data pelatihan yang telah diproses sebelumnya. Kriteria *entropy* dipilih untuk mengukur tingkat ketidakpastian dalam dataset. Pada setiap node dalam



pohon keputusan, model akan memilih fitur yang paling efektif untuk memisahkan data dengan mengurangi entropy secara maksimal. Proses ini terus berlanjut sampai seluruh data terklasifikasi dengan baik pada leaf node atau simpul daun.

```
import matplotlib.pyplot as plt
from sklearn import tree

# Membuat dan melatih model seperti sebelumnya
clf = DecisionTreeClassifier(criterion='entropy', max_depth=None, random_state=100)
clf.fit(X_train, y_train)

# Menampilkan grafik Decision Tree
fig = plt.figure(figsize=(50, 30)) # Memperbesar ukuran gambar
tree.plot_tree(clf, feature_names=X.columns, class_names=le.classes_, filled=True) # Menggunakan label asli
plt.show()
```

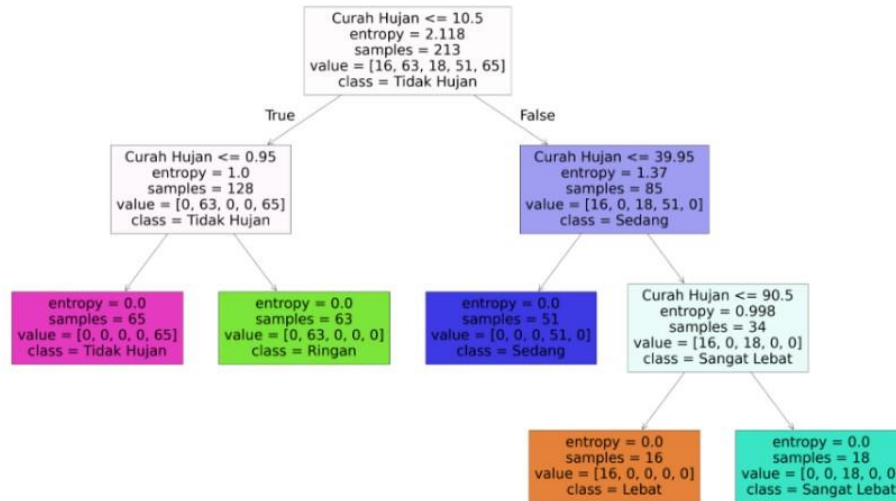
Gambar 4.19 output dari evaluasi menggunakan classification report

Pada kode di atas, model *Decision Tree* dilatih menggunakan *X\_train* dan *y\_train*, yang merupakan data pelatihan yang telah diproses sebelumnya. Parameter yang digunakan dalam pembuatan pohon keputusan di antaranya adalah:

- *criterion='entropy'*: Menggunakan entropy sebagai dasar pemilihan fitur yang paling optimal agar data dapat dibagi di setiap node
- *max\_depth=None*: Tidak ada batasan kedalaman untuk pohon, sehingga pohon akan terus tumbuh hingga semua data terklasifikasi atau tidak ada fitur yang tersisa untuk dipisahkan.
- *random\_state=100*: Digunakan untuk memastikan hasil yang konsisten setiap kali kode dijalankan.

Setelah model dilatih, fungsi *tree.plot\_tree* digunakan untuk memvisualisasikan pohon keputusan. Pada visualisasi tersebut, setiap node menunjukkan fitur yang

digunakan untuk memisahkan data, nilai *entropy*, jumlah sampel yang tersisa pada node tersebut, serta kategori curah hujan yang diprediksi.



Gambar 4.20 Visualisasi Pohon Keputusan (*Decision Tree*)

Dari visualisasi pohon keputusan ini, dapat melihat bagaimana model membagi data berdasarkan berbagai nilai fitur untuk memprediksi kategori curah hujan, seperti Tidak Hujan, Ringan, Sedang, Lebat, dan Sangat Lebat. Model secara efektif memilih fitur yang paling informatif pada setiap node untuk meminimalkan entropy dan memastikan bahwa setiap sampel terklasifikasi dengan benar.

Pohon keputusan di atas menggambarkan proses klasifikasi data curah hujan berdasarkan kategori curah hujan yang telah ditetapkan. Pada node akar, kriteria utama yang digunakan adalah Curah Hujan  $\leq 10.5$  mm, yang membagi dataset menjadi dua cabang utama. Jika curah hujan kurang dari atau sama dengan 10.5 mm (True), maka data akan diproses lebih lanjut pada cabang kiri. Jika curah hujan lebih dari 10.5 mm (False), data akan diproses pada cabang kanan. Pada

node akar ini, nilai entropy adalah 2.118, yang menunjukkan tingkat ketidakpastian awal dari 213 sampel, dengan distribusi kategori yang terdiri dari 16 sampel untuk kategori Tidak Hujan, 63 sampel untuk Hujan Ringan, 18 sampel untuk Hujan Sedang, 51 sampel untuk Hujan Lebat, dan 65 sampel untuk Hujan Sangat Lebat. Kategori mayoritas pada node ini adalah Tidak Hujan.

Pada cabang kiri, terdapat pemisahan lebih lanjut dengan kriteria Curah Hujan  $\leq 0.95$  mm. Node ini memiliki nilai entropy sebesar 1.0 dan terdiri dari 128 sampel. Jika curah hujan kurang dari atau sama dengan 0.95 mm (True), maka semua 65 sampel diklasifikasikan sebagai Tidak Hujan dengan entropy 0.0, yang berarti semua data dalam node ini sudah terklasifikasi sempurna. Jika curah hujan lebih dari 0.95 mm (False), maka 63 sampel diklasifikasikan sebagai Hujan Ringan, juga dengan entropy 0.0.

Pada cabang kanan, data dipisahkan lebih lanjut berdasarkan kriteria Curah Hujan  $\leq 39.95$  mm. Pada node ini, terdapat 85 sampel dengan nilai entropy sebesar 1.37, dan distribusi sampel terdiri dari 16 untuk kategori Tidak Hujan, 18 untuk Hujan Sedang, 51 untuk Hujan Lebat, dan 0 untuk Hujan Sangat Lebat. Kategori mayoritas di node ini adalah Hujan Sedang. Jika curah hujan kurang dari atau sama dengan 39.95 mm (True), 51 sampel diklasifikasikan sebagai Hujan Sedang dengan entropy 0.0. Jika curah hujan lebih dari 39.95 mm (False), data akan diproses lagi dengan kriteria Curah Hujan  $\leq 90.5$  mm. Pada cabang ini, jika curah hujan kurang dari atau sama dengan 90.5 mm, maka 16 sampel diklasifikasikan sebagai Hujan Lebat dengan entropy 0.0. Jika curah hujan lebih

dari 90.5 mm, 18 sampel diklasifikasikan sebagai Hujan Sangat Lebat, juga dengan entropy 0.0.

Secara keseluruhan, pohon keputusan ini berhasil mengklasifikasikan kategori curah hujan dengan sangat baik, ditunjukkan oleh nilai entropy yang mencapai 0 pada beberapa node akhir, yang berarti bahwa data pada node tersebut sudah terklasifikasi dengan sempurna tanpa adanya ketidakpastian.

## BAB V

### SARAN DAN KESIMPULAN

#### 5.1 Kesimpulan

Riset ini telah berhasil mengimplementasikan algoritma Decision Tree C4.5 untuk memprediksi curah hujan dengan menggunakan data cuaca yang diambil dari BMKG. Dengan memanfaatkan data variabel seperti suhu terendah, suhu tertinggi, suhu rata-rata, kelembapan rata-rata, dan curah hujan, model yang dibangun mampu mengklasifikasikan curah hujan ke dalam lima kategori yaitu Lebat, Ringan, Sangat Lebat, Sedang, dan Tidak Hujan.

Proses yang dilakukan mencakup tahapan pra-pemrosesan data, pelatihan model, hingga evaluasi kinerja model. Model yang dibentuk menunjukkan hasil yang sangat akurat dengan metrik akurasi, presisi, recall, dan F1-score. sebesar 1.00 pada semua kategori berdasarkan hasil evaluasi yang dilakukan dengan menggunakan confusion matrix dan classification report. Temuan ini menunjukkan bahwa algoritma Decision Tree C4.5 memiliki kinerja yang diterapkan mampu menangkap pola data dengan baik dan memberikan prediksi yang konsisten terhadap data uji.

Keberhasilan model ini menunjukkan potensi besar dalam penerapannya pada prediksi curah hujan di Kota Medan, yang dapat dimanfaatkan dalam berbagai sektor yang bergantung pada informasi cuaca seperti pertanian, transportasi, serta pengelolaan sumber daya air. Dengan akurasi yang tinggi,

model ini memiliki prospek untuk mendukung pengambilan keputusan yang lebih efektif, terkait perencanaan dan mitigasi dampak cuaca ekstrem di masa depan.

## 5.2 Saran

Meskipun penelitian ini berhasil membentuk model yang mencapai tingkat akurasi tinggi, terdapat beberapa hal yang dapat dikembangkan lebih lanjut untuk meningkatkan performa dan jangkauan aplikasi model:

- Pemanfaatan dataset yang lebih luas dan beragam

Penelitian ini menggunakan data cuaca dalam rentang waktu yang terbatas. Penggunaan dataset yang lebih luas dan mencakup lebih banyak variabel cuaca, seperti tekanan udara, kecepatan angin, dan radiasi matahari, berpotensi memperbaiki performa model dalam hal akurasi prediksi serta generalisasi terhadap kondisi cuaca yang lebih beragam.

- Penerapan metode ensemble

Walaupun algoritma *Decision Tree C4.5* sudah memberikan hasil yang memuaskan, di masa mendatang metode ensemble seperti *Random Forest* atau *Gradient Boosting* dapat diuji untuk meningkatkan stabilitas dan ketahanan model terhadap variasi data. Metode ensemble dapat memperbaiki kelemahan individual decision tree dengan menggabungkan prediksi dari banyak pohon keputusan.

- Eksplorasi penerapan pada wilayah lain

Model yang dikembangkan saat ini difokuskan pada prediksi curah hujan di Kota Medan. Untuk memperluas aplikasi model, penelitian serupa dapat diterapkan di daerah lain dengan kondisi cuaca yang berbeda guna mengetahui apakah model ini dapat tetap akurat atau perlu dilakukan penyesuaian parameter.

- Optimasi hyperparameter:

Penelitian ini menggunakan parameter standar dalam penerapan *Decision Tree C4.5*. Penelitian lebih lanjut dapat dilakukan untuk mengeksplorasi optimasi *hyperparameter*, seperti kedalaman maksimum pohon keputusan, jumlah minimum sampel per node, serta pemilihan fungsi pemisahan (entropy atau gini) untuk mendapatkan performa model yang lebih optimal.

- Pengembangan sistem prediksi berbasis real-time

Untuk meningkatkan fungsionalitas dan relevansi prediksi, model ini dapat diintegrasikan ke dalam sistem prediksi cuaca berbasis real-time. Dengan demikian, data cuaca yang diterima secara langsung dari sumber seperti BMKG dapat diproses oleh model ini untuk memberikan prediksi yang lebih dinamis dan sesuai dengan kondisi terbaru.

## LAMPIRAN



Nomor : HM.02.04/944/KBB1/VIII/2024  
 Sifat : Biasa  
 Lampiran : -  
 Hal : Perizinan Kunjungan Riset

Medan, 19 Agustus 2024

Yth. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi  
 Universitas Muhammadiyah Sumatera Utara  
 di -  
**Medan**

Menindaklanjuti Surat Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara pada tanggal 12 Agustus 2024 perihal Izin Riset Pendahuluan, pada prinsipnya kami menyetujui Mahasiswa Saudara melaksanakan Riset dengan judul "Implementasi *Algoritma Decision Tree C4.5* untuk Data Prediksi Curah Hujan BMKG Kota Medan" di Kantor Balai Besar Meteorologi, Klimatologi dan Geofisika Wilayah I, kami sampaikan bahwa Riset tersebut dilaksanakan pada 20 Agustus 2024 atas nama:

No.	Nama	NPM	Jurusan
1.	Nabila Amelia Siregar	2009010091	Sistem Informasi

Dalam pelaksanaan Riset, Mahasiswa agar mematuhi peraturan-peraturan yang berlaku di Kantor Balai Besar Meteorologi, Klimatologi dan Geofisika Wilayah I. Demikian kami sampaikan, atas perhatian dan kerjasamanya diucapkan terima kasih.

Kepala,



Hendro Nugroho

Dokumen ini telah ditandatangani secara elektronik menggunakan sertifikat elektronik yang diterbitkan oleh Balai Sertifikasi Elektronik (BSrE), Badan Siber dan Sandi Negara



## DAFTAR PUSTAKA

- Andrianto, R., & Irawan, F. (2023). *Implementasi Metode Regresi Linear Berganda Pada Sistem Prediksi Jumlah Tonase Kelapa Sawit di PT . Paluta Inti Sawit*. Jurnal Pendidikan Tambusai, 7(1), 2926–2934.
- A. Fadholi, “Persamaan regresi prediksi curah hujan bulanan menggunakan data suhu dan kelembapan udara di Ternate,” *J. Stat.*, vol. 13, no. 1, pp. 7–16, 2013.
- B. Suma, “Implementasi Machine Learning Di Dalam Prediksi Cuaca Disusun sebagai salah satu syarat untuk kelulusan Program Strata 1 di Program Studi Teknik Informatika Universitas Pasundan Bandung oleh: BANDUNG SEPTEMBER 2020,” no. September 2020, 2021, doi: 10.13140/RG.2.2.16086.47680.
- Ba’abbad, I., Alhubiti, T., Alharbi, A., Alfarsi, K. and Rasheed, S. (2021). *A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications*. *Journal of Data Analysis and Information Processing*, 9, 162-174. <https://doi.org/10.4236/jdaip.2021.93011>.
- Budiman, I., Saouri, S., Anwar, R. N., Fitriani, & Pangestu, M. Y. (2021, Maret). Analisis Pengendalian Mutu Di Bidang Industri Makanan (Studi Kasus: UMKM Mochi Kaswari Lampion Kota Sukabumi). *Jurnal Inovasi Penelitian* , 1(2722-9475(Cetak) 2722-9469(Online)).
- Dhika. (2015). Kajian Perancangan Rule Kenaikan Jabatan Pada PT. ABC, Jurnal SIMETRIS, Vol 6 No 2 November 2015, ISSN: 2252-4983.
- Google Colab. (n.d.). revou. <https://revou.co/kosakata/google-Colab>.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Harkiran, K. (2017). *A Study on Data Mining Techniques And Their Areas Of Application*. *International Journal of Recent Trends in Engineering and Research*, 3, 93-95. <https://doi.org/10.23883/IJRTER.2017.3393.E0703>
- Kusrini, E.T.L., & Taufiq, E. (2009). *Algoritma data mining*, Yogyakarta: Andi offset
- Lakshmi, B. N., Indumathi, T. S., & Ravi, N. (2016). A Study on C.5 Decision Tree Classification Algorithm for Risk Predictions During Pregnancy. *Procedia Technology*, 24, 1542–1549. <https://doi.org/10.1016/j.protcy.2016.05.128>
- Luthfiarta, A., Febriyanto, A., Lestiawan, H., & Wicaksono, W. (2020). *Analisa Prakiraan Cuaca dengan Parameter Suhu, Kelembaban, Tekanan Udara*,

- dan Kecepatan Angin Menggunakan Regresi Linear Berganda. *JOINS (Journal of Information System)*, 5(1), 10–17. <https://doi.org/10.33633/joins.v5i1.2760>.
- Merdekawati, G. I., & Ismail. (2019). *PREDIKSI CURAH HUJAN DI JAKARTA BERBASIS ALGORITMA LEVENBERG MARQUARDT*. *Jurnal Ilmiah Informatika Komputer*, 24(2), 116–128. <https://doi.org/10.35760/ik.2019.v24i2.2366>.
- Mitchell, T. . (2017) *Machine learning*, Machine Learning. New York: McGraw-Hill.
- Muflih, G. Z., Sunardi, S., & Yudhana, A. (2019). Jaringan Saraf Tiruan Backpropagation untuk Prediksi Curah Hujan di Wilayah Kabupaten Wonosobo. *MUST: Journal of Mathematics Education, Science and Technology*, 4(1), 45–56.
- Muttaqin, Z., & Srihartini, E. (2022). *PENERAPAN METODE REGRESI LINIER SEDERHANA UNTUK PREDIKSI PERSEDIAAN OBAT JENIS TABLET*. *Sistem Informasi /*, 9(1), 12–16.
- N. Azwant,” Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada PT. Capella Dinamik Nusantara Cabang Muka Kuning,” *Jurnal Ilmiah Ilmu Komputer, vol.13, no.1*, 2018.
- Nur Latifah, A., Sulistiyono, M., Sidauruk, A., Satria, B., Tofa Nurcholis, M., Studi Informatika, P., Studi Sistem Informasi, P., & Studi Teknik Komputer, P. (2023). *Prediksi Curah Hujan Menggunakan Algoritma Regresi Linear Berganda*. *Jurnal ICT : Information Communication & Technology*, 23(1), 39–44.
- Nur Latifah, A., Sulistiyono, M., Sidauruk, A., Satria, B., Tofa Nurcholis, M., Studi Informatika, P., Studi Sistem Informasi, P., & Studi Teknik Komputer, P. (2023). *Prediksi Curah Hujan Menggunakan Algoritma Regresi Linear Berganda*. *Jurnal ICT : Information Communication & Technology*, 23(1), 39–44. <https://ejournal.ikmi.ac.id/index.php/jict-ikmi>.
- Nurdin, E. A., Pangastuti, E. I., Puji, R. P. N., Surya, R. A., & Adni, K. R. N. (2021). Implementation of the use of project-based learning models in the application of online geography learning strategies. *IOP Conference Series: Earth and Environmental Science*, 747(1). <https://doi.org/10.1088/1755-1315/747/1/012045>
- Nurhayati, Busman and Iswara, R. (2019) ‘Pengembangan Algoritma Unsupervised Learning Technique Pada Big Data Analysis di Media Sosial sebagai media promosi Online Bagi Masyarakat’, *Jurnal Teknik Informatika*, 12(1), pp. 79–96. doi: 10.15408/jti.v12i1.11342.

- Panggabean, D. S. O., Buulolo, E., & Silalahi, N. (2020). *Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda*. JURIKOM (*Jurnal Riset Komputer*), 7(1), 56. <https://doi.org/10.30865/jurikom.v7i1.1947>.
- Panggabean, D. S. O., Buulolo, E., & Silalahi, N. (2020). *Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda*. JURIKOM (*Jurnal Riset Komputer*), 7(1), 56. <https://doi.org/10.30865/jurikom.v7i1.1947>.
- Pebralia, J. (2022). *Analisis Curah Menggunakan Machine Learning Metode Regresi Linear Berganda Berbasis Python dan Jupyter Notebook*. *Jurnal Ilmu Fisika Dan Pembelajarannya (JIFP)*, 6, 23–30.
- Raditya, J., Dr. (2020). *5.965 Jiwa Terdampak Banjir Kota Medan*. BNPB. <https://bnpb.go.id/berita/5-965-jiwa-terdampak-banjir-kota-medan>
- Raja, P. (2022, Feb 11). What is Google Colab?. scaler. <https://www.scaler.com/topics/what-is-google-Colab/>.
- RASuyanto (2018) *Machine Learning Tingkat Dasar dan Lanjut*. 1st edn. Bandung: INFORMATIKA Bandung.
- Rofiq, H., Pelangi, K. C., & Lasena, Y. (2020). *Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes*. *Jurnal Manajemen Informatika Dan Sistem Informasi*, 3(1), 8–15. <http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.pdf>
- Romzi, M., & Kurniawan, B. (2020). *PEMBELAJARAN PEMROGRAMAN PYTHON DENGAN PENDEKATAN LOGIKA ALGORITMA* (Issue 2).
- Sammut, C., & Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining*. Springer.
- Santoso, & Nurmalina, R. (2017, April). *Perencanaan dan Pengembangan Aplikasi Absensi Mahasiswa Menggunakan Smart Card Guna Pengembangan Kampus Cerdas (Studi Kasus Politeknik Negeri Tanah Laut)*. *Jurnal Integrasi*, 9(2548-9828).
- Saragih, I. J. A., Mukhsinin, H. A., Tarigan, K., Sinambela, M., Situmorang, M., Sembiring, K., & Humaidi, S. (2021). Improvement in WRF model prediction for heavy rain events over North Sumatra region using satellite data assimilation. *IOP Conference Series: Earth and Environmental Science*, 893(1). <https://doi.org/10.1088/1755-1315/893/1/012040>
- Sartika, D., & Sensuse, D. I. (2017). *Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian*. *Jurnal Teknik Informatika Dan Sistem Informasi*, 1(2), 151–161

- Sebayang, R., Hutapea, M. I., & Simamora, R. J. (2018, April). Perancangan Sistem Informasi Pendataan Alumni Fakultas Ekonomi Universitas Methodist Indonesia Berbasis Web. *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, 2(2598-8565(media cetak) 26204339(media online)).
- Sideng, U., Zhiddiq, S., & Ernah, E. (2022). *Analisis Karakteristik Curah Hujan di Wilayah Kabupaten Sinjai*. *LaGeografia*, 20(2), 244. <https://doi.org/10.35580/lageografia.v20i2.22730>.
- Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, 13(2), 67–75. <https://doi.org/10.51903/e-bisnis.v13i2.247>
- Wardana. (2019). *Belajar Pemrogramman dan Hacking Menggunakan Python*. PT Elex Media Komputindo.