BAB I
PENDAHULUAN

I.1 Latar Belakang

Energi merupakan salah satu elemen vital dalam mendukung pembangunan ekonomi dan infrastruktur suatu negara. Di tengah tantangan perubahan iklim dan kebutuhan untuk mengurangi emisi karbon, energi terbarukan muncul sebagai solusi strategis untuk memastikan keberlanjutan sumber daya energi di masa depan. Menurut Arafah dkk. (2022), perkembangan ekonomi suatu negara sangat bergantung pada ketersediaan energi, terutama dalam mendukung prioritas pembangunan pemerintah, seperti pembangunan infrastruktur di Indonesia. Salah satu upaya untuk meningkatkan keamanan energi nasional jangka panjang adalah dengan mengurangi ketergantungan pada energi fosil dan beralih ke energi terbarukan. Hal ini sejalan dengan penelitian Kurniawan dkk. (2022) yang menunjukkan bahwa potensi energi terbarukan di Indonesia mencapai lebih dari 400.000 Megawatt (MW). Namun, implementasi energi terbarukan, terutama energi surya, masih menghadapi berbagai tantangan, seperti biaya investasi yang tinggi dan keterbatasan regulasi yang mendukung. Di sisi lain, Lau dkk. (2017) menyoroti pentingnya optimalisasi potensi energi surya di lingkungan tropis, dengan menyebutkan bahwa radiasi matahari yang melimpah di kawasan ini dapat menjadi solusi untuk memenuhi kebutuhan energi di wilayah urban yang terus berkembang. Oleh karena itu, integrasi kebijakan, teknologi, dan strategi pemasaran yang tepat sangat diperlukan untuk mendukung transisi energi di Indonesia (Arafah dkk., 2022; Kurniawan dkk., 2022; Lau dkk., 2017).

Tutupan awan memiliki pengaruh signifikan terhadap radiasi matahari yang mencapai permukaan bumi. Penelitian menunjukkan bahwa adanya awan dapat menyebabkan penurunan intensitas radiasi matahari, tergantung pada jenis dan ketebalan awan tersebut. Fister dkk. (2003) menemukan bahwa tutupan awan yang rendah sering kali menghasilkan peningkatan moderat pada radiasi matahari karena matahari tetap terlihat, sementara awan yang lebih tebal cenderung menghalangi radiasi secara signifikan. Selain itu, variasi tutupan awan juga berkorelasi negatif dengan intensitas radiasi rata-rata bulanan, di mana peningkatan tutupan awan mengurangi nilai radiasi yang diterima. Di sisi lain, penelitian Udelhofen dan Cess (2001) mengamati adanya keterkaitan antara variabilitas tutupan awan dan siklus matahari, menunjukkan bahwa variasi sinar matahari dapat memengaruhi pola atmosfer dan, secara tidak langsung, tutupan awan. Dengan demikian, pemahaman tentang dinamika tutupan awan sangat penting untuk memperkirakan potensi energi surya yang dapat dimanfaatkan di wilayah tertentu.

Penerapan *machine learning* dalam prediksi tutupan awan dan variabel meteorologi lainnya menunjukkan potensi besar untuk meningkatkan akurasi peramalan dan efisiensi berbagai sektor. Baran dkk. (2020) menyelidiki penggunaan

metode seperti *multilayer perceptron* (MLP), *gradient boosting* machines (GBM), dan random forest (RF) untuk post-processing prediksi tutupan awan (TCC) yang sering kali tidak terkalibrasi dengan baik. Hasilnya menunjukkan bahwa semua metode ini memberikan peningkatan yang signifikan dalam keterampilan peramalan, terutama ketika data curah hujan dimasukkan sebagai faktor prediktor tambahan. Hal ini sejalan dengan penelitian Pavuluri dkk. (2020), yang mengaplikasikan algoritma seperti *Decision Tree*, k-nearest neighbors (K-NN), dan random forest untuk memprediksi perubahan suhu dan cuaca menggunakan data historis. Penelitian ini menunjukkan bahwa penggunaan algoritma tersebut dapat memprediksi kondisi cuaca seperti perubahan suhu dan curah hujan dengan akurasi yang baik. Sementara itu, Bochenek dan Ustrnul (2020) dalam ulasan mereka mengenai penerapan *machine learning* dalam penelitian cuaca dan iklim menyoroti pentingnya metode seperti Deep Learning, Random Forest, dan XGBoost dalam memodelkan berbagai variabel meteorologi seperti suhu, tekanan udara, dan radiasi. Ketiga penelitian ini menggarisbawahi bahwa *machine learning* memiliki potensi besar dalam meningkatkan akurasi peramalan cuaca dan tutupan awan, yang pada gilirannya dapat memberikan manfaat besar dalam aplikasi praktis, seperti perencanaan energi terbarukan dan mitigasi dampak perubahan iklim.

Studi oleh Flowers dkk. (2016) menunjukkan bahwa iklim lokal secara langsung mempengaruhi kinerja dan umur panel surya, yang pada akhirnya berdampak pada perhitungan *Levelized Cost of Energy* (LCOE). Di Indonesia, Paradongan dkk. (2020) menemukan bahwa pemberian insentif emisi dapat meningkatkan kelayakan finansial proyek energi surya bagi produsen listrik independen, sementara di sisi lain, PLN memperoleh manfaat dari penghematan biaya operasional bahan bakar fosil, terutama diesel. Dalam konteks ini, prediksi tutupan awan menjadi sangat krusial, karena tingkat radiasi surya yang diterima oleh panel sangat bergantung pada kondisi langit. Penggunaan machine learning dalam memodelkan dan memprediksi tutupan awan menawarkan pendekatan cerdas untuk meningkatkan keakuratan estimasi radiasi surya, yang secara langsung mendukung efisiensi perencanaan dan operasional pembangkit listrik tenaga surya (PLTS). Studi oleh Siregar, Kusumah, dan Ardah (2019) mengenai variabilitas curah hujan dan suhu udara di Tanjungpinang menunjukkan adanya fluktuasi signifikan dari tahun ke tahun, termasuk pergeseran jumlah bulan basah dan kering serta tren anomali suhu yang berkisar antara -0,6 °C hingga 1,4 °C. Temuan ini menegaskan bahwa Tanjung pinang merupakan wilayah dengan dinamika atmosferik yang tinggi, sehingga pendekatan berbasis machine learning dalam memprediksi tutupan awan menjadi sangat relevan sebagai dasar ilmiah dan teknis dalam mempertimbangkan kelayakan pembangunan pembangkit listrik tenaga surya di wilayah tersebut.

I.2 Rumusan Masalah

Adapun rumusan masalah dari penelitian ini:

1 Bagaimana pengaruh variabilitas tutupan awan terhadap intensitas radiasi sinar matahari yang diterima di permukaan bumi?

2 Bagaimana kinerja berbagai model *machine learning* dalam memprediksi tutupan awan, dan model mana yang menunjukkan tingkat akurasi terbaik berdasarkan evaluasi statistik?

I.3 Tujuan dan Manfaat

I.2.1 Tujuan

Adapun tujuan dari penelitian ini:

1. Menganalisis pengaruh tutupan awan terhadap intensitas sinar matahari yang diterima di permukaan, untuk mengidentifikasi sejauh mana pengaruhnya.
2. Mengembangkan dan mengevaluasi beberapa model *machine learning* dalam memprediksi tutupan awan, serta memilih model dengan akurasi terbaik.

I.2.2 Manfaat Penelitian

Adapun manfaat dari penelitian:

1. Pihak Pengembang Energi Surya: Penelitian ini memberikan wawasan yang lebih mendalam tentang pengaruh tutupan awan terhadap sinar matahari yang diterima oleh permukaan bumi. Dengan pemahaman ini, pengembang sistem energi surya dapat merencanakan instalasi panel surya dengan lebih efektif, mengoptimalkan performa energi surya, dan mengurangi ketidakpastian yang disebabkan oleh fluktuasi tutupan awan;
2. Perusahaan Teknologi dan Riset: Hasil penelitian ini dapat membantu perusahaan yang mengembangkan sistem prediksi cuaca berbasis *machine learning* untuk meningkatkan akurasi model prediksi tutupan awan. Hal ini sangat penting dalam meningkatkan keandalan model prediksi cuaca yang digunakan untuk perencanaan energi terbarukan atau pemodelan iklim;
3. Pemerintah dan Pembuat Kebijakan: Dengan memahami pengaruh tutupan awan terhadap energi surya, pemerintah dapat merancang kebijakan yang lebih tepat dalam mendukung pengembangan energi terbarukan, seperti solar power. Penelitian ini juga dapat membantu dalam merumuskan insentif yang lebih efektif untuk pengembangan energi bersih dan berkelanjutan;
4. Peneliti dan Akademisi**:** Penelitian ini dapat memperkaya literatur ilmiah terkait dengan meteorologi, prediksi cuaca, dan energi terbarukan, serta memberikan kontribusi dalam pengembangan model *machine learning* dalam bidang prediksi tutupan awan. Hal ini juga membuka jalan untuk penelitian lanjutan yang lebih mendalam terkait dampak perubahan cuaca dan iklim terhadap energi surya;
5. Masyarakat umum dan Industri yang tergantung pada Cuaca**:** Dengan;adanya model prediksi tutupan awan yang lebih akurat, masyarakat dan industri yang bergantung pada cuaca, seperti pertanian, transportasi, dan sektor pariwisata, dapat merencanakan aktivitas mereka dengan lebih baik, mengurangi risiko dan kerugian akibat cuaca yang tidak terduga.

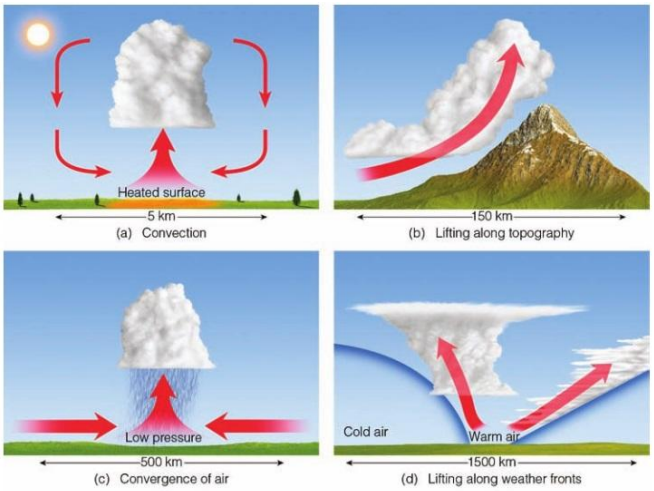
I.4 Landasan Teori

I.3.1 Dinamika Atmosfer dan Tutupan Awan

Awan merupakan kumpulan tetesan air atau kristal es yang tersuspensi di atmosfer dan terbentuk akibat kondensasi uap air di udara, memainkan peran penting dalam siklus hidrologi, cuaca, dan iklim. Dalam pembelajaran meteorologi, klasifikasi awan menjadi konsep penting karena dapat membantu memprediksi kondisi cuaca dan perubahan atmosfer. Menurut jurnal yang ditulis oleh Ajeng Putu Habsah dkk. (2023), awan diklasifikasikan berdasarkan ketinggiannya menjadi tiga kategori utama: awan rendah, menengah, dan tinggi. Awan rendah terbentuk di bawah ketinggian 2.000 meter, contohnya adalah awan stratus dan nimbostratus, yang sering menghasilkan hujan ringan hingga sedang. Awan menengah terbentuk di ketinggian antara 2.000 hingga 6.000 meter, seperti altostratus dan altocumulus, yang sering menunjukkan adanya perubahan cuaca. Sedangkan awan tinggi terbentuk di atas 6.000 meter, misalnya cirrus dan cirrostratus, yang umumnya tipis dan menandakan cuaca cerah.

Tutupan awan adalah indikator persentase langit yang tertutup oleh awan, umumnya dinyatakan dalam skala oktas, dengan nilai 0 untuk langit cerah sepenuhnya dan 8 untuk langit yang sepenuhnya tertutup. Definisi ini digunakan secara luas untuk memahami dan mengklasifikasikan awan berdasarkan pengamatan langsung atau metode berbasis instrumen (Orsini dkk., 2002). Pengukuran tutupan awan dapat dilakukan melalui pengamatan manusia atau perangkat berbasis darat, seperti radiometer, yang mampu mendeteksi radiasi gelombang pendek dan panjang yang masuk. Teknik ini membantu membedakan antara kondisi langit cerah, berawan sebagian, atau tertutup sepenuhnya (Calbó dkk., 2017).

Proses pembentukan awan dimulai dari kondensasi uap air pada partikel kecil yang bertindak sebagai inti kondensasi awan (CCN) atau inti pembentuk es (INPs), di mana partikel-partikel seperti mikroplastik dan nanoplastik (MnPs) juga dapat berperan setelah melalui proses pelapukan kimiawi atau adsorpsi molekul tertentu (Aeschlimann dkk., 2022). Selain itu, faktor meteorologi seperti kelembapan relatif di atas lapisan awan sangat memengaruhi proses ini, karena kelembapan yang tinggi dapat meningkatkan kandungan air awan dengan mengurangi efisiensi hujan, sedangkan kelembapan rendah cenderung mengurangi tutupan awan melalui penguapan (Ackerman dkk., 2004). Interaksi antara sifat fisikokimia partikel aerosol dan kondisi atmosfer ini menunjukkan pentingnya penelitian lebih lanjut untuk memahami dampaknya terhadap pembentukan awan dan keseimbangan energi bumi.



Gambar 1. Proses Pembentukan Awan (Climate4life, 2018)

Tutupan awan memiliki peran signifikan dalam dinamika perubahan iklim dan intensitas radiasi matahari yang mencapai permukaan bumi. Menurut Mendoza dkk. (2021), peningkatan suhu troposfer tengah sebesar 1°C dapat mengurangi tutupan awan hingga 1,5 persen jika kelembapan relatif tetap, dan hingga 7,6 persen jika kelembapan relatif tidak terjaga. Penurunan ini dapat menciptakan umpan balik positif yang memperparah pemanasan global karena lebih banyak radiasi matahari yang mencapai permukaan bumi, suatu mekanisme yang dipengaruhi oleh hubungan suhu dan tekanan uap jenuh seperti dijelaskan oleh persamaan Clausius-Clapeyron. Selain itu, Matuszko (2011) menunjukkan bahwa tutupan awan biasanya melemahkan intensitas radiasi matahari, meskipun pada kondisi tertentu seperti saat langit hanya sebagian tertutup (3/8-6/8) dengan awan konvektif, intensitas radiasi dapat meningkat akibat refleksi dan transmisi awan. Studi oleh Eberhardt dkk. (2016) juga menyoroti dampak musiman tutupan awan di wilayah tropis dan subtropis, yang membatasi penggunaan penginderaan jauh optik untuk pemantauan lahan selama musim hujan. Kombinasi efek refleksi, absorpsi, dan transmisi awan terhadap radiasi matahari menjadikan tutupan awan sebagai elemen kunci dalam memahami perubahan iklim serta pengaruhnya terhadap sistem cuaca, energi global, dan aplikasi praktis seperti pertanian.

I.3.2 Hubungan Tutupan Awan dan Radiasi Matahari

Radiasi matahari merupakan salah satu faktor utama yang menentukan potensi energi surya yang dapat dimanfaatkan di suatu wilayah. Radiasi ini diterima oleh permukaan Bumi dalam bentuk energi yang dapat diubah menjadi listrik dengan menggunakan panel surya. Namun, efisiensi penyerapan energi surya dapat dipengaruhi oleh faktor-faktor atmosfer, khususnya tutupan awan. Awan memiliki kemampuan untuk memantulkan sebagian besar radiasi matahari kembali ke luar angkasa, yang mengurangi jumlah energi yang mencapai permukaan Bumi. Penelitian yang dilakukan oleh Mueller dkk. (2011) menunjukkan bahwa albedo

awan—yang mengukur proporsi radiasi yang dipantulkan kembali oleh awan—berperan penting dalam menentukan jumlah radiasi yang tersedia untuk sistem pembangkit tenaga surya. Semakin banyak tutupan awan, semakin sedikit radiasi yang mencapai permukaan, sehingga mengurangi potensi energi surya.

Di sisi lain, variasi dalam tutupan awan juga dapat memengaruhi fluktuasi radiasi yang diterima oleh panel surya, yang berpotensi mengurangi kestabilan produksi energi. Fountoulakis dkk. (2021) menemukan bahwa awan dapat mengurangi radiasi permukaan matahari antara 25-50% tergantung pada jenis dan ketebalan awan. Hal ini menjadi penting untuk diperhatikan dalam perencanaan pembangunan pembangkit listrik tenaga surya di Tanjung Pinang, karena daerah tersebut memiliki potensi energi surya yang tinggi, namun pengaruh tutupan awan yang bervariasi bisa memengaruhi efektivitas pemanfaatan energi tersebut. Selain itu, penelitian oleh Mauladhania dkk. (2023) menekankan bahwa pemahaman yang mendalam mengenai pola tutupan awan dan radiasi matahari sangat diperlukan untuk merancang sistem tenaga surya yang efisien dan tahan terhadap variabilitas cuaca. Dalam hal ini, penting untuk memonitor dan menganalisis pola awan yang ada untuk memaksimalkan potensi energi terbarukan di wilayah Tanjung Pinang. Manabe dan Moller (1961) juga menunjukkan bahwa perubahan dalam tutupan awan dan radiasi matahari dapat mempengaruhi keseimbangan energi di atmosfer, yang berhubungan langsung dengan efisiensi sistem pembangkit listrik tenaga surya.

I.3.3 Konsep Dasar *Machine learning*

Machine learning adalah cabang kecerdasan buatan yang bertujuan untuk membuat komputer mampu memprediksi secara akurat berdasarkan pengalaman sebelumnya. Dalam konteks ini, pengalaman tersebut direpresentasikan sebagai data, yang digunakan untuk melatih model melalui proses yang dikenal sebagai "training." Model yang terlatih dapat memetakan hubungan kompleks antara variabel input dan output, sehingga memungkinkan prediksi untuk data baru yang tidak pernah terlihat sebelumnya (Baştanlar & Özuysal, 2014). Proses ini sangat penting dalam berbagai disiplin ilmu, termasuk bioinformatika, di mana analisis data yang kompleks seringkali memerlukan pendekatan otomatis untuk meningkatkan efisiensi dan akurasi.

Metode supervised learning, seperti *Logistic Regression*, support vector machines (SVM), dan random forest, menjadi pilihan utama untuk tugas klasifikasi biner karena kemampuannya dalam mengidentifikasi pola dalam data berlabel. Penelitian oleh Graf dkk. (2022) menunjukkan bahwa algoritma seperti random forest seringkali mengungguli metode tradisional seperti linear discriminant analysis (LDA) dalam tugas klasifikasi tertentu, terutama pada data dengan distribusi non-normal. Di sisi lain, pendekatan ranking dengan SVM juga digunakan untuk menentukan kualitas prediksi dalam berbagai domain, menunjukkan fleksibilitas *machine learning* dalam menangani masalah kompleks, seperti yang dijelaskan oleh Xu dkk. (2022). Dengan berbagai algoritma yang tersedia, pemilihan model yang tepat bergantung pada tujuan spesifik, karakteristik data, dan kebutuhan performa.

1. *Logistic Regression*

Logistic Regression adalah metode analisis statistik yang digunakan untuk memodelkan hubungan antara variabel prediktor (baik kontinu maupun kategorikal) dan variabel respons yang bersifat biner atau dikotom. Model ini memiliki keunggulan dalam menganalisis data observasi, terutama ketika terdapat kebutuhan untuk mengontrol bias potensial akibat perbedaan antar kelompok (LaValley, 2008). Dalam konteks meteorologi, *Logistic Regression* sering digunakan untuk klasifikasi tutupan awan, prediksi curah hujan, dan pengklasifikasian peristiwa cuaca berdasarkan variabel lingkungan seperti suhu, kelembapan, dan curah hujan (Shah dkk., 2013). Secara matematis, *Logistic Regression* memodelkan probabilitas kejadian (P(Y=1)P(Y=1)) sebagai fungsi dari logit, yaitu logaritma natural dari odds:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \qquad 1.1$$

di mana PP adalah probabilitas kejadian, β0\beta_0 adalah intercept, dan $\beta 1, \beta 2, \ldots, \beta p, \beta_1, \beta_2, \ldots$ adalah koefisien regresi untuk prediktor X1,X2,…,XpX_1, X_2, \dots, X_p,. Model ini memungkinkan estimasi probabilitas dengan mempertimbangkan efek masing-masing prediktor.

Dalam penelitian ini, *Logistic Regression* digunakan untuk memprediksi pola tutupan awan dengan input dari variabel meteorologi yang relevan. Pendekatan ini mirip dengan penelitian Shah dkk. (2013), di mana model *Logistic Regression* diterapkan untuk memprediksi epidemik Fusarium Head Blight menggunakan variabel cuaca seperti kelembapan relatif, suhu, dan curah hujan. Selain itu, penelitian lain menunjukkan bahwa implementasi *Logistic Regression* pada data besar, seperti klasifikasi gambar hiperspektral menggunakan cloud computing, dapat meningkatkan efisiensi analisis (Haut dkk., 2017). Dengan demikian, *Logistic Regression* menjadi alat yang sangat berguna untuk mendukung analisis dan klasifikasi berbasis cuaca dalam domain meteorologi.

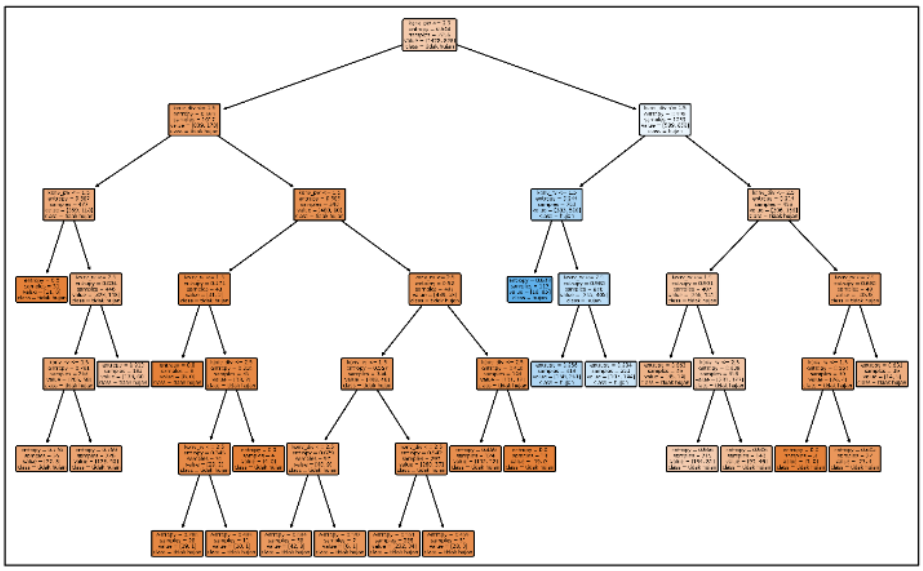
2. *Decision Tree*

Decision Tree adalah algoritma pembelajaran mesin yang telah berkembang sejak era awal perkembangan sirkuit digital pada abad ke-20. Sebagai alat prediksi dan klasifikasi, *Decision Tree* terkenal karena hasilnya yang mudah dipahami, berkat representasinya dalam bentuk pohon hierarkis yang intuitif (Barry de Ville, 2013). Algoritma ini menggunakan pendekatan pemecahan masalah berbasis aturan dengan membagi data menjadi subset berdasarkan atribut tertentu pada setiap node hingga menghasilkan keputusan di daun (leaf). Pendekatan ini tidak hanya fleksibel tetapi juga sangat robust dalam menghadapi data yang hilang atau kualitas data yang bervariasi. Dalam domain statistik dan pembelajaran mesin, *Decision Tree* telah dimanfaatkan secara luas untuk berbagai tugas, termasuk klasifikasi, prediksi, dan penemuan pola, karena kemampuannya untuk menghasilkan model prediksi yang akurat dan mudah diinterpretasi.

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i(log_2\ p_i)$$
$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\ Entropy\ (S_i) \quad 1.2$$

Dalam penelitian terkait meteorologi, *Decision Tree* seperti algoritma C4.5 dapat digunakan untuk memprediksi fenomena cuaca, termasuk curah hujan, awan, atau kondisi atmosfer lainnya. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang digunakan untuk membentuk pohon keputusan berbasis informasi (information gain) dan rasio gain (*gain ratio*). C4.5 bekerja dengan cara membagi data berdasarkan atribut-atribut yang paling informatif untuk menghasilkan keputusan klasifikasi. Keunggulannya terletak pada kemampuannya menangani data dengan nilai numerik maupun kategori, mengatasi data hilang, serta melakukan pemangkasan (*pruning*) untuk menghindari overfitting.

Misalnya, penelitian Adi Prasetiyo (2024) menunjukkan bahwa algoritma C4.5 memiliki potensi untuk mengolah data atmosfer yang kompleks dan tidak kontinu, sehingga menghasilkan akurasi prediksi hujan tahunan sebesar 84%. Hal ini menunjukkan bahwa C4.5 mampu menyederhanakan pola iklim yang rumit menjadi model prediktif yang dapat diinterpretasikan dengan baik, menjadikannya alat yang relevan dalam studi iklim dan prediksi cuaca di wilayah tropis seperti Indonesia.



Gambar 2. Algoritma *Decision Tree* (Prasetyo, 2024)

Studi lain yang menggunakan *Decision Tree*, seperti yang dilakukan oleh Sašo Džeroski dkk. (2024), menunjukkan bahwa algoritma ini mampu memanfaatkan data penginderaan jauh (Landsat dan LIDAR) untuk memprediksi parameter tutupan kanopi hutan dan ketinggian tegakan hutan. Hal ini dapat mendukung peramalan tutupan awan di wilayah Tanjung Pinang sebagai pertimbangan untuk pembangunan panel surya.

3. Naïve Bayes

Naive bayes (NB) adalah metode klasifikasi probabilistik yang berbasis pada teorema Bayes, dengan asumsi bahwa setiap fitur dalam data bersifat independen satu sama lain. Metode ini efektif dalam mengatasi masalah klasifikasi dengan data yang besar dan tidak terstruktur, serta memiliki kecepatan komputasi yang tinggi (Murphy, 2006). *Naive bayes* digunakan dalam berbagai aplikasi, seperti klasifikasi curah hujan (Azmi dkk., 2024) dan ramalan radiasi matahari global (Kwon dkk., 2019), dengan mempertimbangkan variabel cuaca yang relevan. Keunggulan utama dari metode ini adalah kesederhanaannya, di mana model hanya memerlukan informasi dasar mengenai probabilitas kondisi kelas, sehingga memberikan hasil yang cepat tanpa memerlukan perhitungan yang kompleks. Meskipun mengasumsikan independensi antar fitur, *Naive bayes* tetap mampu memberikan hasil yang memadai pada banyak aplikasi praktis, terutama ketika data yang digunakan besar dan variabel-variabelnya saling berhubungan secara jelas.

Dalam konteks penelitian ini, yang bertujuan untuk memprediksi tutupan awan sebagai pertimbangan dalam pembangunan Pembangkit Listrik Tenaga Surya (PLTS), *Naive bayes* dapat digunakan untuk mengklasifikasikan kondisi cuaca berdasarkan variabel-variabel yang mempengaruhi tutupan awan, seperti kelembaban, suhu, dan tekanan udara. Dengan memanfaatkan data cuaca historis, algoritma *Naive bayes* dapat memprediksi kategori tutupan awan, seperti ringan, sedang, atau berat, yang berpengaruh langsung terhadap efisiensi pembangkit listrik tenaga surya. Hasil prediksi yang akurat, seperti yang ditemukan dalam penelitian sebelumnya (Azmi dkk., 2024), menunjukkan bahwa *Naive bayes* merupakan alat yang efektif untuk meramalkan kondisi yang dapat mengoptimalkan pengoperasian PLTS, serta mengurangi ketidakpastian terkait fluktuasi radiasi matahari yang berdampak pada pembangkit listrik tersebut.

4. *Gradient boosting*

Gradient boosting Machines (GBMs) adalah salah satu metode yang efektif dalam pembelajaran mesin untuk tugas regresi dan klasifikasi. Dalam artikel yang ditulis oleh Alexey Natekin dan Alois Knoll (2013), GBM dijelaskan sebagai teknik ensemble yang membangun model secara bertahap, dengan setiap model baru memfokuskan pembelajaran pada kesalahan yang dibuat oleh model sebelumnya. Penurunan gradien digunakan untuk mengoptimalkan fungsi kerugian, dan ini memungkinkan metode ini untuk sangat fleksibel dan dapat disesuaikan dengan berbagai jenis data dan tugas. Ini menjadikannya sangat efektif dalam berbagai aplikasi, dari pengenalan pola hingga kontrol robotik dan klasifikasi teks.

Penelitian oleh Suri Babu Nuthalapati dan Aravind Nuthalapati (2024) menyoroti penerapan GBM dalam prakiraan cuaca, dimana teknik ini menunjukkan hasil yang mengesankan dalam memproses data cuaca yang kompleks. Dengan menggabungkan GBM dengan model prediksi numerik cuaca, mereka berhasil meningkatkan akurasi prediksi kecepatan angin hingga 80,95%. Dalam konteks ini, kemampuan GBM untuk menangani hubungan non-linear dan pola kompleks sangat

penting, karena cuaca melibatkan dinamika atmosfer yang sangat rumit. Sementara itu, studi oleh Wenqing Xu dkk. (2020) juga menunjukkan aplikasi GBM dalam memperbaiki hasil prediksi kecepatan angin dengan memanfaatkan data dari model WRF (Weather Research and Forecasting). Hasilnya menunjukkan bahwa GBM dapat meningkatkan prediksi kecepatan angin yang lebih akurat dibandingkan dengan model lain seperti *Decision Tree* regression dan multi-layer perceptron regression, dengan mengurangi kesalahan prediksi secara signifikan.

5. *Simple Linear Regression*

Simple regresi linier merupakan metode statistik yang digunakan untuk memodelkan hubungan linier antara dua variabel kontinu. Tujuan utama dari analisis ini adalah untuk memprediksi nilai variabel respons (variabel terikat) berdasarkan nilai variabel prediktor (variabel bebas). Model regresi linier sederhana dapat dijelaskan dengan persamaan berikut:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1.3)$$

di mana:

- Y adalah variabel respons
- X adalah variabel prediktor
- β0 adalah intersep (nilai Y ketika X = 0)
- β1 adalah koefisien regresi (kemiringan garis regresi)
- ε adalah galat acak (kesalahan dalam prediksi)

Koefisien regresi (β1) menunjukkan perubahan rata-rata pada Y untuk setiap unit perubahan pada X. Analisis regresi linier memungkinkan peneliti untuk menguji signifikansi hubungan antara variabel prediktor dan respons serta memperkirakan nilai respons untuk nilai-nilai tertentu dari variabel prediktor. (Zou, Tuncali, & Silverman, 2003)

I.3.4 Relevansi Meteorologi Terhadap PLTS

Meteorologi memiliki peran signifikan dalam mendukung pembangunan dan pengelolaan Pembangkit Listrik Tenaga Surya (PLTS), khususnya melalui analisis parameter seperti radiasi matahari, tutupan awan, dan pengaruh aerosol atmosfer. Menurut Marzouk (2023), penilaian intensitas generasi energi (Energy Generation Intensity, EGI) dari berbagai teknologi surya menunjukkan bahwa variabilitas radiasi harian dan musiman sangat memengaruhi efisiensi sistem PV, terutama di wilayah dengan tingkat radiasi tinggi seperti dekat Tropic of Cancer. Hal ini relevan dengan penelitian di Tanjung Pinang, Kepulauan Riau, di mana analisis tutupan awan diperlukan untuk mengoptimalkan potensi radiasi matahari dan desain sistem PLTS yang adaptif terhadap fluktuasi lokal. Penelitian Marzouk menyoroti pentingnya orientasi optimal dan pelacakan panel surya, yang dapat dikembangkan lebih lanjut dengan pendekatan berbasis *machine learning* untuk memprediksi kondisi meteorologi spesifik seperti tutupan awan.

Selain itu, studi Shen dkk. (2013) tentang Turpan Solar Eco-City menunjukkan bahwa peramalan radiasi matahari jangka pendek berbasis model meteorologi seperti WRF dapat digunakan untuk mengatur grid pintar dan meningkatkan efisiensi sistem PLTS. Dengan mempertimbangkan bahwa tutupan awan adalah salah satu faktor utama yang memengaruhi radiasi matahari, penelitian Anda berfokus pada prediksi tutupan awan menggunakan *machine learning* akan memberikan kontribusi signifikan terhadap desain dan pengelolaan PLTS di Tanjung Pinang. Temuan Norris dan Wild (2007) juga menegaskan relevansi penelitian ini, di mana tren "dimming" dan "brightening" radiasi matahari akibat aerosol dan variabilitas awan memberikan wawasan penting tentang dampak perubahan atmosfer terhadap energi surya. Integrasi pendekatan meteorologi dan *machine learning* yang Anda lakukan dapat menjadi solusi inovatif untuk meningkatkan keandalan perencanaan energi surya di wilayah tropis.

I.3.5 *Heatmap correlation* dan Data Meteorologi

Heatmap korelasi adalah metode visualisasi yang digunakan untuk melihat hubungan antar variabel dalam dataset, terutama dalam analisis data meteorologi. Teknik ini mempermudah pengguna untuk memahami pola dan hubungan antar variabel seperti suhu, kelembaban, tekanan udara, dan curah hujan melalui warna. Variabel dengan korelasi tinggi ditunjukkan oleh warna mencolok, sehingga pola hubungan yang kuat dapat diidentifikasi lebih mudah. Dalam penelitian energi, *heatmap* korelasi digunakan untuk menganalisis faktor-faktor yang memengaruhi efisiensi energi, seperti yang dilakukan oleh Leni dkk. (2023), yang menunjukkan bagaimana alat ini membantu mengidentifikasi variabel signifikan seperti penggunaan energi dan kapasitas pendinginan dalam meningkatkan efisiensi AC.

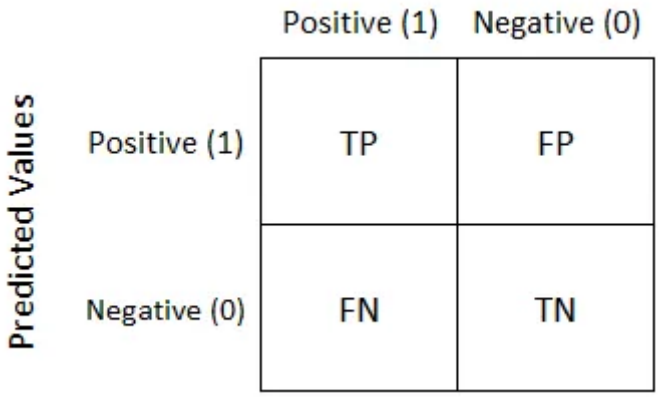
$$r_{xy} = \frac{\Sigma xy}{(n-1)S_x S_y} \tag{1.4}$$

Leni dkk. (2023) menggunakan korelasi Pearson untuk menganalisis hubungan linear antara variabel input (misalnya, luas permukaan, orientasi) dan output (beban pemanasan dan pendinginan) dalam pengembangan model Artificial Neural Network (ANN). Dengan *heatmap* korelasi, variabel dengan korelasi rendah terhadap target dihilangkan untuk mengurangi dimensi data tanpa mengorbankan akurasi prediksi. Seleksi fitur berbasis korelasi ini meningkatkan efisiensi model, mencegah overfitting, dan menghasilkan model ANN dengan performa tinggi (nilai R-squared mendekati 1). Pendekatan ini menunjukkan bahwa *heatmap* korelasi adalah alat yang efektif untuk seleksi fitur dalam berbagai bidang, termasuk meteorologi dan pemodelan prediktif.

I.3.6 Train Test Split dan *Confusion matrix*

Train-test split adalah teknik pembagian dataset menjadi dua subset: data pelatihan (training set) dan data pengujian (testing set). Data pelatihan digunakan untuk melatih model agar mengenali pola, sementara data pengujian digunakan untuk

mengevaluasi performa model pada data baru. Biasanya, proporsi pembagian data adalah 70:30, 80:20, atau 90:10, dengan bagian lebih besar untuk pelatihan. Dalam penelitian yang dilakukan oleh Jimin Tan dkk. (2023), penulis mengevaluasi kelemahan metode split konvensional, terutama pada kasus di mana dataset terbatas dan mahal untuk dikumpulkan, seperti dalam bidang farmasi. Mereka mengusulkan pendekatan Active Learning (AAL), yang lebih adaptif dibandingkan metode tradisional. Penelitian oleh Quang Hung Nguyen dkk. (2021) menunjukkan bahwa performa model *machine learning*, seperti ANN, Extreme Learning Machine (ELM), dan Boosted Trees, sangat dipengaruhi oleh rasio train-test split. Dalam penelitian tersebut, rasio 70:30 memberikan hasil terbaik dengan model ANN menunjukkan akurasi tertinggi.



Gambar 3. *Confusion matrix* (TowardDataScience, 2018)

Confusion matrix adalah matriks evaluasi yang menggambarkan performa klasifikasi model pada data pengujian. Matriks ini terdiri atas empat komponen: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dari matriks ini, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-*score* dapat dihitung. Akurasi dihitung dengan rumus:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad 1.5$$

Relevansi train-test split dan *Confusion matrix* terhadap penelitian ini sangat signifikan, terutama dalam konteks evaluasi performa model *machine learning*. Memilih rasio train-test split yang sesuai dan menggunakan *Confusion matrix* sebagai alat evaluasi dapat membantu menilai efektivitas model dengan lebih baik, memastikan bahwa model dapat menangani data nyata dengan akurat. Metode seperti KNN yang digunakan pada penelitian Argina (2020) dan ANN pada penelitian Sanjaya dkk. (2023) memberikan wawasan bagaimana memilih metode dan evaluasi yang tepat untuk dataset yang digunakan.

I.3.7 Outgoing Longwave Radiation

Outgoing Longwave Radiation (OLR) merupakan proxy satelit untuk menggambarkan konveksi mendalam di atmosfer tropis. Nilai OLR tinggi (misalnya > 240 W m^{-2}) merepresentasikan langit cerah atau awan tipis di mana radiasi inframerah dapat keluar ke ruang angkasa dengan bebas, sedangkan nilai OLR rendah (< 220 W m^{-2}) menandakan awan konvektif tebal yang menutupi permukaan dan menyerap kembali radiasi inframerah (Kiladis et al., 2014). Dengan demikian, daerah OLR rendah berkorespondensi langsung dengan aktivitas konveksi kuat, sementara daerah OLR tinggi mencerminkan kondisi tenggat (suppressed) konveksi.

Dalam studi variabilitas intramuskim, Nakazawa (1986) menunjukkan bahwa fluktuasi OLR rendah yang bergerak ke arah timur sepanjang ekuator menjadi ciri khas gelombang Kelvin dan siklon tropis super-clusters. Korelasi negatif kuat antara OLR rendah dan curah hujan harian menegaskan bahwa semakin rendah nilai OLR, semakin intens konveksi dan potensi hujan lebat. Sebaliknya, episod OLR tinggi (> 250 W m^{-2}) sering berlangsung 10–20 hari dan berkaitan dengan penurunan kelembaban kolom, peningkatan radiasi permukaan, serta penurunan laju penguapan laut.

Kiladis et al. (2014) memperkuat temuan ini dengan membandingkan indeks OLR-basis dan indeks sirkulasi untuk melacak Madden–Julian Oscillation (MJO). Mereka menetapkan ambang OLR ≤ 200 W m^{-2} sebagai kriteria awan konvektif aktif MJO, sedangkan OLR ≥ 235 W m^{-2} digunakan untuk mengidentifikasi fase tenggat. Analisis komposit menunjukkan bahwa transisi dari fase OLR tinggi ke rendah berlangsung sekitar 5–7 hari dan diikuti oleh peningkatan kecepatan angas permukaan barat serta pembentangan sirkulasi Walker. Hasil ini menegaskan bahwa pemantauan perubahan OLR menjadi kunci pemahaman dinamika skala besar tropis.

Kesimpulannya, interpretasi OLR rendah dan tinggi tidak hanya memberikan informasi spasial lokasi konveksi, tetapi juga temporal—seberapa cepat sistem konvektif berkembang atau menurun. Oleh karena itu, kombinasi ambang OLR < 220 W m^{-2} untuk konveksi aktif dan > 240 W m^{-2} untuk kondisi tenggat telah menjadi standar dalam literatur iklim tropis (Nakazawa, 1986; Kiladis et al., 2014). Pemanfaatan ambang ini memungkinkan peneliti secara konsisten memantau evolusi MJO, gelombang Kelvin, dan fenomena intramuskim lainnya.

I.3.8 Pengaruh Tutupan Awan terhadap Performa Pembangkit Listrik Tenaga Surya (PLTS) di Tanjung Pinang

Panel surya bekerja dengan memanfaatkan efek fotovoltaik, di mana sinar matahari yang mengenai lapisan semikonduktor silikon pada panel melepaskan elektron sehingga menghasilkan arus listrik searah (DC) yang disimpan dalam baterai melalui charge controller untuk menjaga kestabilan tegangan; selanjutnya, arus DC diubah menjadi arus bolak-balik (AC) oleh inverter sehingga dapat menyuplai berbagai perangkat otomasi stadion seperti lampu, sensor, dan motor, sementara ketika

cahaya tidak tersedia baterai secara otomatis menjadi sumber cadangan untuk menjaga sistem tetap beroperasi (Julisman dkk., 2017).

Tutupan awan merupakan faktor meteorologi yang paling dominan menurunkan intensitas radiasi matahari yang sampai ke permukaan modul fotovoltaik (Gunoto & Sofyan, 2020). Ketika awan tebal menutupi langit, nilai radiasi global horizontal irradiance (GHI) dapat turun hingga 70 % dari kondisi cerah, sehingga daya keluaran PLTS berkurang secara proporsional. Hutajulu et al. (2020) mencatat bahwa pada PLTS on-grid 100 kWp di Ecopark Ancol, fluktuasi harian output akibat awan dapat menyebabkan deviasi 15–25 % dari prakiraan energi yang dihasilkan sistem monitoring pusat. Fenomena ini menegaskan pentingnya informasi awan real-time untuk menjaga stabilitas jaringan dan memaksimalkan aliran daya ke grid.

Di Tanjung Pinang—yang memiliki pola awan konvektif tropis dengan puncak pada sore hari—ketidakpastian irradiance akibat awan menjadi lebih besar. Nurjaman & Purnama (2022) menunjukkan bahwa sistem PLTS rumah tangga 1 kWp di wilayah pesisir dapat kehilangan 0,5–0,7 kWh per hari ketika indeks kecerahan langit (clearness index) turun di bawah 0,4 karena awan kumulonimbus. Oleh sebab itu, model machine learning yang mampu memprediksi fraksi tutupan awan 1–3 jam ke depan sangat dibutuhkan agat operator dapat mengatur baterai atau beban dinamis untuk menutupi kekurangan daya tersebut.

Selain dampak langsung terhadap energi, variabilitas akibat awan juga memengaruhi komponen Balance of System (BOS). Gunoto & Sofyan (2020) melaporkan bahwa pergantian cepat antara kondisi cerah dan berawan menyebabkan thermal cycling pada modul 100 Wp, yang berpotensi menurunkan umur pakai inverter sebesar 10 % setiap tahun bila tidak ada kontrol prediktif. Dengan memanfaatkan prediksi tutupan awan berbasis machine learning, sistem dapat mengantisipasi lonjakan tegangan dan arus melalui algoritma Maximum Power Point Tracking (MPPT) adaptif, sehingga memperpanjang masa operasional peralatan.

BAB 2
METODOLOGI PENELITIAN

2.1 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari situs *metomanz.com*. Data tersebut mencakup informasi meteorologi Kota Tanjung Pinang, Kepulauan Riau, dengan variabel-variabel meliputi suhu udara, kelembapan relatif, kecepatan angin, durasi penyinaran matahari, tutupan awan, serta curah hujan.

2.2 Prosedur Penelitian

2.1.1 Pembacaan Data dan Quality Control

1. Pembacaan Data

Langkah awal dimulai dengan membaca file data mentah yang berformat Excel. Proses ini memastikan bahwa seluruh data dari file telah dimuat ke dalam struktur yang dapat diolah lebih lanjut. Informasi awal tentang dataset, seperti jumlah baris dan kolom, tipe data pada setiap kolom, dan jumlah entri, diperiksa untuk memberikan gambaran umum kondisi data.

2. Pembersihan Kolom Tidak Relevan

Kolom yang tidak relevan atau tidak memiliki keterkaitan langsung dengan analisis, seperti kolom otomatis yang dihasilkan oleh perangkat lunak (*e.g.*, kolom "Unnamed"), dihapus dari dataset. Hal ini dilakukan untuk menyederhanakan struktur data dan menghindari kebingungan selama proses analisis.

3. Konversi Tipe Data

Kolom dengan tipe data yang tidak sesuai, seperti kolom yang dikenali sebagai teks (*object*), diperiksa dan dikonversi ke tipe data yang benar, seperti angka desimal (*float*). Jika terdapat data yang tidak dapat dikonversi, sistem akan mengidentifikasi kolom yang bermasalah untuk diatasi lebih lanjut.

4. Identifikasi dan Penanganan Data Kosong

Setiap kolom dalam dataset diperiksa untuk mengetahui jumlah data kosong atau nilai yang hilang. Persentase data kosong dihitung untuk setiap kolom. Berdasarkan hasil analisis ini: Kolom dengan kurang dari 50% data kosong diisi dengan nilai rata-rata kolom tersebut, sehingga distribusi data tetap terjaga. Kolom dengan lebih dari 50% data kosong dihapus karena data yang terlalu sedikit dianggap tidak cukup representatif untuk analisis.

5. Pemeriksaan Akhir Dataset

Setelah proses pembersihan, struktur dan kondisi dataset diperiksa kembali. Hal ini mencakup pemeriksaan jumlah kolom, jumlah baris, tipe data, dan memastikan bahwa semua data kosong telah diatasi.

6. Dokumentasi dan Penyimpanan Dataset Bersih

Dataset yang telah bersih disimpan untuk digunakan pada tahap analisis selanjutnya. Dokumentasi terkait langkah-langkah pembersihan data disimpan untuk referensi dan transparansi proses.

Catatan:
Untuk detail implementasi teknis, lihat Lampiran 1, yang berisi skrip lengkap terkait pembacaan data dan kontrol kualitas data. Skrip ini berfungsi sebagai pendukung proses dan dapat digunakan kembali untuk pengolahan data serupa.

2.1.2 Prosedur Klasifikasi kondisi awan: Berawan dan Tidak Berawan

Prosedur Pengelompokan Data Berdasarkan Kondisi Awan

1. Identifikasi Variabel

Variabel yang menjadi fokus pengelompokan adalah Cloud c., yang berisi data kuantitatif terkait kondisi awan. Nilai dalam variabel ini digunakan untuk menentukan kelompok awan berdasarkan rentang tertentu.

2. Penentuan Kategori Kelompok Awan

Prosedur pengelompokan dilakukan dengan mendefinisikan aturan berikut: Kategori 0: Kondisi awan dengan nilai antara 1 hingga 4. Ini merepresentasikan kondisi dengan tutupan awan rendah. Kategori 1: Kondisi awan dengan nilai antara 5 hingga 8. Ini merepresentasikan kondisi dengan tutupan awan tinggi. Kategori Tidak Valid (None): Untuk nilai di luar rentang yang ditentukan (di bawah 1 atau di atas 8).

3. Penerapan Fungsi Pengelompokan

Sebuah fungsi khusus bernama kelompok_awan() didefinisikan untuk mengelompokkan nilai dalam variabel Cloud c.. Fungsi ini melakukan: Memeriksa setiap nilai dalam kolom Cloud c.. Menentukan kode kelompok awan berdasarkan aturan kategori yang telah ditetapkan. Mengembalikan hasil pengelompokan ke dalam kolom baru bernama kelompok_awan_code.

4. Analisis Distribusi Kelompok Awan

Setelah pengelompokan selesai, dilakukan perhitungan jumlah setiap kelompok awan menggunakan metode *value counts*. Hasil distribusi ini memberikan gambaran tentang frekuensi data dalam masing-masing kelompok awan (kategori 0 atau 1).

Proses ini bertujuan untuk menyederhanakan analisis data dengan mengelompokkan variabel awan ke dalam kategori tertentu, sehingga memudahkan

proses eksplorasi dan analisis lebih lanjut, seperti korelasi dengan variabel cuaca lainnya.

2.1.3 Pengelompokan Variabel lain

Prosedur pengelompokan dan klasifikasi data cuaca ini dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka *pandas* untuk pengelolaan data. Tujuan dari proses ini adalah untuk menyederhanakan analisis data dengan mengelompokkan variabel numerik menjadi kategori yang dapat direpresentasikan sebagai kode angka, sehingga dapat digunakan lebih efisien dalam analisis statistik dan pemodelan. Modul yang digunakan dalam proses ini adalah `pandas`, yang diimpor dengan sintaks `import pandas as pd`.

Fungsi-fungsi utama yang digunakan dalam proses ini antara lain `pd.cut()` untuk membagi data numerik menjadi interval kategori, `cat.codes` untuk mengubah label kategori menjadi angka, `apply()` untuk menerapkan fungsi khusus ke nilai kolom, `mode()` untuk menemukan nilai paling sering muncul, `fillna()` untuk mengisi nilai kosong, dan `to_excel()` untuk menyimpan hasil akhir ke dalam file Excel.

Pengelompokan pertama dilakukan terhadap variabel suhu udara harian yang terdapat dalam kolom `Ave. T. (°C)`. Nilai suhu dikelompokkan ke dalam enam kategori, yaitu kurang dari 15°C, 15–20°C, 20–25°C, 25–30°C, 30–35°C, dan lebih dari 35°C. Hasil pengelompokan ini disimpan dalam kolom `kelompok_suhu`, sedangkan versi numeriknya disimpan dalam kolom `kelompok_suhu_code`.

Selanjutnya, variabel kelembapan relatif (`RH2M`) juga dikelompokkan ke dalam lima kategori berdasarkan tingkat kelembapan, yaitu dari rentang kelembapan sangat rendah (0–20%) hingga sangat tinggi (80–100%). Hasilnya disimpan dalam kolom `kelompok_kelembapan` dan versi numeriknya pada `kelompok_kelembapan_code`.

Variabel kecepatan angin (`Wind sp. (Km/h)`) dikelompokkan dalam enam kategori berdasarkan kecepatan dari angin sangat tenang hingga sangat kencang, dan hasilnya dimasukkan ke dalam `kelompok_angin` dan `kelompok_angin_code`.

Untuk variabel curah hujan (`Prec. (mm)`), pengelompokan dilakukan berdasarkan rentang 0–5 mm hingga lebih dari 100 mm. Data yang telah diklasifikasi disimpan pada kolom `kelompok_curah_hujan` dan versi numeriknya pada `kelompok_curah_hujan_code`.

Selanjutnya, variabel lama penyinaran matahari (`Insolat. (hours)`) dikelompokkan berdasarkan durasi harian dari kategori sangat rendah (0–2 jam) hingga sangat tinggi (lebih dari 10 jam). Hasil klasifikasi ini disimpan dalam `kelompok_insolate` dan versi angka dalam `kelompok_insolate_code`. Nilai

kosong yang ditemukan dalam data insolasi diisi menggunakan nilai modus dari kolom tersebut agar tidak terjadi gangguan dalam proses analisis.

Pengelompokan juga diterapkan pada variabel tekanan udara (`S.L.Press./Gheopot.`) ke dalam tujuh kategori berdasarkan tekanan atmosfer dari 950 hingga lebih dari 1030 hPa. Data hasil kategorisasi disimpan dalam `kelompok_pressure` dan `kelompok_pressure_code`, dengan nilai kosong diisi oleh modus kolom tersebut.

Terakhir, variabel kondisi tutupan awan (`Cloud c.`) diklasifikasikan berdasarkan dua kelompok, yaitu tutupan awan rendah (kode 0 untuk nilai 1–4) dan tutupan awan tinggi (kode 1 untuk nilai 5–8). Fungsi `kelompok_awan()` digunakan untuk menghasilkan kolom `kelompok_awan_code`.

Sebagai output akhir, hanya kolom-kolom yang memiliki akhiran `code` yang disimpan ke dalam file Excel dengan nama `data_kelompok_code.xlsx` menggunakan metode `to_excel()`. File ini dapat digunakan sebagai input lebih lanjut untuk analisis statistik atau pemodelan prediktif berbasis machine learning.

2.1.4 *Heatmap correlation*

Prosedur untuk membuat *heatmap* korelasi dimulai dengan mempersiapkan data, menghitung matriks korelasi antar variabel, dan kemudian memvisualisasikan hasilnya dalam bentuk *heatmap*. Berikut adalah tahapan-tahapan yang dilakukan:

Data dibaca dari file Excel menggunakan pd.read_excel. Pada tahap ini, file yang digunakan adalah data mentah.xlsx. Kolom yang tidak relevan atau yang memiliki nama 'Unnamed' akan dihapus dengan menggunakan df2.loc[:, ~df2.columns.str.contains('^Unnamed')]. Hal ini memastikan bahwa hanya kolom yang memiliki data yang relevan yang akan digunakan dalam analisis lebih lanjut.

1 Mengonversi Data ke Tipe Numerik
Setelah data dibersihkan dari kolom yang tidak diperlukan, langkah selanjutnya adalah mengonversi semua kolom menjadi tipe data numerik menggunakan apply(pd.to_numeric, errors='coerce'). Proses ini memastikan bahwa data yang ada dapat digunakan untuk perhitungan matematis, seperti menghitung korelasi antar variabel. Jika ada data yang tidak dapat dikonversi ke angka, nilai tersebut akan menjadi NaN (*Not a Number*).

2 Menghitung Matriks Korelasi
Korelasi antar variabel dihitung menggunakan metode *.corr*() yang menghasilkan matriks korelasi. Matriks ini menggambarkan hubungan antara variabel-variabel yang ada, dengan nilai korelasi yang menunjukkan sejauh mana dua variabel saling berhubungan. Nilai korelasi berada dalam rentang -1 hingga 1, dengan 1 menunjukkan hubungan positif yang sempurna, -1 menunjukkan hubungan negatif yang sempurna, dan 0 menunjukkan tidak ada hubungan.

3 Membuat *Heatmap* Korelasi
Setelah matriks korelasi dihitung, langkah selanjutnya adalah memvisualisasikan hasilnya dalam bentuk *heatmap*. *Heatmap* ini dibuat menggunakan sns.*heatmap* dari pustaka Seaborn, yang menghasilkan peta warna berdasarkan nilai korelasi. Parameter annot=True digunakan untuk menampilkan nilai korelasi pada setiap sel dalam *heatmap*, dan cmap='coolwarm' menentukan palet warna yang digunakan untuk menunjukkan kekuatan dan arah korelasi. fmt='.2f' menentukan format angka yang ditampilkan, dengan dua angka di belakang koma.

4 Penyesuaian Tampilan *Heatmap*
Ukuran gambar *heatmap* disesuaikan dengan plt.figure(figsize=(10, 10)) agar lebih mudah dibaca. Selain itu, ukuran font untuk label pada sumbu x dan y juga disesuaikan dengan plt.xticks(fontsize=15) dan plt.yticks(fontsize=15) untuk memastikan keterbacaan yang baik. Terakhir, judul *heatmap* ditambahkan dengan plt.title('*Heatmap* Korelasi Variabel').
Detail skrip dapat dilihat pada Lampiran 1. Skrip tersebut menunjukkan langkah-langkah teknis dalam pembuatan *heatmap* korelasi yang digunakan dalam penelitian ini.

2.1.5 Model *Machine learning*

1 Pembacaan Data
Data yang digunakan adalah hasil pengolahan dan pengelompokan variabel cuaca yang telah diproses sebelumnya. Data ini memuat fitur-fitur cuaca yang telah dikelompokkan dalam kategori numerik serta target yang akan diprediksi, yaitu klasifikasi kelompok awan.

2 Penentuan Fitur dan Target
Dalam model ini, fitur-fitur yang digunakan meliputi kode kelompok dari beberapa variabel cuaca, seperti suhu, kelembapan, kecepatan angin, curah hujan, insolat, dan tekanan udara. Target yang akan diprediksi adalah kelompok awan yang juga telah dikelompokkan dalam kode kategori.

3 Pembagian Data (Train-Test Split)
Untuk memastikan model dapat menggeneralisasi dengan baik, data dibagi menjadi dua bagian: data pelatihan dan data pengujian. Proses pembagian ini dilakukan dengan proporsi 70% untuk data pelatihan dan 30% untuk data pengujian. Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk menguji akurasi model yang telah dilatih.

4 Pembuatan Model
Beberapa model klasifikasi yang berbeda diterapkan untuk memprediksi target. Model yang digunakan di antaranya adalah regresi logistik, pohon keputusan, *gradient boosting* (dengan metode AdaBoost), *simple Linear Regression* dan naïve bayes dan *Naive bayes*. Setiap model dilatih menggunakan data pelatihan, dengan tujuan untuk membandingkan performa masing-masing model.

5 Prediksi dan Perhitungan Akurasi
Setelah melatih model, langkah selanjutnya adalah memprediksi target pada data pengujian. Hasil prediksi kemudian dibandingkan dengan nilai sebenarnya

pada data pengujian untuk menghitung akurasi. Akurasi dihitung dengan cara membandingkan jumlah prediksi yang benar dengan total data pengujian.

6 Evaluasi Hasil
Hasil evaluasi dari setiap model, berupa nilai akurasi, dicatat dan disusun dalam tabel. Tabel ini memuat nama model dan nilai akurasi yang diperoleh masing-masing model pada data pengujian.

7 Visualisasi Hasil Evaluasi
Untuk memudahkan perbandingan antar model, hasil evaluasi ditampilkan dalam bentuk grafik batang. Grafik ini menunjukkan perbandingan akurasi antara model-model yang diuji, sehingga memudahkan pemilihan model dengan performa terbaik.

2.1.6 Confusion Matrix

1. Prediksi dengan Data Pengujian

Pada tahap ini, setelah model dilatih dengan data pelatihan, kita akan melakukan prediksi terhadap data pengujian. Data pengujian ini belum pernah dilihat oleh model sebelumnya, sehingga prediksi yang dihasilkan oleh model akan menunjukkan seberapa baik model dapat mengeneralisasi terhadap data baru.

- Prediksi dilakukan dengan menggunakan fungsi predict() dari model yang telah dilatih.
- Model akan mengeluarkan hasil prediksi yang kemudian dibandingkan dengan nilai target yang sebenarnya (y_test).

2. Penghitungan *Confusion matrix*

Setelah kita mendapatkan hasil prediksi (y_pred), langkah selanjutnya adalah menghitung *Confusion matrix*. *Confusion matrix* adalah alat untuk mengevaluasi hasil klasifikasi dengan membandingkan antara prediksi dan nilai sebenarnya dari data uji. Matriks ini menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas.

Confusion matrix memberikan empat nilai utama:

- *True Positives* (TP): Prediksi yang benar-benar positif (model benar memprediksi kelas positif).
- *False Positives* (FP): Prediksi yang salah positif (model memprediksi kelas positif padahal seharusnya negatif).
- *True Negatives* (TN): Prediksi yang benar-benar negatif (model benar memprediksi kelas negatif).
- *False Negatives* (FN): Prediksi yang salah negatif (model memprediksi kelas negatif padahal seharusnya positif).

Matriks ini memberikan gambaran yang lebih detail tentang kinerja model dalam mengklasifikasikan setiap kelas.

3. Visualisasi *Confusion matrix*

Confusion matrix dapat divisualisasikan untuk mempermudah interpretasi. Dengan menggunakan pustaka *matplotlib* dan *seaborn*, kita dapat menampilkan *confusion matrix* dalam bentuk *heatmap*. *Heatmap* akan menampilkan setiap elemen dalam matriks dengan skala warna yang menunjukkan besar angka pada setiap elemen, sehingga memudahkan identifikasi kesalahan klasifikasi.

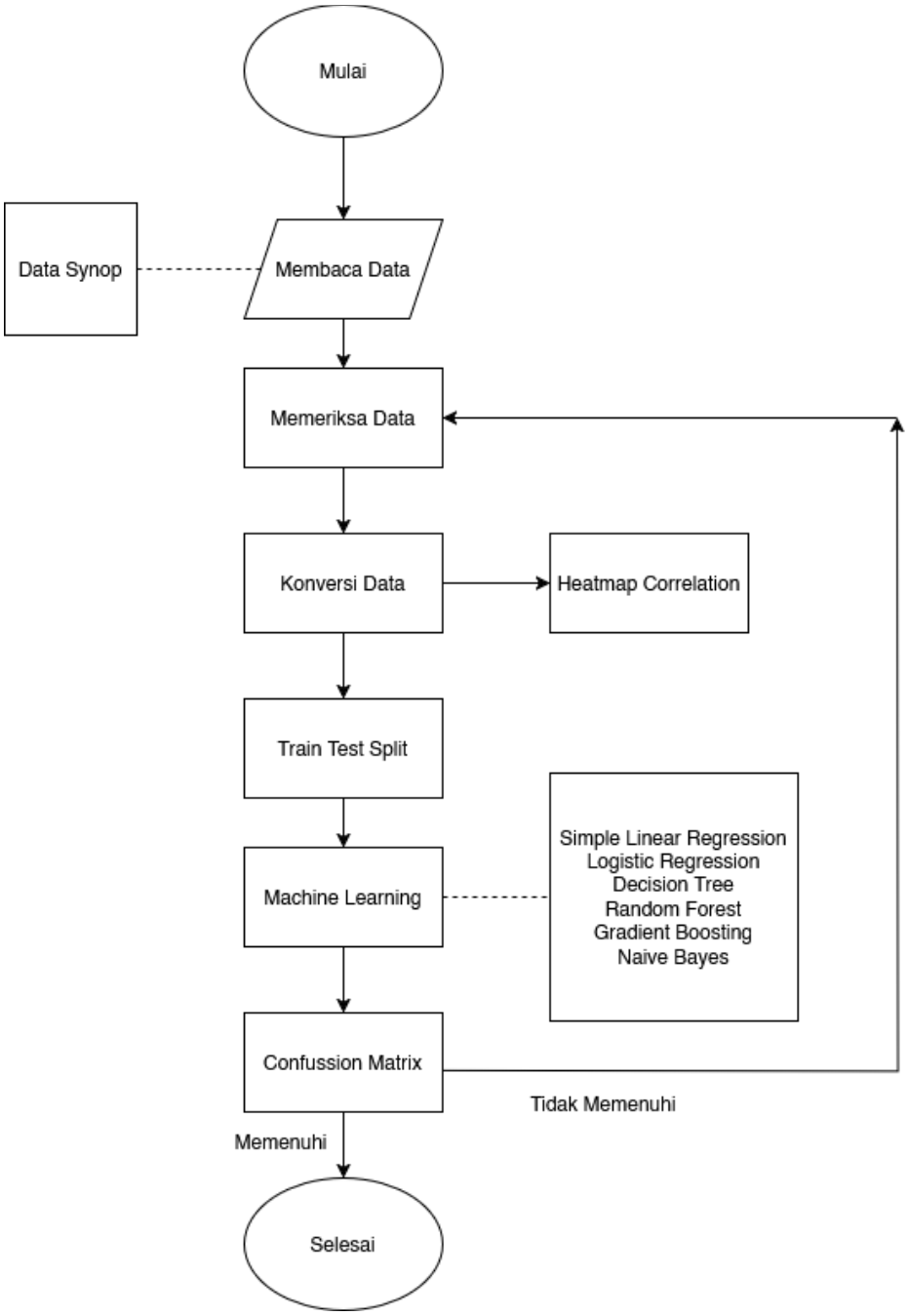
4. Evaluasi Model

Evaluasi dilakukan dengan mengukur **akurasi** dari model yang dihasilkan. Akurasi dihitung sebagai persentase dari prediksi yang benar dibandingkan dengan jumlah total data. Namun, akurasi saja tidak selalu cukup, terutama dalam kasus data yang tidak seimbang (misalnya, satu kelas lebih dominan). *Confusion matrix* memungkinkan kita untuk lebih memahami bagaimana model menangani setiap kelas.

5. Menampilkan *Confusion matrix* untuk Setiap Model

Setelah menghitung dan menampilkan *Confusion matrix*, kita dapat mengamati bagaimana setiap model berperforma dalam hal kesalahan klasifikasi. Hal ini berguna untuk membandingkan antara model yang berbeda dan menentukan model mana yang memiliki kinerja terbaik dalam mengklasifikasikan data.

2.2 Bagan Alir



Gambar 4. Bagan Alir