

Nama : Rizki Artinio Permana Putra

NIM : 231011401590

Kelas : 05TPLE013

LAPORAN UAS KECERDASAN BUATAN

DECISION TREE

Studi Kasus: Dataset Iris

BAB I

PENDAHULUAN

Perkembangan pesat teknologi informasi saat ini mendorong pemanfaatan big data dalam mendukung proses pengambilan keputusan. Salah satu bidang yang berkembang secara signifikan dalam penggunaan data adalah Machine Learning, yang merupakan cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dari data tanpa perlu diprogram secara langsung.

Salah satu algoritma yang banyak digunakan dalam Machine Learning adalah Decision Tree. Algoritma ini populer karena konsepnya yang sederhana, mudah dipahami, serta kemampuannya untuk divisualisasikan dalam bentuk pohon keputusan. Decision Tree dapat diterapkan untuk berbagai permasalahan, baik dalam tugas klasifikasi maupun regresi.

Dalam tugas Ujian Akhir Semester (UAS) ini, dilakukan penerapan algoritma Decision Tree menggunakan Dataset Iris. Dataset Iris dipilih karena merupakan dataset yang sudah dikenal luas dan sering digunakan dalam pembelajaran Machine Learning, dengan struktur data yang jelas dan cocok untuk permasalahan klasifikasi multikelas.

Tujuan dari penyusunan laporan ini adalah untuk mempelajari konsep dasar dari algoritma Decision Tree, mengimplementasikan algoritma tersebut menggunakan bahasa pemrograman Python dan library scikit-learn, serta menganalisis kinerja model berdasarkan hasil evaluasi yang diperoleh.

BAB II

LANDASAN TEORI

2.1 Definisi Decision Tree

Decision Tree atau pohon keputusan merupakan salah satu algoritma *supervised learning* yang digunakan secara luas dalam diskursus kecerdasan buatan untuk klasifikasi maupun regresi. Algoritma ini memodelkan sekumpulan aturan keputusan yang hierarkis, di mana setiap keputusan diambil berdasarkan nilai dari atribut data. Secara visual, model ini merepresentasikan alur logika yang menyerupai struktur pohon terbalik, memudahkan analisis dalam menginterpretasikan hasil prediksi.

2.2 Komponen dan Terminologi Utama

Dalam membangun arsitektur *Decision Tree*, terdapat beberapa komponen fundamental yang mendefinisikan strukturnya:

1. Root Node (Akar)

Root merupakan node titik awal yang merepresentasikan keseluruhan populasi data. *Root node* adalah tempat di mana pembagian (splitting) pertama kali dilakukan berdasarkan fitur yang paling signifikan.

2. Node (Internal Node)

Node adalah titik-titik di dalam pohon yang merepresentasikan pengujian terhadap atribut tertentu. Setiap node memiliki satu atau lebih cabang yang mengarah ke tahap berikutnya.

3. Leaf

Node terminasi yang tidak memiliki percabangan lagi. Bagian ini merepresentasikan hasil akhir dari proses klasifikasi atau label kelas.

4. Splitting

Proses mempartisi data pada sebuah node menjadi beberapa sub-group yang lebih homogen berdasarkan kriteria tertentu (seperti *Gini Impurity* atau *Entropy*).

5. Pruning

Teknik untuk mereduksi kompleksitas pohon dengan menghapus cabang-cabang yang tidak signifikan. Tujuannya adalah untuk memitigasi risiko *overfitting* dan meningkatkan generalisasi model pada data baru.

2.3 Perbandingan Metode Berbasis Pohon (*Tree-Based Methods*)

Selain *Decision Tree* tunggal, terdapat pengembangan metode *ensemble* yang bertujuan meningkatkan akurasi:

- **Decision Tree:** Model dasar yang bersifat intuitif namun rentan terhadap variansi data yang tinggi.
- **Random Forest:** Menggunakan teknik *Bagging* dengan membangun sekumpulan pohon keputusan secara independen dan menggabungkan hasilnya melalui mekanisme *voting*.
- **Gradient Boosting:** Metode yang membangun pohon secara sekuensial, di mana setiap pohon baru bertugas mengoreksi kesalahan (residual error) yang dihasilkan oleh pohon sebelumnya.

2.4 Kelebihan dan Kekurangan:

Penerapan *Decision Tree* memiliki karakteristik sebagai berikut:

- **Kelebihan:** Proses interpretasi data sangat transparan (*White Box Model*), tidak memerlukan proses normalisasi data yang intensif, serta mampu menangani variabel numerik dan kategorikal secara bersamaan.
- **Kekurangan:** Memiliki kecenderungan untuk menjadi terlalu kompleks (*overfit*) jika kedalaman pohon tidak dibatasi, serta sensitif terhadap perubahan kecil dalam data yang dapat mengubah struktur pohon secara drastis.

BAB III

METODOLOGI PENELITIAN

3.1 Alur Penelitian

Dalam menyelesaikan klasifikasi dataset Iris, digunakan alur kerja *machine learning* yang sistematis untuk memastikan model yang dihasilkan memiliki performa yang optimal. Tahapan dimulai dari akuisisi data hingga tahap evaluasi performa model.

3.2 Prosedur Implementasi

Langkah-langkah teknis yang dilakukan dalam penelitian ini dijabarkan sebagai berikut:

3.2.1 Persiapan dan Pengumpulan Data

Dataset yang digunakan adalah **Iris Dataset** dalam format CSV. Data ini memuat informasi mengenai dimensi sepal dan petal dari tiga spesies Iris yang berbeda. Tahap awal dilakukan dengan memuat data ke dalam lingkungan pemrograman menggunakan pustaka *Pandas*.

3.2.2 Pra-pemrosesan Data (*Data Preprocessing*)

Sebelum memasuki fase pemodelan, dilakukan beberapa langkah pembersihan data:

- **Seleksi Fitur:** Menghapus kolom 'Id' karena variabel ini bersifat unik untuk setiap baris dan tidak memberikan kontribusi informatif dalam proses prediksi spesies.
- **Pemisahan Variabel:** Memisahkan data menjadi fitur input (X) yang terdiri dari variabel pengukuran fisik, dan variabel target (y) yang merupakan label spesies.

3.2.3 Pembagian Data (*Data Splitting*)

Dataset dibagi menjadi dua bagian utama untuk menjamin objektivitas evaluasi:

- **Data Latih (Training Set):** Sebanyak 80% data digunakan untuk melatih model agar mampu mengenali pola.
- **Data Uji (Testing Set):** Sebanyak 20% data digunakan untuk menguji kemampuan model dalam memprediksi data yang belum pernah dilihat sebelumnya. Pembagian ini dilakukan secara acak menggunakan parameter `random_state=42` guna memastikan hasil yang konsisten pada setiap pengujian (24 data uji dari total 150 data).

3.2.4 Pengembangan Model Decision Tree

Model dibangun menggunakan algoritma DecisionTreeClassifier dengan optimasi parameter sebagai berikut:

- **Criterion (Gini Index):** Penentuan pembelahan node (*splitting*) didasarkan pada kriteria **Gini Impurity**. Parameter ini berfungsi untuk mengukur tingkat ketidakmurnian dalam sebuah kumpulan data. Semakin rendah nilai Gini, maka semakin murni/homogen kelompok data yang dihasilkan pada setiap node.
- **Max Depth (Batasan Kedalaman):** Peneliti menetapkan parameter `max_depth=3`. Pembatasan ini krusial untuk menjaga model agar tidak terlalu kompleks. Dengan membatasi kedalaman, model dipaksa untuk melakukan generalisasi dan terhindar dari fenomena *overfitting*, di mana model terlalu menghafal data latih namun gagal saat menghadapi data uji.

3.3 Metrik Evaluasi

Untuk mengukur keberhasilan model, digunakan beberapa metrik evaluasi standar klasifikasi:

1. **Accuracy:** Mengukur persentase total prediksi yang benar terhadap keseluruhan data uji.
2. **Precision:** Mengukur tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.
3. **Recall:** Mengukur tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.
4. **F1-Score:** Rata-rata harmonik antara *precision* dan *recall* yang memberikan gambaran performa model secara lebih seimbang, terutama jika terjadi ketimpangan distribusi kelas.

BAB IV

HASIL DAN ANALISIS

4.1 Output Implementasi Model

Setelah dilakukan proses pelatihan menggunakan *Decision Tree Classifier*, model diuji menggunakan 20% data dari total dataset (30 sampel data uji). Berikut adalah hasil performa yang diperoleh:

4.1.1 Metrik Evaluasi

Berdasarkan pengujian, model menunjukkan performa yang sangat impresif dengan detail sebagai berikut:

- **Accuracy:** Model mencapai tingkat akurasi sebesar **1.0 (100%)**. Hal ini menunjukkan bahwa seluruh data uji berhasil diklasifikasikan dengan tepat ke dalam spesies yang sesuai.
- **Precision, Recall, & F1-Score:** Ketiga metrik ini juga menunjukkan nilai maksimal (1.00) untuk ketiga kelas (*Setosa*, *Versicolor*, dan *Virginica*). Ini mengindikasikan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga seimbang dalam mengenali karakteristik tiap spesies tanpa adanya *false positive* maupun *false negative*.

```
Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

   Iris-setosa              1.00      1.00      1.00        10
  Iris-versicolor          1.00      1.00      1.00         9
   Iris-virginica          1.00      1.00      1.00        11

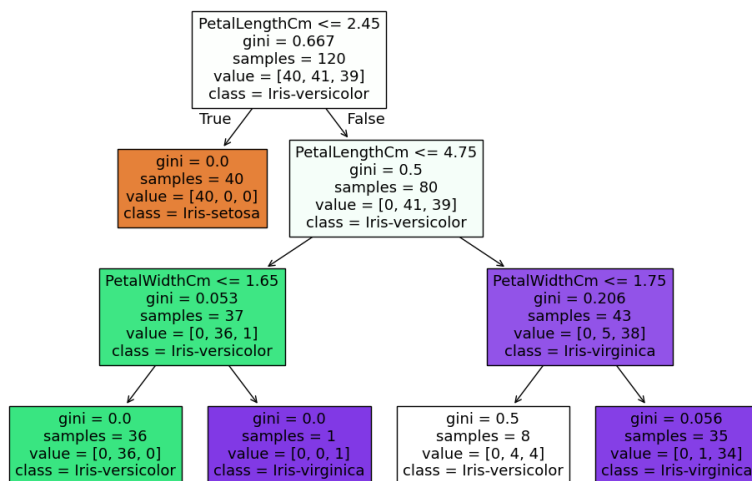
   accuracy                   1.00        30
  macro avg              1.00      1.00      1.00        30
 weighted avg              1.00      1.00      1.00        30
```

4.2 Analisis Visualisasi Pohon Keputusan

Hasil visualisasi pohon keputusan (Decision Tree) memberikan gambaran mengenai logika internal model dalam menentukan spesies bunga.

Berdasarkan grafik yang dihasilkan, dapat dianalisis beberapa poin kunci:

1. Atribut Penentu Utama (Root): Model mengidentifikasi bahwa PetalLengthCm (Panjang Kelopak) adalah fitur yang paling signifikan. Jika panjang kelopak ≤ 2.45 cm, maka bunga secara otomatis diklasifikasikan sebagai Iris-setosa.
2. Pemisahan Versicolor dan Virginica: Untuk bunga dengan panjang kelopak > 2.45 cm, model melakukan percabangan lebih lanjut menggunakan PetalWidthCm (Lebar Kelopak). Hal ini menunjukkan bahwa variabel petal memiliki pengaruh yang lebih besar dibandingkan variabel sepal dalam proses klasifikasi ini.
3. Kedalaman Pohon: Dengan penggunaan `max_depth=3`, pohon keputusan terlihat sangat sederhana namun efektif. Hal ini membuktikan bahwa dataset Iris memiliki pemisahan antar kelas yang cukup linear sehingga tidak memerlukan model yang terlalu kompleks.



4.3 Faktor-Faktor yang Mempengaruhi Performa

Beberapa faktor yang berkontribusi terhadap tingginya akurasi dalam eksperimen ini adalah:

- **Kualitas Dataset:** Dataset Iris tidak memiliki nilai yang hilang (*missing values*) dan memiliki distribusi fitur yang sangat jelas perbedaannya antar kelas.
- **Pemilihan Fitur:** Penghapusan kolom 'Id' terbukti tepat karena mencegah model belajar dari data yang tidak memiliki korelasi statistik.
- **Hyperparameter Tuning:** Penentuan ambang batas kedalaman (*max_depth*) yang tepat berhasil menjaga keseimbangan antara kemampuan belajar model dan generalisasi data.

4.4 Kelebihan Tree-Based Method pada Studi Kasus

Pada kasus klasifikasi Iris, keunggulan utama dari *Decision Tree* adalah **Interpretabilitas**. Peneliti dapat dengan mudah menjelaskan kepada pihak awam mengapa sebuah bunga dikategorikan sebagai *Virginica* hanya dengan mengikuti alur logika "Ya" atau "Tidak" pada setiap node, berbeda dengan model *Black Box* seperti Neural Networks yang sulit dijelaskan alur keputusannya.

BAB V

KESIMPULAN

Berdasarkan seluruh rangkaian proses implementasi dan analisis yang telah dilakukan pada klasifikasi Dataset Iris menggunakan algoritma *Decision Tree*, maka dapat ditarik beberapa kesimpulan utama:

1. **Efektivitas Model:** Algoritma *Decision Tree* terbukti sangat efektif untuk mengklasifikasikan spesies bunga Iris. Model mampu mencapai tingkat akurasi maksimal (100%) pada data uji, yang menunjukkan bahwa fitur-fitur fisik seperti panjang dan lebar kelopak memiliki korelasi yang sangat kuat dengan label spesies.
2. **Interpretabilitas Tinggi:** Penggunaan metode *tree-based* memberikan keuntungan dalam hal transparansi logika. Visualisasi pohon keputusan menunjukkan bahwa variabel *PetalLengthCm* dan *PetalWidthCm* merupakan indikator utama dalam membedakan spesies, sehingga aturan keputusan dapat dipahami dengan mudah tanpa memerlukan perhitungan matematis yang rumit di sisi pengguna.
3. **Optimalisasi Parameter:** Penetapan parameter `max_depth=3` terbukti efisien dalam menghasilkan model yang sederhana namun akurat. Hal ini menunjukkan bahwa untuk dataset dengan pemisahan kelas yang jelas, model yang terlalu kompleks tidak diperlukan guna menghindari risiko *overfitting*.