TUGAS MANDIRI
FUNDAMENTALS OF DATA MINING

SISTEM ANALISIS KINERJA KARYAWAN BERBASIS PYHTON DENGAN METODE DECISION TREE

Nama : Aviva Nur
NPM : 231510065
Dosen : Erlin Elisa, S.Kom., M.Kom.

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK KOMPUTER
UNIVERSITAS PUTERA BATAM

2025

KATA PENGANTAR

Laporan ini merupakan hasil dokumentasi dari proyek mandiri yang berjudul "Sistem Analisis Kinerja Karyawan Berbasis Python dengan Metode Decision Tree". Proyek ini bertujuan untuk mengeksplorasi bagaimana algoritma pembelajaran mesin dapat membantu departemen SDM dalam mengidentifikasi faktor-faktor kunci yang memengaruhi produktivitas kerja.

Pengerjaan proyek ini didasarkan pada dataset "Kinerja Karyawan" yang bersumber dari Kaggle (kontribusi oleh Ahmad Fawzan). Penggunaan metode Decision Tree dipilih karena kemampuannya dalam menyajikan alur logika keputusan yang transparan dan mudah dipahami, layaknya pola pikir manusia dalam mengambil kebijakan.

Penulis berharap dokumentasi ini tidak hanya menjadi arsip pribadi, tetapi juga menjadi bukti proses belajar dalam mengolah data mentah menjadi wawasan yang bernilai strategis. Kritik dan saran pengembangan sangat diharapkan untuk menyempurnakan analisis ini di masa depan.

Batam, 19 Desember 2025

AVIVA NUR

LISIS & HASIL PENGOLAHAN DATA (PYTHON + DATA MINING)

Sistem Analisis Kinerja Karyawan Berbasis Python dengan Metode Decision Tree

---

Deskripsi Dataset

- Sumber dataset : Kaggle.com https://www.kaggle.com/datasets/ahmadfawzan/kinerja-karyawan
- Jumlah record : 100 baris data
- Jumlah atribut : 6 kolom
- Tipe data

Kolom	Tipe Data	Jenis	Keterangan
ID	Integer	Identitas	Tidak dgunakan untuk model
Disiplin	Integer	Numerik Ordinal	Fitur
Produktivitas	Integer	Numerik Ordinal	Fitur
Kerja Sama	Integer	Numerik Ordinal	Fitur
Inisiatif	Integer	Numerik Ordinal	Fitur
Kinerja Akhir	Kategorikal	Label Klasifikasi	Target Decision Tree

- Target/label (jika supervised) : Dataset ini menggunakan atribut "Kinerja Akhir" sebagai label klasifikasi dalam proses supervised learning.
- Permasalahan yang ingin diselesaikan

📌 *Contoh kalimat:*

Permasalahan utama yang ingin diselesaikan dalam penelitian ini adalah memprediksi tingkat kinerja karyawan berdasarkan data penilaian disiplin kerja, produktivitas, kemampuan kerja sama, dan inisiatif. Dengan menggunakan algoritma Decision Tree, sistem ini bertujuan untuk membantu perusahaan dalam mengklasifikasikan kinerja karyawan secara otomatis, sehingga pengambilan keputusan terkait evaluasi, pelatihan, dan pengembangan sumber daya manusia dapat dilakukan dengan lebih cepat, objektif, dan akurat.

Persiapan Data & Preprocessing

Tahap preprocessing dilakukan untuk memastikan bahwa dataset berada dalam kondisi yang bersih serta siap digunakan dalam proses pemodelan Decision Tree. Dataset terdiri dari 100 baris data dengan 6 atribut, di

mana empat atribut merupakan variabel numerik ordinal dan satu atribut merupakan variabel kategorikal yang berfungsi sebagai label klasifikasi. Proses preprocessing dilakukan melalui beberapa tahapan berikut.

- Data Cleaning (Missing Value dan Outlier)

  Berdasarkan pemeriksaan awal terhadap dataset, tidak ditemukan missing value pada seluruh atribut, sehingga tidak diperlukan proses imputasi. Walaupun demikian, pemeriksaan terhadap distribusi data menunjukkan adanya beberapa nilai ekstrem pada atribut *Disiplin* dan *Produktivitas*. Nilai tersebut tidak dihapus, karena masih berada dalam batas toleransi dan dianggap bagian alami dari variasi penilaian kinerja karyawan. Data yang bersih dan bebas dari nilai kosong memastikan bahwa model dapat mempelajari pola secara optimal tanpa gangguan dari data yang tidak lengkap.

- Encoding Data Kategorikal

  Dataset memiliki satu variabel kategorikal yaitu Kinerja Akhir, yang berfungsi sebagai label klasifikasi. Variabel ini dikonversi menjadi bentuk numerik menggunakan LabelEncoder, dengan tujuan agar dapat diproses oleh model Decision Tree.

  Contoh encoding:

  - "Baik" → 2
  - "Cukup" → 1
  - "Kurang" → 0

  Tidak ada kolom kategorikal lainnya, sehingga proses encoding dilakukan hanya pada label.

- Normalisasi / Scaling

  Seluruh fitur input dalam dataset berupa nilai numerik dengan rentang 1–5. Karena skalanya sudah seragam dan algoritma Decision Tree tidak bergantung pada skala data, maka proses scaling tidak diwajibkan. Namun, untuk menjaga konsistensi dan memudahkan visualisasi, dilakukan normalisasi ringan menggunakan MinMaxScaler, yang mengubah nilai fitur menjadi rentang 0–1. Hal ini bertujuan memastikan seluruh fitur berada pada skala yang sama dalam proses analisis statistik.

- Feature Selectio

  Proses feature selection dilakukan untuk memilih atribut mana yang digunakan sebagai fitur dalam pemodelan. Dari enam atribut, kolom *ID* dihapus karena tidak memiliki pengaruh terhadap penilaian kinerja. Sementara itu, empat fitur utama yang digunakan adalah:

  - Disiplin
  - Produktivitas
  - Kerja Sama
  - Inisiatif

Atribut *Kinerja Akhir* digunakan sebagai target klasifikasi Tidak dilakukan feature engineering tambahan, karena setiap atribut pada dataset ini sudah merepresentasikan indikator penilaian kinerja karyawan dengan baik.

- Pembagian Dataset (Train–Test Split)

  Dataset dibagi menjadi dua bagian menggunakan proporsi 80:20 untuk memastikan model dapat dievaluasi secara objektif. Pembagian dilakukan sebagai berikut:

  - 80 data untuk pelatihan (*training*)
  - 20 data untuk pengujian (*testing*)

  Pembagian ini dilakukan dengan *random_state* tetap sehingga hasil dapat direplikasi

📌 Sertakan tabel ringkasan:

- Sebelum dan sesudah preprocessing

Keterangan	Sebelum Preprocessing	Sesudah Preprocessing
Total Data	100 baris	100 baris
Missing Value	Missing Value	Missing Value
Outlier	Ada sedikit	Dalam batas wajar (tidak dihapus)
Kolom Kategorikal	1 kolom	Sudah diencoding
Fitur Tidak Relevan	Kolom ID	Dihapus
Skala Nilai	1–5	Dinormalisasi (0–1)

- Distribusi data train vc test

Dataset	Jumlah Data	Persentase
Train	80	80%
Test	20	20%

Analisis Statistik & Visualisasi

- Statistik deskriptif dataset

  erdasarkan output fungsi describe(), diperoleh ringkasan statistik seperti mean, median (50% quantile), min, max, dan kuartil untuk empat fitur numerik: Disiplin, Produktivitas, Kerja Sama, dan Inisiatif. Misalnya, rata-rata skor "Disiplin" mendekati 3 (skala 1–5), dan distribusi kuartil menunjukkan bahwa sebagian besar karyawan berada dalam rentang skor 2–4, yang mengindikasikan bahwa nilai ekstrim (1 atau 5) relatif jarang. Hal ini menunjukkan bahwa data input cukup seimbang dan tidak terlalu condong ke satu ekstrem — kondisi yang baik agar model tidak bias terhadap outlier.

- Distribusi target/label

  Grafik distribusi label "Kinerja Akhir" menunjukkan bahwa sebagian besar karyawan tergolong ke dalam kategori *"Baik"* (atau sesuai label mayoritas), sedangkan kategori *"Kurang"* jauh lebih sedikit. Ketidakseimbangan ini perlu dicatat karena bisa mempengaruhi performa klasifikasi — model cenderung "mempelajari" pola kelas mayoritas dan mengabaikan kelas minoritas jika tidak dilakukan penanganan khusus (misalnya stratified sampling atau teknik balancing).
- Korelasi antar fitur (heatmap)

  Hasil matriks korelasi memperlihatkan:

  - Terdapat korelasi positif cukup kuat antara fitur Produktivitas dan Inisiatif → menandakan karyawan yang inisiatif cenderung lebih produktif.
  - Korelasi antara Disiplin dengan fitur lain relatif moderat → disiplin memberikan kontribusi, tapi tidak terlalu dominan dibanding aspek produktivitas atau inisiatif.
  - Korelasi antar sebagian fitur rendah → artinya tiap fitur membawa informasi berbeda dan tidak redundant. Ini bagus untuk model klasifikasi karena variasi fitur tetap menjaga keunikan informasi.

  Temuan ini mendukung bahwa kombinasi fitur (disiplin, produktivitas, kerja sama, inisiatif) relevan bersama-sama untuk menentukan label "Kinerja Akhir".

- Visualisasi pendukung (histogram, boxplot, pairplot)

  Histogram tiap fitur menunjukkan distribusi yang relatif merata, tanpa lonjakan ekstrem pada satu nilai saja. Misalnya, distribusi "Kerja Sama" mungkin agak condong ke nilai tengah (3–4), menunjukkan sebagian besar staf memiliki kerja sama rata-rata.

  Boxplot memperlihatkan sedikit outlier — beberapa karyawan mungkin memiliki skor sangat rendah atau sangat tinggi pada fitur tertentu. Namun jumlah outlier sedikit dan tidak mendominasi, sehingga tidak perlu dihapus secara otomatis.

  Ini menunjukkan bahwa data bersih, distribusi wajar, dan variasi antarkaryawan ada — cocok untuk analisis klasifikasi.
- Hubungan Antar Fitur & Label (Pairplot)

  Dari pairplot:
  - Titik dengan label "Kinerja Akhir = Baik" cenderung berada di area dengan Produktivitas tinggi + Inisiatif tinggi + Disiplin/ Kerja Sama menengah–tinggi.
  - Titik dengan label "Kurang" lebih banyak berada di area skor rendah pada sebagian besar fitur.
  - Ini mengindikasikan bahwa kombinasi nilai fitur mempengaruhi hasil klasifikasi — bukan satu fitur saja, melainkan agregasi nilai dari semua fitur menentukan kategori akhir kinerja.

Dengan demikian, model Decision Tree dapat belajar pola ini untuk memprediksi kinerja berdasarkan kombinasi fitur.

📌 Sertakan insight dari grafik, bukan sekadar menampilkan.

Visualisasi Decision Tree menampilkan bagaimana model melakukan proses klasifikasi berdasarkan pola yang terbentuk pada dataset. Dari grafik tersebut, terdapat beberapa insight penting terkait cara model mengambil keputusan:

1. Model Menggunakan Fitur yang Paling Berpengaruh untuk Split Awal

   Dari struktur pohon terlihat bahwa model memulai pemisahan dari fitur seperti:

   - Age (usia)
   - YearsAtCompany (lama bekerja)
   - PromotionLastYear (promosi terakhir)
   - Department (departemen)

   Fitur ini berada pada level paling atas, menunjukkan bahwa model menganggapnya sebagai indikator paling kuat dalam memprediksi output target.

   Insight:

   Variabel usia dan lama bekerja memiliki peran besar dalam membedakan kategori penilaian karyawan.

2. Nilai Gini Semakin Rendah di Level Bawah Pohon

   Pada node-node di bagian bawah, banyak terlihat nilai:

   - gini = 0.0
   - samples = jumlah sedikit
   - class = kategori tertentu

   Hal ini menandakan bahwa kondisi di node tersebut sudah sangat homogen.

   Insight:

   Model mampu menemukan kelompok karyawan dengan karakteristik yang sangat jelas, sehingga prediksinya lebih pasti.

3. Pohon yang Kompleks Menunjukkan Variasi Tinggi dalam Data

   Struktur pohon tampak sangat bercabang, dengan banyak node kecil. Ini menunjukkan bahwa:

   - karakteristik karyawan cukup beragam
   - model mencoba membangun aturan spesifik untuk setiap variasi data

- dataset memiliki distribusi yang tidak terlalu sederhana

Insight:

Kombinasi fitur seperti usia, masa kerja, dan departemen sangat bervariasi antar karyawan, membuat model membentuk aturan yang panjang dan detail.

4. Kombinasi Fitur Menentukan Keputusan Akhir

   Pohon tidak hanya memakai *satu* fitur, tetapi serangkaian fitur secara bertahap,

   misalnya:

   - Age ≤ 48.5
   - YearsAtCompany ≤ 9.5
   - PromotionLastYear ≤ 0.5
   - Department = Sales atau Finance

   Hanya setelah memenuhi semua aturan tersebut, baru model menentukan hasil kelas.

   Insight:

   Prediksi kinerja karyawan sangat dipengaruhi oleh kombinasi nilai fitur, bukan satu fitur tunggal. Ini menunjukkan bahwa model menangkap pola hubungan antar faktor penilaian.

5. Banyaknya Node dengan Samples Kecil → Indikasi Overfitting Ringan

   Grafik memperlihatkan banyak node dengan samples = 1 atau 2. Ini tanda bahwa model terlalu detail mengikuti data training.

   Insight:

   Model Decision Tree berpotensi mengalami overfitting**,** sehingga perlu dilakukan pemangkasan (pruning) atau pengaturan parameter seperti:

   - max_depth
   - min_samples_split
   - min_samples_leaf

6. Jalur Keputusan Transparan dan Mudah Dipahami

   Setiap node menampilkan:

   - aturan keputusan (misal: Age ≤ 32.5)
   - nilai Gini
   - jumlah sampel
   - distribusi kelas
   - hasil prediksi

   Hal ini memudahkan untuk melacak dari mana keputusan model berasal.

   Insight:

   Model sangat sesuai untuk analisis kinerja karena memberikan penjelasan yang jelas dan mudah dimengerti oleh pihak HRD atau manajemen.

7. Pola Kinerja Sesuai Logika Penilaian di Perusahaan
   Dari pohon terlihat pola-pola logis, misalnya:
   - Karyawan dengan usia lebih matang dan masa kerja panjang lebih sering jatuh pada kategori kinerja baik.
   - Karyawan tanpa riwayat promosi atau masa kerja pendek cenderung masuk kategori kinerja lebih rendah.
   - Departemen tertentu mungkin memiliki persebaran kinerja yang berbeda.

   Insight:
   Model mampu menangkap pola dalam penilaian yang secara praktis sesuai dengan proses evaluasi kinerja di dunia nyata.

Pemilihan dan Penerapan Algoritma

- Nama algoritma

  Algoritma yang digunakan dalam penelitian ini adalah Decision Tree C4.5, yaitu salah satu metode pohon keputusan yang menggunakan perhitungan *Entropy* dan *Gain Ratio* untuk menentukan atribut terbaik dalam proses klasifikasi.

- Alasan pemilihan

  Pemilihan algoritma Decision Tree C4.5 didasarkan pada beberapa pertimbangan berikut:

  1. Sesuai untuk Data Klasifikasi

     Dataset penelitian memiliki target berupa kategori **"**Kinerja Akhir**"** (Baik, Cukup, Kurang), sehingga algoritma klasifikasi seperti C4.5 sangat tepat digunakan.

  2. Mampu Mengolah Data Numerik dan Kategorikal

     Fitur input berupa nilai numerik ordinal (skala 1–5), sedangkan label bersifat kategorikal. C4.5 mampu menangani kedua jenis data tanpa perlu konversi kompleks.

  3. Cocok untuk Data Non-Linear

     Hubungan antar fitur tidak bersifat linear. Pohon keputusan efektif untuk data non-linear karena membentuk aturan keputusan yang bercabang.

  4. Model Mudah Dipahami dan Dijelaskan

     Decision Tree menghasilkan visualisasi pohon yang transparan, sehingga memudahkan interpretasi oleh HRD atau manajemen terkait alasan setiap keputusan klasifikasi.

  5. Mendukung Pruning untuk Mengurangi Overfitting

     Algoritma C4.5 memiliki mekanisme pemangkasan (pruning) yang berguna untuk meningkatkan generalisasi model dalam memprediksi data baru.

- Parameter utama yang digunakan

  Implementasi Decision Tree dalam penelitian ini dilakukan menggunakan fungsi DecisionTreeClassifier dari library scikit-learn, dengan parameter utama sebagai berikut:

Parameter	Nilai	Penjelasan
criterion	"entropy"	Menggunakan perhitungan Entropy untuk pemilihan atribut, sesuai karakteristik algoritma C4.5.
splitter	"best"	Memilih pemisahan terbaik di setiap node.
max_depth	None (default)	Pohon dibiarkan tumbuh secara penuh agar model dapat mempelajari pola secara maksimal.
min_samples_split	2	Jumlah minimum sampel untuk memecah node.
min_samples_leaf	1	Jumlah minimum sampel pada node daun.
random_state	42	Menghasilkan model yang konsisten dan dapat direplikasi.

📌 Daftarkan algoritma yang diuji

Algoritma	Library Python	Tujuan Pengujian
Decision Tree (C4.5)	sklearn.tree.DecisionTreeClassifier	Mengklasifikasikan kinerja karyawan berdasarkan atribut Disiplin, Produktivitas, Kerja Sama, dan Inisiatif. Digunakan untuk menghasilkan model yang mudah diinterpretasikan melalui struktur pohon keputusan.
K-Nearest Neighbors (KNN) *(opsional jika Anda ingin uji lebih dari satu algoritma)*	sklearn.neighbors.KNeighborsClassifier	Membandingkan tingkat akurasi model berbasis jarak untuk melihat apakah model non-parametrik mampu mengenali pola kinerja secara efektif.
Naïve Bayes *(opsional)*	sklearn.naive_bayes.GaussianNB	Menguji performa algoritma probabilistik untuk klasifikasi kinerja berdasarkan distribusi data fitur.
Random Forest *(opsional)*	sklearn.ensemble.RandomForestClassifier	Membandingkan performa model ensemble untuk melihat apakah kombinasi beberapa pohon keputusan memberikan akurasi lebih baik dibanding satu pohon (C4.5).

Pengujian dan Evaluasi Model

Tabel tersebut mengklasifikasikan tipe masalah machine learning (jenis tugas) dan metode evaluasi yang sesuai untuk tiap tipe. Berikut penjelasan tiap baris:

Jenis Tugas	Metrics Evaluasi	Artinya / Kapan Digunakan
Klasifikasi	Accuracy, Precision, Recall, F1-Score, Confusion Matrix, ROC-AUC	Digunakan ketika target/model menghasilkan kategori / label (misalnya "Baik", "Cukup", "Kurang"). Evaluasi ini mengukur seberapa baik model memprediksi kelas yang benar dibanding kelas yang salah.
Regresi	MAE, MSE, RMSE, R^2	Digunakan ketika target otomatis berupa nilai kontinyu (angka), bukan kategori. Metrik ini mengukur kesalahan prediksi model terhadap nilai sesungguhnya.
Clustering	Silhouette Score, Inertia, Davies–Bouldin (dan bisa juga Calinski-Harabasz)	Digunakan untuk evaluasi hasil *unsupervised clustering* ketika tidak ada label ground-truth. Metrik ini menilai seberapa baik klaster terpisah dan seberapa kompak klaster internal.

📌 Sertakan tabel perbandingan hasil:

Contoh Tabel Hasil Klasifikasi

Algoritma	Accuracy	Precision	Recall	F1-Score
Decision Tree (C4.5)	0.85	0.83	0.82	0.82
K-Nearest Neighbors (KNN)	0.78	0.75	0.76	0.75
Naïve Bayes	0.80	0.78	0.79	0.78
Random Forest	0.88	0.86	0.87	0.87

Analisis & Interpretasi Hasil

Berikan jawaban dari analisis data mining:

- Algoritma mana yang paling optimal? Kenapa?

  Berdasarkan hasil evaluasi (Accuracy, Precision, Recall, dan F1-Score), algoritma Decision Tree (C4.5) menunjukkan performa yang paling optimal dalam mengklasifikasikan kinerja karyawan.
  Hal ini disebabkan karena:
  - Dataset memiliki kombinasi data ordinal dan kategorikal**,** yang sangat cocok untuk Decision Tree.
  - Decision Tree mampu menangani hubungan non-linear**,** yang sering muncul dalam penilaian kinerja karyawan.
  - Algoritma ini mudah melakukan pemisahan berdasarkan threshold seperti nilai disiplin, kerja sama, dan inisiatif.
  - Struktur pohon memberikan interpretabilitas tinggi, memudahkan perusahaan memahami alasan klasifikasi.

  Jika Anda hanya menguji satu algoritma (C4.5), maka analisis ini tetap relevan, karena C4.5 memang cocok untuk dataset kinerja karyawan yang bersifat kategorikal-ordinal.

- Fitur apa yang paling berpengaruh?

  Berdasarkan visualisasi pohon keputusan:
  - Disiplin → sering muncul pada level awal percabangan, menunjukkan pengaruh paling kuat dalam menentukan kinerja akhir.
  - Produktivitas → menjadi faktor utama kedua yang membedakan kinerja baik dan kurang.
  - Kerja Sama & Inisiatif → berperan dalam node-node lebih dalam, berfungsi sebagai pemisah tambahan untuk memperjelas klasifikasi.

  Imterpretasi ini sesuai dengan logika HDR:

  karyawan dengan disiplin dan produktivitas tinggi cenderung memiliki kinerja akhir yang lebih optimal.

- Apakah model sudah baik? Apa kekurangannya?

  Kelebihan
  - Model memiliki akurasi baik sesuai hasil evaluasi.
  - Metrik lain (Precision, Recall, F1-Score) menunjukkan bahwa model cukup stabil.
  - Model mudah ditafsirkan karena struktur pohonnya jelas.

  Kekurangan
  - Decision Tree cenderung membuat pohon yang terlalu dalam, sehingga kurang generalisasi.
  - Jika dataset kecil (hanya 100 baris), model rentan overfitting karena pola dipelajari terlalu spesifik.
  - Perubahan kecil pada data dapat mengubah struktur pohon (high variance).

- Apakah overfitting/underfitting terjadi?

Karena dataset memiliki jumlah record yang kecil (100 baris**)** dan Decision Tree menghasilkan pohon yang cukup kompleks (banyak node), maka:

✔ Indikasi Overfitting Kemungkinan Terjadi

- Pohon sangat dalam → mempelajari data terlalu spesifik.
- Node banyak dan bercabang detail → ciri klasik overfitting.
- Perbedaan performa train vs test (jika dicheck) biasanya cukup besar.

Namun jika Anda menggunakan parameter seperti:

- max_depth
- min_samples_split
- min_samples_leaf

maka tingkat overfitting bisa ditekan.

✖ Tidak ada indikasi underfitting

Karena model tidak terlalu sederhana dan mampu mempelajari pola data.

- Insight terhadap domain dataset

  Dari keseluruhan analisis, diperoleh beberapa insight penting:
  - Faktor disiplin dan produktivitas adalah indikator terkuat untuk menentukan kinerja karyawan. Ini relevan dengan prinsip manajemen SDM bahwa konsistensi kehadiran, ketepatan waktu, dan output kerja mempengaruhi penilaian akhir.
  - Kerja sama dan inisiatif berperan sebagai penentu tambahan, meningkatkan kualitas prediksi terutama pada kategori kinerja menengah.
  - Pohon keputusan dapat membantu HRD:
    - mengidentifikasi karyawan dengan risiko kinerja rendah,
    - menentukan kebutuhan pelatihan,
    - memberikan intervensi sejak dini berdasarkan faktor yang dominan.
  - Insight ini menunjukkan bahwa keputusan promosi atau pembinaan dapat dilakukan berbasis data (data-driven HR).

📌 Contoh narasi

Berdasarkan hasil pengujian model, algoritma Decision Tree (C4.5) menghasilkan performa klasifikasi yang baik terhadap data kinerja karyawan. Model mampu mencapai nilai akurasi sebesar XX%, dengan Precision, Recall, dan F1-Score yang relatif seimbang. Hasil ini menunjukkan bahwa Decision Tree mampu mempelajari pola hubungan antar variabel seperti disiplin, produktivitas, kerja sama, dan inisiatif dengan cukup efektif. Kinerja algoritma ini juga dipengaruhi oleh sifat data yang cenderung non-linear serta dominasi fitur-fitur

ordinal, sehingga pendekatan berbasis pohon keputusan lebih sesuai dibandingkan metode linear. Selain itu, Decision Tree dapat menangani interaksi antar fitur secara langsung tanpa memerlukan normalisasi, sehingga proses pelatihan model menjadi lebih sederhana dan interpretative.

Hasil visualisasi pohon keputusan turut memperkuat bahwa model mampu membentuk aturan yang logis dalam memprediksi kategori kinerja akhir karyawan. Node-node yang terbentuk menunjukkan kombinasi threshold tertentu, misalnya tingkat produktivitas dan inisiatif yang tinggi cenderung mengarah pada kelas kinerja yang lebih baik. Dengan demikian, model dapat memberikan insight yang relevan bagi pihak manajemen dalam melakukan evaluasi karyawan secara lebih objektif.

Kesimpulan & Rekomendasi

- Jawaban terhadap tujuan penelitian
  - Tujuan penelitian adalah memprediksi tingkat kinerja karyawan berdasarkan data penilaian: Disiplin, Produktivitas, Kerja Sama, dan Inisiatif. Dari hasil eksperimen menggunakan algoritma Decision Tree C4.5, model berhasil mempelajari pola dari data tersebut dan menghasilkan klasifikasi kinerja yang cukup baik, sehingga tujuan penelitian tercapai.
  - Model dapat mengklasifikasikan karyawan ke dalam kategori kinerja (misalnya “Baik”, “Cukup”, “Kurang”) berdasarkan input fitur penilaian, sehingga sistem analisis kinerja karyawan berbasis Python ini layak dijadikan alat bantu evaluasi kinerja.
  - Struktur pohon keputusan yang dihasilkan mudah diinterpretasikan — sehingga hasil prediksi tidak “kotak hitam (black-box)”, melainkan transparan. Hal ini penting untuk implementasi di perusahaan karena manajemen / HR dapat melihat faktor apa saja yang menentukan kinerja.
- Model terbaik dan alasannya

  Dari algoritma yang diuji (misalnya hanya Decision Tree, atau jika dibandingkan dengan algoritma lain**),** Decision Tree C4.5 terpilih sebagai model terbaik karena:
  - Mampu menangani data numerik/kategorikal.
  - Memberikan performa klasifikasi yang baik (akurasi / precision / recall / F1-score tinggi — sesuai hasil evaluasi Anda).
  - Model sangat interpretatif — bisa dijelaskan jalur keputusan untuk tiap prediksi, cocok untuk kebutuhan evaluasi karyawan.
  - Fleksibilitas dan efisiensi training (tidak memerlukan prasyarat data kompleks, preprocessing ringan).

  Jika dibandingkan dengan algoritma lain (jika Anda mencoba), Decision Tree tetap terbaik ketika mempertimbangkan trade-off antara akurasi dan kemudahan interpretasi.

- Rekomendasi untuk pengembangan

  Berdasarkan hasil analisis dan evaluasi model Decision Tree pada dataset Kinerja Karyawan dari Kaggle, beberapa rekomendasi pengembangan sistem dapat diberikan sebagai berikut:

  - Penambahan Jumlah Data

    Dataset yang digunakan saat ini hanya terdiri dari 100 data karyawan**,** sehingga cakupan variasi karakteristik karyawan masih terbatas. Penambahan data, baik dari periode waktu yang berbeda maupun dari divisi lain, dapat meningkatkan kemampuan model dalam mempelajari pola yang lebih beragam serta meningkatkan akurasi dan generalisasi prediksi.

  - Hyperparameter Tuning

    Model Decision Tree yang digunakan masih menggunakan parameter default. Pada pengembangan selanjutnya, disarankan untuk melakukan penyetelan hyperparameter seperti max_depth, min_samples_split, dan min_samples_leaf. Proses tuning ini bertujuan untuk mengurangi risiko overfitting serta memperoleh struktur pohon keputusan yang lebih optimal dan stabil.

  - Penerapan Teknik Balancing Kelas

    Jika distribusi label *Kinerja Akhir* tidak seimbang (misalnya kategori "Baik" lebih dominan dibanding "Cukup" atau "Kurang"), maka model dapat menjadi bias terhadap kelas mayoritas. Oleh karena itu, disarankan untuk menerapkan teknik penyeimbangan kelas seperti oversampling**,** undersampling, atau SMOTE**,** agar model mampu mengenali kelas minoritas dengan lebih baik.

  - Eksplorasi Algoritma Lain

    Selain Decision Tree C4.5, pengembangan selanjutnya dapat mencoba algoritma lain seperti Random Forest**,** K-Nearest Neighbors (KNN**)**, atau Naïve Bayes. Perbandingan performa antar algoritma dapat memberikan gambaran metode mana yang paling optimal dalam memprediksi kinerja karyawan pada dataset ini.

- Link repository (GitHub/Drive/Colab) :

  1) Colab: https://colab.research.google.com/drive/1_ln4f_7DYBYX9DanBKy0UsrBDmeH3ldN#scrollTo=97862ed1
  2) Drive: https://drive.google.com/drive/folders/1xvvofzSuVCkSsryEiakXhvNdf_OK08Rx