LAPORAN TUGAS BESAR DATA MINING

"Klasifikasi dan Clustering"

Oleh:

Kelompok 3

Ahmad Syahir	(60900122024)
Nurul Ilmi	(60900122023)
Israyani	(60900122021)
Muh. Afdal Nas	(60900122027)

Dosen Pengampu :

Adhy Rizaldy, S.Kom., M.Kom

JURUSAN SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI ALAUDDIN MAKASSAR

2024/2025

BAB I
PENDAHULUAN

A. Latar Belakang

Data mining adalah proses eksplorasi dan analisis data dalam skala besar untuk menemukan pola, hubungan, atau informasi berharga yang sebelumnya tidak terlihat secara langsung. Metode ini mengintegrasikan teknik dari berbagai disiplin ilmu, termasuk kecerdasan buatan, statistik, dan database(Asfa, 2024).

Exploratory Data Analysis (EDA) merupakan langkah penting dalam proses analisis data yang bertujuan untuk memahami struktur, karakteristik, pola, dan anomali dalam dataset sebelum memulai analisis lanjutan atau penerapan model. EDA sering melibatkan penggunaan teknik statistik deskriptif dan visualisasi data untuk mendapatkan wawasan awal tanpa asumsi spesifik(Hidayat et al., 2023).

Klasifikasi adalah salah satu metode dalam pembelajaran mesin yang digunakan untuk memprediksi kategori atau kelas dari suatu data berdasarkan pola yang ditemukan pada data pelatihan. Proses klasifikasi terdiri dari dua tahap utama yaitu tahap pelatihan dan tahap pengujian.(Rabbani et al., 2023).

Clustering adalah teknik dalam data mining yang digunakan untuk mengelompokkan objek atau data ke dalam grup atau cluster berdasarkan kesamaan karakteristik tertentu, sehingga objek dalam satu cluster memiliki kemiripan yang lebih tinggi dibandingkan dengan objek di cluster lain. Metode ini sering digunakan dalam berbagai bidang, seperti analisis pasar, segmentasi pelanggan, dan pengelompokan dokumen(Ihsan Ahmad Fauzi & Raditya Danar Dana, 2023).

Langkah-langkah seperti Exploratory Data Analysis (EDA) sangat penting untuk memahami karakteristik dan struktur dataset sebelum analisis lebih lanjut. Selain itu, metode klasifikasi dan clustering memainkan peran krusial dalam membedakan dan mengelompokkan data

berdasarkan pola yang ditemukan, yang bermanfaat dalam berbagai aplikasi, termasuk analisis pasar dan segmentasi pelanggan. Dengan demikian, integrasi berbagai teknik dan metode ini memungkinkan peneliti dan profesional untuk mendapatkan wawasan yang lebih mendalam dari data yang dianalisis.

B. Tujuan

Berdasarkan latar belakang di atas , adapun tujuan dari analisis ini di antaranya :

1. Menganalisis Dataset

   Menganalisis dataset adalah proses sistematis untuk mengeksplorasi dan memahami data guna menemukan informasi berharga. Ini melibatkan pengumpulan data, pembersihan, eksplorasi, dan menerapkan analisis lanjutan. Tujuannya adalah untuk mengidentifikasi pola, tren, atau anomali yang dapat memberikan wawasan untuk pengambilan keputusan.

2. Preprocessing Data

   Preprocessing data adalah tahap persiapan yang sangat penting sebelum analisis. Ini mencakup berbagai langkah seperti:

   - **Pembersihan Data**: Menghapus atau memperbaiki data yang hilang, duplikat, atau tidak konsisten.
   - **Transformasi Data**: Mengubah format atau skala data, seperti normalisasi dan standardisasi.
   - **Pengkodean**: Mengubah kategori data menjadi format numerik agar dapat digunakan dalam model analisis.
   - **Pembagian Data**: Memisahkan dataset menjadi subset untuk pelatihan dan pengujian model.

3. Modeling

   Modeling adalah tahap di mana algoritma statistik atau machine learning diterapkan pada dataset untuk membangun model yang dapat membuat prediksi atau klasifikasi. Ini mencakup pemilihan

algoritma yang sesuai, pelatihan model dengan data pelatihan, dan penyesuaian parameter untuk meningkatkan akurasi. Contoh metode modeling termasuk regresi, pohon keputusan, dan clustering.

4. Evaluasi Model

   Evaluasi model adalah proses untuk menilai kinerja model yang telah dibangun. Ini dilakukan dengan menggunakan data pengujian dan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Tujuannya adalah untuk memastikan bahwa model dapat generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Evaluasi juga dapat mencakup validasi silang untuk mendapatkan gambaran yang lebih akurat tentang kinerja model. -

5. Memberikan Insight

   Memberikan insight adalah langkah yang melibatkan interpretasi hasil analisis dan modeling untuk menyimpulkan informasi yang berarti. Ini bertujuan untuk menjawab pertanyaan bisnis atau penelitian, serta mengidentifikasi tindakan yang dapat diambil. Insight ini dapat berupa laporan, visualisasi, atau rekomendasi strategis yang membantu dalam pengambilan keputusan yang lebih baik.

C. Dataset Description

1. Online Retailer.Csv

Adapun nama dataset yang kami gunakan yaitu "Online Retailer.Csv", dataset ini berisi informasi mengenai transaksi penjualan barang yang mencakup berbagai atribut terkait pelanggan dan produk. Dataset ini bertujuan untuk menganalisis pola pembelian dan kinerja penjualan. Dataset ini memiliki jumlah 8 kolom dan sejumlah record yang bervariasi. Adapun penjelasan dari masing-masing kolomnya:

a. InvoiceNo**:** Nomor identifikasi unik untuk setiap transaksi yang terjadi.
b. StockCode: Kode unik untuk setiap produk yang dijual, digunakan untuk mengidentifikasi produk dalam sistem.
c. Description: Deskripsi dari produk yang menjelaskan jenis barang yang dijual.

d. Quantity: Jumlah unit produk yang dibeli dalam satu transaksi.
e. InvoiceDate: Tanggal dan waktu saat transaksi dilakukan, dalam format YYYY-MM-DD HH:MM:SS.
f. UnitPrice: Harga per unit produk pada saat transaksi, menunjukkan biaya untuk satu unit.
g. CustomerID**:** ID unik untuk pelanggan yang melakukan pembelian, membantu melacak transaksi pelanggan.
h. Country**:** Negara tempat pelanggan berada, memberikan informasi tentang lokasi pelanggan untuk analisis pasar.

Dataset ini dapat digunakan untuk mengidentifikasi pola dalam pemilihan produk berdasarkan lokasi, waktu transaksi, dan jumlah pembelian, serta untuk menganalisis kinerja penjualan berdasarkan harga dan kuantitas.

2. Student Depression Dataset.csv

Dataset yang kami gunakan yaitu "Student Depression Dataset.csv". Dataset ini berisi informasi mengenai kondisi kesehatan mental mahasiswa, khususnya terkait depresi, dengan berbagai atribut yang mencakup data demografis dan faktor-faktor yang memengaruhi kesehatan mental. Dataset ini bertujuan untuk menganalisis prevalensi depresi di kalangan mahasiswa dan faktor-faktor yang berkontribusi terhadapnya. Dataset ini memiliki 17 kolom dan sejumlah record yang bervariasi. Adapun penjelasan dari masing-masing kolomnya:

a. id: Nomor identifikasi unik untuk setiap responden dalam dataset.
b. Gender: Jenis kelamin responden, yang dapat berupa 'Male' atau 'Female'.
c. Age: Usia responden dalam tahun.
d. City: Kota tempat tinggal responden, memberikan informasi tentang lokasi geografis.
e. Profession: Pekerjaan atau status pendidikan responden, umumnya 'Student'.

f. Academic Pressure: Tingkat tekanan akademis yang dirasakan responden, dinyatakan dalam skala tertentu.
g. Work Pressure: Tingkat tekanan kerja yang dialami responden, biasanya dinyatakan dalam skala.
h. CGPA: Cumulative Grade Point Average, yang menunjukkan prestasi akademis responden.
i. Study Satisfaction: Tingkat kepuasan responden terhadap metode belajar mereka, dinyatakan dalam skala.
j. Job Satisfaction: Tingkat kepuasan responden terhadap pekerjaan (jika ada), dinyatakan dalam skala.
k. Sleep Duration: Durasi tidur responden dalam jam per hari, memberikan gambaran tentang kebiasaan tidur.
l. Dietary Habits: Kebiasaan makan responden, yang dapat dikategorikan sebagai 'Healthy', 'Moderate', atau 'Unhealthy'.
m. Degree: Jenjang pendidikan yang sedang dijalani responden, seperti 'B.Sc', 'M.Tech', dll.
n. Have you ever had suicidal thoughts?: Informasi apakah responden pernah mengalami pikiran untuk bunuh diri, dijawab dengan 'Yes' atau 'No'.
o. Work/Study Hours: Jumlah jam yang dihabiskan untuk bekerja atau belajar, memberikan informasi tentang waktu yang diinvestasikan.
p. Financial Stress: Tingkat stres finansial yang dialami responden, biasanya dinyatakan dalam skala.
q. Family History of Mental Illness: Apakah ada riwayat penyakit mental dalam keluarga responden, dijawab dengan 'Yes' atau 'No'.
r. Depression: Status depresi responden, biasanya dinyatakan dalam bentuk angka (0 untuk tidak depresi, 1 untuk depresi).

Dataset ini dapat digunakan untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap kesehatan mental mahasiswa, serta untuk

menganalisis hubungan antara tekanan akademis, kebiasaan tidur, dan kepuasan hidup dengan tingkat depresi.

BAB II
METODOLOGI

A. Preprocesing

Preprocessing merupakan langkah pertama dalam pengolahan data yang bertujuan untuk mengubah data mentah menjadi format yang lebih bersih, terorganisir, dan siap digunakan dalam analisis atau model pembelajaran mesin. Proses ini sangat penting karena data mentah sering kali memiliki masalah seperti ketidaklengkapan, redundansi, atau inkonsistensi format.

Berikut Langkah – Langkah Preprocession yang dialami pada dataset Online Retailer :

- Menampilkan jumlah kolom dan data sebelum preprocessing
- Menangani nilai yang hilang melalui imputasi
- Mendeteksi dan menghapus data duplikat

B. Model Selection

Model yang digunakan pada dataset Health Insurance Lead Prediction Raw Data ada dua yaitu :

a. Klasifikasi

Klasifikasi adalah proses untuk mengidentifikasi kategori atau kelas dari data yang tidak terstruktur, dengan tujuan untuk memprediksi kelas dari data baru berdasarkan model yang telah dibangun.

Adapun algoritma yang di gunakan pada proses klasifikasi pada dataset ini adalah sebagai berikut :

1) Desicion Tree

Decision Tree adalah sebuah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Struktur utama dari decision tree adalah pohon yang terdiri dari simpul *(nodes)* dan cabang *(branches)*.

2) Logistic regression

Logistic Regression adalah metode statistik dan algoritma pembelajaran mesin yang digunakan untuk analisis dan prediksi variabel dependen biner, yaitu variabel yang hanya memiliki dua kategori atau hasil (misalnya, berhasil/gagal, ya/tidak, positif/negatif).

3) Random Forrest

Random Forest adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, yang terdiri dari banyak pohon keputusan (decision trees) yang bekerja sama dalam membuat prediksi.

b. Clustering

Clustering adalah metode untuk mengelompokkan data ke dalam kelompok (cluster) berdasarkan kesamaan di antara data. Tidak ada label yang ditentukan sebelumnya, sehingga digunakan untuk menemukan pola atau struktur tersembunyi dalam data.

Adapun algoritma yang di gunakan pada proses clustering pada dataset ini adalah sebagai berikut :

1) K-Means

K-means adalah algoritma pengelompokan yang mulai dengan k centroid acak. Titik data diassign ke kelompok terdekat berdasarkan jarak Euclidean, kemudian centroid diperbarui hingga posisi stabil.

2) DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

DBSCAN adalah algoritma pengelompokan yang mengidentifikasi kelompok berdasarkan kepadatan data, memisahkan area dengan kepadatan rendah sebagai noise. Algoritma ini tidak memerlukan jumlah kelompok yang ditentukan sebelumnya, sehingga fleksibel untuk berbagai bentuk dan ukuran kelompok.

1. Implementation Details

Detail implementasi untuk model-model di atas meliputi langkah-langkah berikut :

a. Data Prepocessing
    1) Membersihkan data (menghapus missing values, mengubah tipe data jika diperlukan).
    2) Melakukan normalisasi atau standarisasi data untuk algoritma yang membutuhkan skala seragam.
b. Model Training
    1) Klasifikasi
        - Memisahkan dataset menjadi data train dan test (misalnya, 80% untuk training dan 20% untuk testing).
        - Melatih model Decision Tree dan Random Forest menggunakan data training.
    2) Clustering :
        - Mengaplikasikan algoritma K-Means dengan memilih jumlah cluster (K) yang optimal.
        - Menerapkan DBSCAN untuk mengidentifikasi cluster berdasarkan kepadatan
c. Model evaluation
    1) Cross Validation untuk evaluasi model dalam pembelajaran mesin yang digunakan untuk mengukur bagaimana hasil analisis statistik.
    2) Untuk clustering, menggunakan silhouette score atau Davies-Bouldin index untuk menilai kualitas cluster.
d. Visualisasi

    Menggunakan scaterploot dan confusion matrix untuk memvisualisasikan hasil pengolahan data.

BAB III
HASIL DAN ANALISIS

A. EDA Findings

Dari analisis EDA yang dilakukan pada dataset depresi mahasiswa, beberapa temuan penting dapat diidentifikasi. Pertama, terdapat variasi yang signifikan dalam tekanan akademis dan pekerjaan di antara mahasiswa, dengan beberapa kelompok melaporkan tingkat tekanan yang lebih tinggi, yang dapat berkontribusi pada kondisi depresi.

Selanjutnya, analisis demografis menunjukkan perbedaan antara gender, di mana mahasiswa perempuan cenderung melaporkan lebih banyak gejala depresi dibandingkan mahasiswa laki-laki. Rentang usia juga menunjukkan hubungan dengan tingkat kepuasan belajar dan pekerjaan, di mana mahasiswa yang lebih muda cenderung memiliki kepuasan lebih rendah.

Selain itu, pola tidur dan kebiasaan diet mahasiswa berhubungan erat dengan kesehatan mental mereka; mahasiswa yang tidur kurang dari lima jam cenderung mengalami depresi lebih tinggi. Temuan ini menyoroti pentingnya faktor-faktor eksternal seperti tekanan akademis dan gaya hidup dalam mempengaruhi kesehatan mental mahasiswa, yang dapat menjadi dasar untuk intervensi yang lebih efektif dalam mendukung kesejahteraan mereka.

B. Model Performance

1. Klasifikasi
    a. Algoritma Decision Tree

        Insight dan Hasil

- Jumlah Data

  Dataset terdiri dari 18 kolom dan 27,901 baris, yang cukup untuk melatih model. Data kategori berhasil diubah menjadi numerik untuk kompatibilitas dengan algoritma machine learning.

- Proses Encoding

  Semua kolom kategori telah diubah menggunakan Label Encoding, memungkinkan model untuk memproses informasi dengan baik.

- Pembagian Data

  Dataset dibagi menjadi 80% untuk data pelatihan dan 20% untuk data pengujian. Ini memastikan model memiliki data yang cukup untuk belajar dan diuji secara adil.

- Pelatihan Decision Tree

  Model Decision Tree dilatih menggunakan kriteria Gini Impurity dengan kedalaman maksimum 5. Ini membantu model menangkap pola tanpa overfitting.

- Evaluasi Model

  Akurasi Model: Tingkat akurasi mencapai 81.26%, menunjukkan bahwa model cukup baik dalam memprediksi hasil berdasarkan data input.

- Laporan Klasifikasi: Model menunjukkan distribusi yang baik pada precision, recall, dan f1-score untuk setiap kelas, menunjukkan kinerja yang stabil dalam membedakan kelas.

- Visualisasi Decision Tree

- Grafik decision tree memberikan wawasan tentang logika pemisahan data. Fitur yang paling berpengaruh ditampilkan di puncak pohon, dan cabang menunjukkan aturan keputusan yang dibuat untuk memisahkan kelas.

Kesimpulan

- Kinerja Model: Decision Tree memberikan hasil akurasi yang baik pada dataset ini, mampu menangkap pola kompleks dalam data.
- Fitur Penting: Dari visualisasi, fitur yang sering muncul di node atas menjadi indikator paling signifikan untuk prediksi.

Analisis Hasil Evaluasi Model

- Akurasi

  Akurasi model mencapai 81.26%, menunjukkan bahwa model mampu memprediksi dengan benar untuk sebagian besar data dalam dataset pengujian.
- Interpretasi

  Hasil ini baik, menunjukkan kinerja yang memadai dalam prediksi.

Laporan Klasifikasi

- Precision

  Kelas 0 (2,343 sampel): Precision sebesar 0.76, artinya dari semua prediksi kelas 0, 76% benar-benar merupakan kelas 0.

  Kelas 1 (3,238 sampel): Precision sebesar 0.86, artinya 86% dari semua prediksi kelas 1 adalah benar.

Kesimpulan: Model efektif dalam meminimalkan kesalahan prediksi positif palsu.

- Recall

  Kelas 0: Recall sebesar 0.81, menunjukkan model mampu mengenali 81% sampel kelas 0.

  Kelas 1: Recall sebesar 0.82, artinya dari seluruh sampel kelas 1, 82% diidentifikasi dengan benar oleh model.

Kesimpulan: Model baik untuk kedua kelas, meskipun ada sedikit kekurangan dalam mengenali semua sampel.

- F1-score

  Kelas 0: 0.78, menunjukkan keseimbangan yang baik antara precision dan recall.

  Kelas 1: 0.83, juga menunjukkan kinerja yang baik meskipun sedikit lebih rendah dibandingkan kelas 0.

- Macro Avg

  Rata-rata precision, recall, dan f1-score di seluruh kelas menunjukkan kinerja konsisten di seluruh kelas. Hasilnya adalah precision 0.81, recall 0.81, dan f1-score 0.81, menunjukkan model bekerja secara baik secara keseluruhan.

b. Logistik Regression

Insight dan Hasil

- Jumlah Data

  Dataset terdiri dari 18 kolom dan 27,901 baris, yang cukup untuk melatih model. Data kategori telah berhasil diubah menjadi numerik agar dapat digunakan dalam algoritma machine learning.

- Proses Encoding

  Semua kolom kategori telah diubah menggunakan Label Encoding, memungkinkan model untuk memproses informasi ini secara efektif.

- Pembagian Data

  Dataset dibagi menjadi 80% untuk data pelatihan dan 20% untuk data pengujian. Ini memastikan model memiliki cukup data untuk belajar dan diuji secara adil.

- Pelatihan Logistic Regression

Model Logistic Regression dilatih dengan parameter optimal yang ditemukan melalui Grid Search. Model ini mampu menangkap hubungan linear antara fitur dan target.

- Evaluasi Model

  Akurasi Model: Tingkat akurasi mencapai 83.64%, menunjukkan bahwa model cukup baik dalam memprediksi hasil berdasarkan data input.

  Laporan Klasifikasi: Model menunjukkan performa yang baik pada precision, recall, dan f1-score untuk setiap kelas, dengan kinerja yang stabil dalam membedakan kelas.

- Visualisasi Hasil

  Grafik hasil klasifikasi dan metrik evaluasi memberikan wawasan tentang kinerja model dan distribusi kesalahan.

Kesimpulan

- Kinerja Model: Logistic Regression memberikan hasil akurasi yang baik pada dataset ini, mampu menangkap pola hubungan linear dalam data.
- Fitur Penting: Fitur-fitur yang berkontribusi pada model dapat dianalisis melalui koefisien model, memberikan insight tentang faktor-faktor yang mempengaruhi prediksi.

Analisis Hasil Evaluasi Model

- Akurasi

  Akurasi model mencapai 83.64%, menunjukkan bahwa model mampu memprediksi dengan benar untuk sebagian besar data dalam dataset pengujian.

- Interpretasi: Hasil ini baik dan menunjukkan bahwa model memiliki kinerja yang solid dalam prediksi.

Laporan Klasifikasi

- Precision

Kelas 0 (2,343 sampel): Precision sebesar 0.82, artinya dari semua prediksi kelas 0, 82% benar-benar merupakan kelas 0.

Kelas 1 (3,238 sampel): Precision sebesar 0.85, artinya 85% dari semua prediksi kelas 1 adalah benar.

Kesimpulan: Model menunjukkan kinerja yang baik dalam meminimalkan kesalahan prediksi positif palsu.

- Recall

  Kelas 0: Recall sebesar 0.79, menunjukkan model mampu mengenali 79% sampel kelas 0.

  Kelas 1: Recall sebesar 0.87, artinya dari seluruh sampel kelas 1, 87% diidentifikasi dengan benar oleh model.

Kesimpulan: Model sangat baik dalam mengenali kelas 1, tetapi ada sedikit kekurangan dalam mengenali semua sampel kelas 0.

- F1-score:

  Kelas 0: 0.80, menunjukkan keseimbangan yang baik antara precision dan recall.

  Kelas 1: 0.86, juga menunjukkan kinerja yang baik meskipun lebih tinggi dibandingkan kelas 0.

- Macro Avg:
- Rata-rata precision, recall, dan f1-score di seluruh kelas menunjukkan kinerja yang konsisten. Hasilnya adalah precision 0.83, recall 0.83, dan f1-score 0.83, menunjukkan model bekerja dengan baik secara keseluruhan.

c. Random Forrest

Insight dan Hasil

- Jumlah Data

  Dataset terdiri dari 18 kolom dan 27,901 baris, cukup untuk melatih model. Data kategori telah berhasil diubah menjadi numerik untuk digunakan dalam algoritma

machine learning.

- Prose Encoding

  Semua kolom kategori telah diubah menggunakan Label Encoding, memungkinkan model untuk memproses informasi ini secara efektif.

- Pembagian Data

  Dataset dibagi menjadi 80% untuk data pelatihan dan 20% untuk data pengujian. Ini memastikan model memiliki cukup data untuk belajar dan diuji secara adil.

- Pelatihan Random Forest

  Model Random Forest dilatih dengan parameter optimal yang ditemukan melalui Grid Search. Model ini menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting.

- Evaluasi Model

  Akurasi Model: Tingkat akurasi mencapai 82.99%, menunjukkan bahwa model cukup baik dalam memprediksi hasil berdasarkan data input.

  Laporan Klasifikasi: Model menunjukkan performa yang baik pada precision, recall, dan f1-score untuk setiap kelas, dengan kinerja yang stabil dalam membedakan kelas.

- Visualisasi Hasil

  Grafik hasil klasifikasi dan metrik evaluasi memberikan wawasan tentang kinerja model dan distribusi kesalahan. Visualisasi penting untuk memahami fitur yang berkontribusi pada prediksi.

Kesimpulan

- Kinerja Model: Random Forest memberikan hasil akurasi yang baik pada dataset ini, mampu menangkap pola kompleks dalam data.

- Fitur Penting: Fitur yang sering muncul di pohon keputusan menjadi indikator signifikan untuk prediksi.

Analisis Hasil Evaluasi Model

- Akurasi

  Akurasi model mencapai 82.99%, menunjukkan bahwa model mampu memprediksi dengan benar untuk sebagian besar data dalamdatasetpengujian.

- Interpretasi

  Hasil ini baik, menunjukkan bahwa model memiliki kinerja yang solid dalam prediksi.

Laporan Klasifikasi

- Precision

  Kelas 0 (2,343 sampel): Precision sebesar 0.81, artinya dari semua prediksi kelas 0, 81% benar-benar merupakan kelas 0.

  Kelas 1 (3,238 sampel): Precision sebesar 0.84, artinya 84% dari semua prediksi kelas 1 adalah benar.

Kesimpulan: Model menunjukkan kinerja yang baik dalam meminimalkan kesalahan prediksi positif palsu.

- Recall

  Kelas 0: Recall sebesar 0.78, menunjukkan model mampu mengenali 78% sampel kelas 0.

  Kelas 1: Recall sebesar 0.87, artinya dari seluruh sampel kelas 1, 87% diidentifikasi dengan benar oleh model.

Kesimpulan: Model sangat baik dalam mengenali kelas 1, tetapi ada sedikit kekurangan dalam mengenali semua sampel kelas 0.

- F1-score

  Kelas 0: 0.79, menunjukkan keseimbangan yang baik antara precision dan recall.

  Kelas 1: 0.86, juga menunjukkan kinerja yang baik meskipun lebih tinggi dibandingkan kelas 0.

- MacroAvg

  Rata-rata precision, recall, dan f1-score di seluruh kelas menunjukkan kinerja yang konsisten. Hasilnya adalah precision 0.82, recall 0.83, dan f1-score 0.82, menunjukkan model bekerja dengan baik secara keseluruhan.

2. Clustering
   a. K-Means

      Insight dan Hasil:

      - Jumlah Data:

        Dataset memiliki sejumlah kolom 14 dan baris yang cukup untuk melatih model (397924 Data). Data kategori berhasil diubah menjadi numerik agar kompatibel dengan algoritma machine learning.

      - Proses Encoding:

        Semua kolom kategori telah diubah ke bentuk numerik menggunakan Label Encoding, sehingga model dapat memproses informasi ini dengan baik. Ini memastikan bahwa setiap fitur yang bersifat kategorikal dapat digunakan dalam analisis clustering.

      - Pembagian Data:

        Analisis dilakukan pada keseluruhan dataset untuk menemukan pola tersembunyi. Dengan menggunakan data yang telah diproses, model dapat mengeksplorasi berbagai segmen dalam dataset.

      - Pelatihan K-Means:

        Model K-Means dilatih dengan menentukan jumlah cluster (K) yang optimal. Proses ini melibatkan iterasi untuk memperbarui posisi centroid hingga konvergensi tercapai. Hasil dari pelatihan ini adalah label cluster yang diberikan kepada setiap data.

Evaluasi Model:

- Performance:

  Silhouette Score: Untuk algoritma ini, nilai mencapai 0,380. Ini menunjukkan bahwa cluster cukup baik dengan jarak yang jelas antar cluster.

  Davies-Bouldin Index: Nilai mencapai 0,85, menunjukkan pemisahan yang baik antara cluster, dengan cluster yang lebih padat dan jarak antar cluster yang lebih besar.

- Visualisasi K-Means:

  Grafik yang menunjukkan cluster membantu dalam memahami pola distribusi data. Misalnya, cluster dengan karakteristik tertentu dapat diidentifikasi berdasarkan fitur yang paling berpengaruh, seperti Quantity dan UnitPrice.

- Kesimpulan:

  Kinerja Model: K-Means berhasil mengelompokkan data dengan jelas, dan hasil menunjukkan pola yang berarti dalam data. Model memberikan pemisahan yang baik antara cluster, dengan metrik evaluasi yang menunjukkan kualitas clustering yang memuaskan.

  Fitur Penting: Analisis cluster mengungkap fitur-fitur yang sering muncul dalam cluster tertentu, memberikan wawasan tentang kelompok pelanggan dan preferensi mereka. Misalnya, karakteristik tertentu dari produk mungkin menunjukkan pola pembelian yang relevan dengan segmen pasar tertentu.

b. DB Scan

Insight dan Hasil:

- Jumlah Data:

  Dataset yang sama (397924 Data) digunakan untuk analisis clustering dengan algoritma DBSCAN. Ini memungkinkan perbandingan yang tepat dengan hasil dari K-Means.

- Proses Encoding:

  Seperti pada K-Means, semua kolom kategori telah diubah ke bentuk numerik menggunakan Label Encoding. Ini memastikan bahwa DBSCAN dapat memproses data tanpa masalah.
- Pembagian Data:

  Proses analisis dilakukan pada keseluruhan dataset untuk menemukan pola yang tidak terlihat. DBSCAN berfokus pada area dengan kepadatan tinggi, yang dapat mengungkap kluster yang tidak teratur.
- Pelatihan DBSCAN:

  Model DBSCAN dilatih dengan parameter eps (radius pencarian untuk lingkungan) dan min_samples (jumlah minimal titik dalam lingkungan untuk membentuk cluster). Dengan menetapkan parameter ini, DBSCAN dapat mengidentifikasi kluster berdasarkan kepadatan.

Evaluasi Model:

- Performance:

  Silhouette Score: Untuk algoritma ini, nilai mencapai 0,145. Nilai ini menunjukkan bahwa pemisahan antar cluster tidak sejelas pada K-Means, dengan beberapa titik mungkin tidak terklasifikasi dengan baik.

  Davies-Bouldin Index: Nilai yang dihitung menunjukkan pemisahan antara cluster, meskipun tidak sebaik pada K-Means. Ini mengindikasikan bahwa ada beberapa tumpang tindih antar cluster.
- Visualisasi DBSCAN:

  Grafik yang menunjukkan hasil clustering DBSCAN membantu dalam memahami distribusi data. DBSCAN dapat mengidentifikasi pola yang lebih kompleks, termasuk noise (data yang tidak terklasifikasi) yang tidak ditemukan dalam metode K-Means.

- Kesimpulan:

  Kinerja Model: DBSCAN berhasil menemukan beberapa cluster dengan baik, tetapi juga mengidentifikasi banyak noise. Metrik evaluasi menunjukkan bahwa meskipun ada beberapa kluster yang terdeteksi, pemisahan antar cluster tidak sekuat yang ditemukan dengan K-Means.

  Fitur Penting: Analisis cluster mengungkap bahwa DBSCAN dapat menangkap struktur data yang lebih kompleks. Namun, tantangan dalam pemilihan parameter eps dan min_samples mempengaruhi hasil akhir yang optimal.

C. Comparative Analysis

1. Klasifikasi

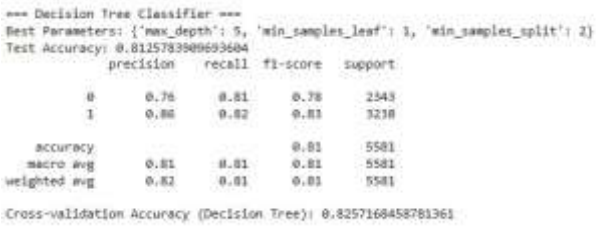
Berikut perbandingan hasil evaluasi model antara algoritma Decision Tree, Logistic Regression dan Random Forest berdasarkan metrik performa:

```
Model Comparison
Decision Tree Test Accuracy: 0.8125783909693604
Logistic Regression Test Accuracy: 0.8364092456549006
Random Forest Test Accuracy: 0.8299587887475363
```

Decision Tree Test Accuracy: 0.8125783909693604

Logistic Regression Test Accuracy: 0.8364092456549006

Random Forest Test Accuracy: 0.8299587887475363

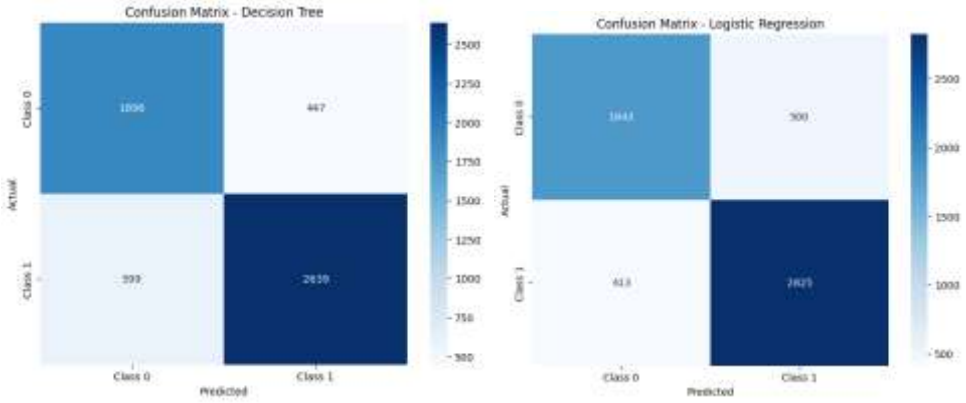


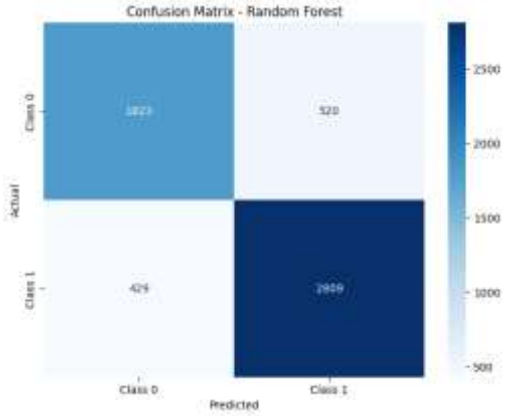
```
=== Decision Tree Classifier ===
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}
Test Accuracy: 0.8125783909693604
              precision    recall  f1-score   support

           0       0.76      0.81      0.78      2343
           1       0.86      0.82      0.83      3238

    accuracy                           0.81      5581
   macro avg       0.81      0.81      0.81      5581
weighted avg       0.82      0.81      0.81      5581

Cross-validation Accuracy (Decision Tree): 0.8257168458781361
```

```
=== Logistic Regression ===
Best Parameters: {'C': 0.1, 'solver': 'lbfgs'}
Test Accuracy: 0.8364092456549006
              precision    recall  f1-score   support

           0       0.82      0.79      0.80      2343
           1       0.85      0.87      0.86      3238

    accuracy                           0.84      5581
   macro avg       0.83      0.83      0.83      5581
weighted avg       0.84      0.84      0.84      5581

Cross-validation Accuracy (Logistic Regression): 0.8484767025089607
```

```
=== Random Forest Classifier ===
Best Parameters: {'bootstrap': True, 'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
Test Accuracy: 0.8299587887475363
              precision    recall  f1-score   support

           0       0.81      0.78      0.79      2343
           1       0.84      0.87      0.86      3238

    accuracy                           0.83      5581
   macro avg       0.83      0.83      0.83      5581
weighted avg       0.83      0.83      0.83      5581

Cross-validation Accuracy (Random Forest): 0.8474938994205134
```

Visualisasi :





2. Clustering

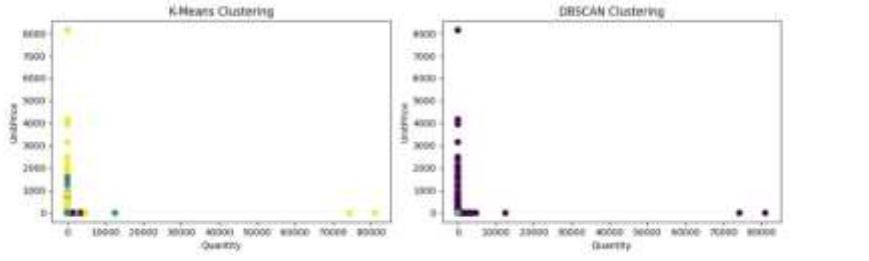
Berikut perbandingan hasil evaluasi model antara K-Means dan DB Scan berdasarkan metrik performa:

```
Jumlah baris data: 397924
Jumlah label K-Means: 397924
Jumlah label DBSCAN: 397924
```

```
K-Means Silhouette Score: 0.380
DBSCAN Silhouette Score: 0.145
```

```
Columns in the dataset: Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country', 'KMeans_Cluster',
       'DBSCAN_Cluster'],
      dtype='object')
First 5 rows with clustering results:
   InvoiceNo  StockCode  Description  Quantity         InvoiceDate  UnitPrice  \
0   536365.0       3249         3716         6 2010-12-01 08:26:00       2.55
1   536365.0       2649         3724         6 2010-12-01 08:26:00       3.39
2   536365.0       2855          861         8 2010-12-01 08:26:00       2.75
3   536365.0       2803         1813         6 2010-12-01 08:26:00       3.39
4   536365.0       2802         2776         6 2010-12-01 08:26:00       3.39

   CustomerID  Country  KMeans_Cluster  DBSCAN_Cluster
0     17850.0       35               2               0
1     17850.0       35               2               1
2     17850.0       35               1               2
3     17850.0       35               2               3
4     17850.0       35               2               4
```

Visualiasi



REFERENSI

Asfa, A. N. (2024). *Potensi Teknologi Blockchain Bagi Perusahaan Antar Barang Atau Dokumen. April*. Https://Doi.Org/10.13140/RG.2.2.32974.47686

Hidayat, I., Tolago, A. I., Dako, R. D. R., & Ilham, J. (2023). Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik. *Jambura Journal Of Electrical And Electronics Engineering*, *5*(2), 185–191. Https://Doi.Org/10.37905/Jjeee.V5i2.18397

Ihsan Ahmad Fauzi, & Raditya Danar Dana. (2023). Implementasi Data Mining Clustering Dalam Mengelompokan Kasus Perceraian Yang Terjadi Di Provinsi Jawa Barat Menggunakan Algoritma K-Means. *Maeswara : Jurnal Riset Ilmu Manajemen Dan Kewirausahaan*, *1*(4), 58–72. Https://Doi.Org/10.61132/Maeswara.V1i4.64

Rabbani, S., Safitri, D., Rahmadhani, N., Sani, A. A. F., & Anam, M. K. (2023). Perbandingan Evaluasi Kernel SVM Untuk Klasifikasi Sentimen Dalam Analisis Kenaikan Harga BBM. *MALCOM: Indonesian Journal Of Machine Learning And Computer Science*, *3*(2), 153–160. Https://Doi.Org/10.57152/Malcom.V3i2.897