PROKASDADIK: Prosiding Kecerdasan Artifisial, Sains Data, dan Pendidikan Masa Depan
Vol. 3, No. 2, (2025) E-ISSN: 3063-5845

Implementasi Model HDBSCAN dan XGBoost untuk Segmentasi Pelanggan dan Prediksi Produk Terlaris pada Data Transaksi *Retail* PT. XYZ

Raden Gumilar Riyansyah[1], Surya Permana[2], Masniari Samosir[3]
[1,2,3]Program Studi Teknik Informatika S-2, Universitas Pamulang
e-mail: radengumilarr@gmail.com[1], suryapermana.095018@gmail.com[2], samosirmasniari@gmail.com[3]

***Abstrak*— Transformasi digital dalam industri *retail* menghasilkan volume data transaksi yang sangat besar. Data ini berpotensi memberikan wawasan strategis bagi perusahaan, khususnya dalam memahami perilaku pelanggan dan mengelola stok produk. Namun, banyak perusahaan menghadapi kendala dalam mengolah data secara efisien, terutama dalam segmentasi pelanggan dan prediksi penjualan produk. Penelitian ini menawarkan solusi berupa sistem analitik berbasis data mining yang menggabungkan algoritma HDBSCAN untuk segmentasi pelanggan dan XGBoost untuk prediksi produk terlaris. *Dataset* transaksi pelanggan dan produk dari PT. XYZ digunakan sebagai objek studi. Proses analisis meliputi prapemrosesan data, pemodelan, dan visualisasi hasil menggunakan *framework Flask* dan pustaka *Plotly*. Hasil penelitian menunjukkan bahwa sistem ini mampu mengidentifikasi pola pembelian pelanggan secara efektif serta memprediksi produk dengan penjualan tertinggi secara akurat. Sistem ini diharapkan dapat mendukung pengambilan keputusan bisnis secara berbasis data.**

***Kata Kunci*— Data Mining, HDBSCAN, Prediksi Penjualan, *Retail*, XGBoost**

I. PENDAHULUAN

Industri *retail* saat ini berada dalam tekanan transformasi digital yang cepat, menghasilkan volume data transaksi dalam jumlah besar setiap harinya. Data ini, apabila diolah dengan tepat, dapat memberikan informasi berharga untuk mendukung pengambilan keputusan strategis, seperti identifikasi pelanggan prioritas, penyusunan strategi promosi, serta pengelolaan stok barang [1]. Namun demikian, banyak perusahaan menghadapi tantangan dalam mengelola dan menganalisis data tersebut secara efisien karena keterbatasan teknologi dan sumber daya manusia yang memahami analisis data secara mendalam [2].

Permasalahan utama yang sering dihadapi adalah kurangnya pemahaman mendalam terhadap pola perilaku pelanggan, serta kesulitan dalam memprediksi produk mana yang akan memiliki tingkat penjualan tinggi di masa mendatang [3]. Strategi pemasaran seringkali dilakukan secara umum tanpa mempertimbangkan segmentasi pelanggan yang tepat, dan keputusan pembelian stok cenderung bersifat spekulatif [4]. Oleh karena itu, dibutuhkan pendekatan analitik yang mampu menangani data berskala besar dan tidak beraturan secara otomatis.

Beberapa penelitian telah mengusulkan pendekatan berbeda dalam menangani permasalahan serupa. Kasmari dan Taryadi (2023) melakukan segmentasi pelanggan dengan model RFM dan *cluster*ing *K-Means* serta analisis demografi. Hasilnya menunjukkan bahwa kombinasi bobot RFM dan atribut demografi menghasilkan 2491 *rule* dengan skor tertinggi sebesar 0.284 [5]. Selanjutnya, Muhammad Syam Al Ghifari et al. (2024) mengelompokkan data transaksi penjualan aksesoris HP dan pulsa di Bagus Celluler menggunakan algoritma *K-Means*. Dengan nilai *Davies Bouldin Index* (DBI) sebesar -0.205, diperoleh jumlah *cluster* optimal sebanyak 3 [6].

Kemudian Fatia A. Maresti et al. (2025) mengimplementasikan *K-Means* dengan metode *elbow* pada data transaksi *retail* untuk segmentasi produk. Hasil segmentasi menunjukkan 3 *cluster* optimal dengan variabel harga sebagai faktor paling dominan dalam pembentukan klaster [7]. Sementara itu, Fadhlan Agus Setiawan et al. (2024) mengidentifikasi pola pembelian parfum dengan algoritma Apriori. Itemset yang paling sering dibeli bersamaan adalah *Jasmine* dan *Fantasy* dengan support minimum sebesar 15% [8]. Berikutnya Arwa Ulayya Haspriyanti et al. (2021) memprediksi produk layanan *Indihome* terlaris menggunakan metode *K-Nearest Neighbor* (KNN). Dengan akurasi sebesar 99,99%, layanan internet menjadi produk paling dominan [9].

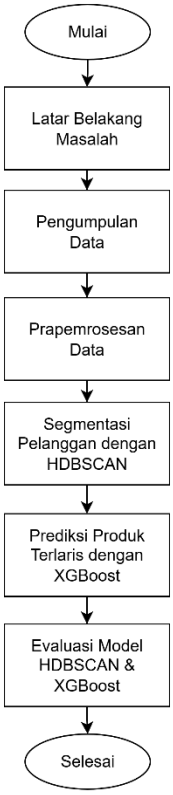
Berdasarkan latar belakang dan studi sebelumnya, penelitian ini mengusulkan solusi berupa sistem analitik berbasis *web* yang mengintegrasikan algoritma HDBSCAN dan XGBoost dalam satu platform. HDBSCAN dipilih karena kemampuannya dalam melakukan segmentasi pelanggan tanpa perlu menentukan jumlah *cluster* di awal dan dapat menangani *noise*. Sementara itu, XGBoost dipilih karena kecepatan dan akurasinya dalam melakukan prediksi numerik, khususnya pada data penjualan. Sistem ini diimplementasikan menggunakan *framework Flask* yang dilengkapi visualisasi interaktif dari pustaka *Plotly* untuk

PROKASDADIK: Prosiding Kecerdasan Artifisial, Sains Data, dan Pendidikan Masa Depan
Vol. 3, No. 2, (2025) E-ISSN: 3063-5845

memudahkan analisis oleh pengguna bisnis non-teknis.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksploratif untuk membangun sebuah sistem analitik berbasis data transaksi *retail*. Tujuan utama penelitian adalah melakukan segmentasi pelanggan secara otomatis serta memprediksi produk dengan nilai penjualan tertinggi. Adapun tahapan metodologi penelitian secara sistematis dijelaskan sebagai berikut:



Gambar 1. Tahapan Penelitian

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari data transaksi harian PT. XYZ. Terdapat dua *Dataset* utama, yaitu:

a. **Data pelanggan**, yang memuat informasi: nama pelanggan, kota, jumlah transaksi (QtyTransaksi), nilai transaksi (ValueTransaksi), dan jenis pembayaran.
b. **Data produk**, yang memuat: nama produk, jumlah produk terjual (*Qty*), dan nilai penjualan (*Sales*). Total data mentah yang terkumpul berjumlah 11.625 baris, yang kemudian dilakukan proses seleksi dan validasi untuk digunakan dalam analisis.

B. Prapemrosesan Data

Tahapan prapemrosesan dilakukan untuk memastikan kualitas data sesuai dengan kebutuhan algoritma yang digunakan. Proses ini meliputi:

a. Penghapusan data kosong dan duplikat.
b. Transformasi kolom ValueTransaksi dari format string ke numerik.
c. Penerapan *Label Encoding* untuk mengonversi fitur kategorikal seperti Kota dan JenisPembayaran ke bentuk numerik.
d. Standardisasi data numerik menggunakan *StandardScaler* guna menormalkan skala antar fitur sehingga tidak mendominasi perhitungan jarak dalam proses klasterisasi.

C. Segmentasi Pelanggan dengan HDBSCAN

Segmentasi pelanggan dilakukan dengan algoritma HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). Algoritma ini dipilih karena memiliki keunggulan dalam mengelompokkan data yang memiliki

kepadatan tidak merata dan tidak memerlukan penentuan jumlah klaster di awal. Selain itu, HDBSCAN juga mampu mengidentifikasi data yang dianggap sebagai *noise*, yang ditandai dengan label -1. Fitur yang digunakan dalam proses klasterisasi ini adalah QtyTransaksi, ValueTransaksi, *CityEncoded*, dan *PayEncoded*. Hasil klasterisasi divisualisasikan dalam bentuk *scatter plot* dan histogram distribusi klaster.

D. Prediksi Produk Terlaris dengan *XGBoost*

Prediksi produk unggulan dilakukan menggunakan algoritma *XGBoost* berbasis regresi. Model ini dilatih menggunakan data historis penjualan produk dengan fitur input berupa Qty dan target prediksi adalah Sales. Model *XGBoost* dipilih karena performanya yang tinggi dalam pemodelan prediksi pada data dengan kompleksitas tinggi serta kemampuannya dalam menangani data numerik tanpa perlu banyak prapemrosesan tambahan. Hasil prediksi ditampilkan dalam bentuk grafik batang untuk mengidentifikasi produk dengan nilai penjualan tertinggi.

E. Evaluasi Model

Evaluasi dilakukan terhadap dua model utama, yaitu:

a. **Model HDBSCAN**, yang dievaluasi menggunakan metrik *Silhouette Score* untuk mengukur validitas pemisahan klaster. Nilai Silhouette yang positif menunjukkan pemisahan klaster yang baik, sedangkan nilai negatif mengindikasikan adanya tumpang tindih antar klaster.
b. **Model XGBoost**, yang dievaluasi menggunakan tiga metrik regresi, yaitu: *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *R-squared* (R^2). Ketiga metrik tersebut digunakan untuk menilai seberapa akurat model dalam memprediksi nilai penjualan.

F. Implementasi Sistem Aplikasi

Sistem aplikasi dikembangkan dalam bentuk *web* berbasis *framework* ***Flask*** dengan antarmuka interaktif menggunakan pustaka ***Plotly***. Sistem ini memungkinkan pengguna untuk:

a. Mengunggah *Dataset* transaksi dalam format CSV,
b. Menjalankan proses segmentasi pelanggan dan prediksi penjualan secara otomatis,
c. Melihat hasil analisis dalam bentuk visualisasi grafik yang interaktif,
d. Mengunduh hasil analisis dalam bentuk laporan HTML.

G. Perangkat dan Lingkungan Pengembangan

Seluruh proses analisis dilakukan menggunakan bahasa pemrograman **Python** dengan pustaka-pustaka utama sebagai berikut:

a. pandas dan numpy untuk manipulasi dan analisis data,
b. hdbscan, xgboost, dan scikit-learn untuk proses pemodelan,
c. *Plotly* untuk visualisasi hasil dalam format interaktif,
d. *Flask* untuk pengembangan sistem antarmuka berbasis *web*.

III. HASIL DAN PEMBAHASAN

A. Hasil Pengumpulan Data dan Hasil Prapemrosesan Data

Data yang digunakan dalam penelitian ini diperoleh dari catatan transaksi harian pelanggan PT. XYZ. *Dataset* terdiri dari dua kelompok utama, yaitu data pelanggan dan data produk. Data pelanggan mencakup atribut nama pelanggan, kota domisili (*City*), jenis pembayaran (JenisPembayaran), jumlah produk yang dibeli (QtyTransaksi), dan nilai total transaksi (ValueTransaksi). Sementara itu, data produk memuat informasi nama produk, jumlah produk terjual (Qty), dan nilai penjualan (Sales).

Dari hasil pengumpulan data diperoleh total sebanyak 11.625 baris data transaksi mentah. Data tersebut mencerminkan aktivitas transaksi ritel dalam periode tertentu yang bersifat heterogen dan tersebar di berbagai wilayah kota. Sebelum dilakukan analisis, seluruh data dievaluasi untuk menjamin kelengkapan dan kesesuaiannya dengan kebutuhan pemodelan. Berikut dibawah ini merupakan Tabel *Dataset*.

Tabel 1 *Dataset* Transaksi Penjualan By Customer

No	CustName	*City*	NoReceipt	JenisPembayaran	QtyTransaksi	ValueTransaksi	Dept
1	KID	Serang	UF2.2.20241027.18	BNI Debit	1	12999000	Mattress
2	WRA	Serang	UF2.2.20241028.6	Home Credit	1	2999000	Kitchen
3	WWRA	Serang	UF2.2.20241028.6	Home Credit	1	3499000	Kitchen

No	CustName	*City*	NoReceipt	JenisPembayaran	QtyTransaksi	ValueTransaksi	Dept
4	YS	Cilegon	UF2.2.20241028.9	KREDIVO	2	279000	Comm Chair Table
5	YS	Cilegon	UF2.2.20241028.9	KREDIVO	1	2599000	Bedroom
6	MFA	Cilegon	UF2.3.20241029.1	Cash	1	5399000	Living Upholstered
7	S	Cilegon	UF2.2.20241030.1	Cash	1	7199000	Office
8	S	Cilegon	UF2.2.20241030.1	Cash	2	7199000	Office
9	M	Cilegon	UF2.2.20241030.3	BNI QR	1	899000	Office Seating
10	M	Cilegon	UF2.2.20241030.3	BNI QR	5	499000	Office Seating
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1809	HRN	Serang	UF2.45.20241028.1	MANDIRI Debit	1	8999000	Bedroom

Pada Tabel 1 menyajikan data mentah hasil transaksi ritel yang dilakukan oleh pelanggan PT. XYZ. Atribut utama yang tercakup meliputi nama pelanggan (CustName), kota domisili (*City*), nomor transaksi (NoReceipt), jenis pembayaran (JenisPembayaran), jumlah produk yang dibeli (QtyTransaksi), nilai total transaksi (ValueTransaksi), dan departemen produk (Dept). Data ini menggambarkan karakteristik awal pelanggan berdasarkan perilaku transaksional mereka, yang menjadi dasar dalam segmentasi pelanggan menggunakan algoritma HDBSCAN. Penyusunan data secara struktural dalam tabel ini memungkinkan penelusuran pola pembelian berdasarkan lokasi dan metode pembayaran, serta membantu dalam mengidentifikasi pelanggan potensial berdasarkan nilai transaksinya.

Tabel 2 *Dataset* Transaksi Penjualan By Product

No	Dept Name	Stock Code	Produck Name	Qty	Sales
1	Sofa	10023758	LUMI ARM CHAIR SAND-DALLAS420	1	3728901
2	Sofa	10027167	BIELLA SOFA 3S SLV GRY#BV/PVC-BV-041E	2	30629100
3	Sofa	10027168	BIELLA SOFA 2S SLV GRY#BV/PVC-BV-041E	1	14805450
4	Sofa	10027169	BIELLA SOFA 1S SLV GRY#BV/PVC-BV-041E	2	17457928
5	Sofa	10031492	KRINGLE SOFA 2S LIGHT BROWN#CU203-5	5	35310900
6	Sofa	10031493	KRINGLE SOFA 3S LIGHT BROWN#CU203-5	3	27902749
7	Sofa	10032365	MACLAINE SOFA 3S RC BROWN-803(S/2)	4	40580584
8	Sofa	10041971	SPAREPART OTHERS LIVING UP	4	2522522
9	Sofa	10060984	KAMMA ARM CHAIR GREY-HOLLY7	3	16213514
10	Sofa	10069031	MADISON SOFA 2S RC BROWN#2527	1	11621081
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
9816	Pos Transaction	70125163	BIAYA INSTALLASI HCIR	134	18518311

Pada Tabel 2 menyajikan rincian penjualan pada level produk. Atribut yang ditampilkan mencakup nama departemen (Dept Name), kode stok (Stock Code), nama produk (Product Name), jumlah produk yang terjual (Qty), dan nilai penjualan (Sales). *Dataset* ini menjadi dasar dalam proses prediksi penjualan menggunakan model XGBoost. Ketersediaan informasi granular terkait performa masing-masing produk memfasilitasi identifikasi produk unggulan (best-seller) dan dapat dijadikan sebagai referensi dalam pengambilan keputusan manajerial terkait pengadaan barang dan perencanaan promosi.

Tahapan prapemrosesan data dilakukan untuk memastikan bahwa data yang digunakan memiliki kualitas dan format yang sesuai untuk kebutuhan algoritma yang diterapkan. Dari total 11.625 data transaksi yang terkumpul, dilakukan proses seleksi untuk menghapus baris yang bersifat duplikat dan tidak lengkap. Setelah proses validasi, diperoleh sebanyak 1.809 data transaksi yang layak digunakan untuk analisis lebih lanjut. Pada tahap ini, seluruh fitur numerik seperti ValueTransaksi dan QtyTransaksi dikonversi dari format teks ke numerik. Fitur kategorikal seperti nama kota (*City*) dan jenis pembayaran (JenisPembayaran) dikodekan menggunakan metode *Label Encoding*, sehingga menghasilkan atribut baru *CityEncoded* dan *PayEncoded*. Langkah ini

PROKASDADIK: Prosiding Kecerdasan Artifisial, Sains Data, dan Pendidikan Masa Depan
Vol. 3, No. 2, (2025) E-ISSN: 3063-5845

bertujuan untuk memungkinkan algoritma machine learning mengenali informasi kategorikal dalam bentuk numerik. Kemudian, seluruh fitur numerik dinormalisasi menggunakan metode *StandardScaler*, guna memastikan skala antar fitur seragam dan mencegah dominasi satu fitur terhadap yang lain dalam proses perhitungan jarak atau pemodelan prediktif.

B. Hasil Segmentasi Pelanggan dengan HDBSCAN

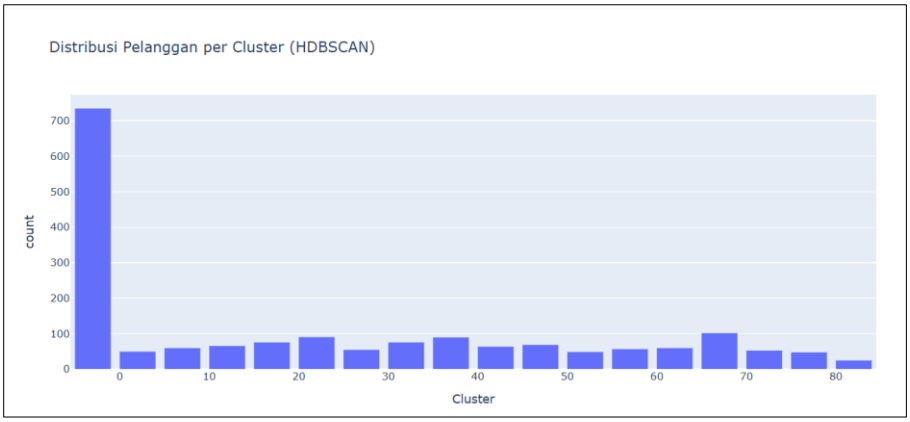
Segmentasi pelanggan dilakukan menggunakan algoritma HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) yang dipilih karena keunggulannya dalam menangani data yang tidak memiliki distribusi homogen, serta kemampuannya dalam mendeteksi *noise* atau anomali secara otomatis. Algoritma ini tidak memerlukan penentuan jumlah klaster di awal, yang menjadi kelebihan dibandingkan metode konvensional seperti *K-Means*.

Fitur yang digunakan untuk proses segmentasi terdiri dari empat atribut utama: QtyTransaksi, ValueTransaksi, *CityEncoded*, dan *PayEncoded*. Hasil segmentasi divisualisasikan dalam bentuk ***scatter plot***, yang memperlihatkan distribusi pelanggan ke dalam beberapa klaster berdasarkan kemiripan pola transaksi mereka. Selain itu, histogram distribusi klaster menunjukkan persebaran jumlah anggota pada masing-masing klaster.

Tabel 3 Hasil Segmentasi Pelanggan

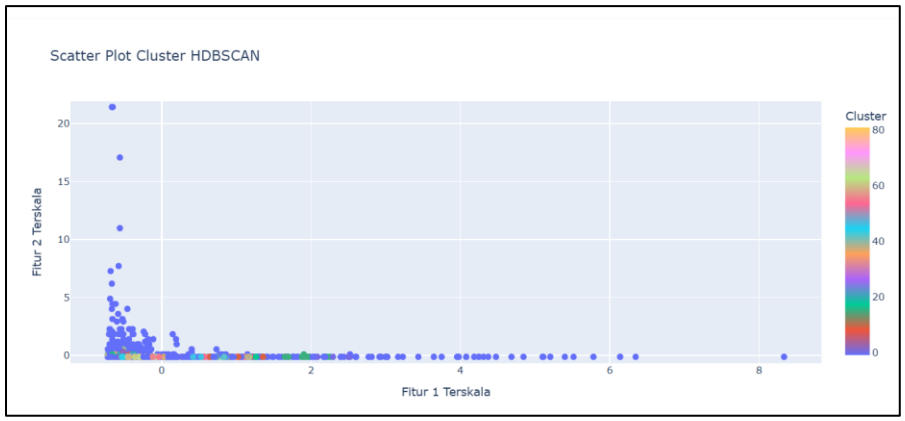
Cust Name	*City*	No Receipt	Jenis Pembayaran	Qty Transaksi	Value Transaksi	Dept	*City Encoded*	*Pay Encoded*	*Cluster*
KID	Serang	UF2.2.20241027.18	BNI Debit	1	12999000	Mattress	10	10	-1
WRAF	Serang	UF2.2.20241028.6	Home Credit	1	2999000	Kitchen	10	29	70
WRAF	Serang	UF2.2.20241028.6	Home Credit	1	3499000	Kitchen	10	29	-1
YS	Cilegon	UF2.2.20241028.9	KREDIVO	2	279000	Comm Chair Table	1	30	21
YS	Cilegon	UF2.2.20241028.9	KREDIVO	1	2599000	Bedroom	1	30	-1
MFAA	Cilegon	UF2.3.20241029.1	Cash	1	5399000	Living Upholstered	1	22	24
S	Cilegon	UF2.2.20241030.1	Cash	1	7199000	Office	1	22	22
S	Cilegon	UF2.2.20241030.1	Cash	2	7199000	Office	1	22	-1
IM/PA	Cilegon	UF2.2.20241030.3	BNI QR	1	899000	Office Seating	1	12	-1
IM/PA	Cilegon	UF2.2.20241030.3	BNI QR	5	499000	Office Seating	1	12	-1

Pada Tabel 3 menampilkan hasil akhir segmentasi pelanggan yang diperoleh melalui algoritma HDBSCAN. Setiap baris mencerminkan data pelanggan yang telah diproses, lengkap dengan hasil encoding kota (*City Encoded*) dan jenis pembayaran (*Pay Encoded*), serta label klaster hasil segmentasi (*Cluster*). Keberadaan label -1 menunjukkan bahwa data pelanggan tersebut dikategorikan sebagai *noise* atau outlier oleh algoritma. Melalui tabel ini, dapat diidentifikasi kelompok pelanggan dengan perilaku serupa yang dapat ditindaklanjuti lebih lanjut dalam strategi promosi yang dipersonalisasi atau program loyalitas. Tabel ini juga berperan sebagai bukti efektivitas pemodelan unsupervised learning dalam memetakan pola transaksi pelanggan. Berikut dibawah ini merupakan hasil visualisasi distribusi pelanggan per *cluster* (HDBSCAN).



Gambar 2 Visualisasi Distribusi Pelanggan per *Cluster* (HDBSCAN)

Pada Gambar 2 menampilkan histogram distribusi pelanggan berdasarkan hasil klaster HDBSCAN. Setiap batang merepresentasikan jumlah pelanggan yang tergabung dalam masing-masing klaster. Histogram ini penting untuk melihat proporsi distribusi, mendeteksi adanya dominasi klaster tertentu, serta mengidentifikasi besarnya jumlah data yang termasuk ke dalam *noise* (label -1). Informasi ini membantu dalam evaluasi kualitas segmentasi dan menjadi dasar pertimbangan untuk perbaikan parameter model apabila diperlukan.



Gambar 3 *Scatter plot Cluster* HDBSCAN

Pada Gambar 3 *scatter plot* ini merupakan representasi visual dari hasil segmentasi pelanggan yang dilakukan oleh HDBSCAN. Setiap titik pada grafik menunjukkan satu pelanggan, yang dipetakan berdasarkan kombinasi fitur numerik (misalnya QtyTransaksi dan ValueTransaksi) yang telah distandarisasi. Warna yang berbeda menandakan perbedaan klaster. Klaster dengan label -1 biasanya ditampilkan dengan warna abu-abu atau gelap sebagai indikasi *noise*. *Scatter plot* ini memberikan pandangan intuitif terhadap separabilitas antar klaster dan menunjukkan adanya kelompok pelanggan dengan karakteristik yang serupa dalam hal perilaku transaksi.

C. Hasil Prediksi Produk Terlaris dengan *XGBoost*

Prediksi nilai penjualan produk dilakukan dengan memanfaatkan algoritma **XGBoost Regressor**, sebuah metode berbasis pohon keputusan yang terkenal memiliki akurasi tinggi dan efisiensi komputasi yang baik dalam memproses data numerik. Dalam penelitian ini, model dilatih dengan menggunakan fitur Qty (jumlah produk terjual) sebagai variabel prediktor, dan Sales (nilai penjualan) sebagai target yang diprediksi.

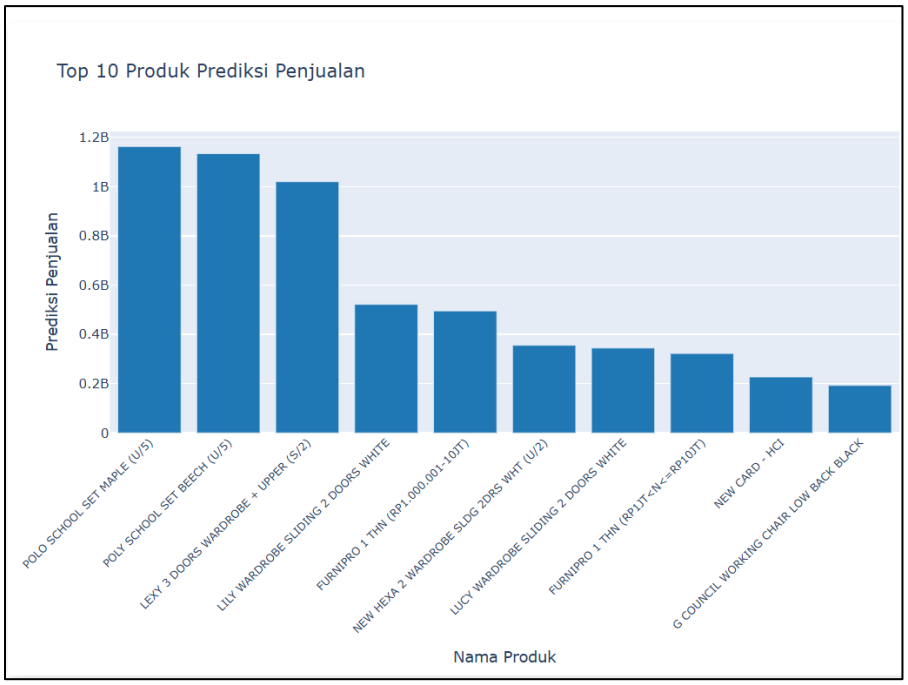
Hasil prediksi ditampilkan dalam bentuk **grafik batang (*bar chart*)** yang menunjukkan daftar 10 produk dengan prediksi nilai penjualan tertinggi. Visualisasi ini membantu manajer operasional untuk memahami produk mana yang diperkirakan akan menghasilkan pendapatan terbesar, sehingga dapat digunakan sebagai dasar dalam perencanaan stok maupun strategi promosi.

Tabel 4 Prediksi Produk *Best Seller*

No	Dept Name	Stock Code	Produck Name	Qty	Sales	PredictedSales
1	Sofa	10023758	LUMI ARM CHAIR SAND-DALLAS420	1	3728901	1.795.242.375
2	Sofa	10027167	BIELLA SOFA 3S SLV GRY#BV/PVC-BV-041E	2	30629100	2.707.142.250
3	Sofa	10027168	BIELLA SOFA 2S SLV GRY#BV/PVC-BV-041E	1	14805450	1.795.242.375
4	Sofa	10027169	BIELLA SOFA 1S SLV GRY#BV/PVC-BV-041E	2	17457928	2.707.142.250
5	Sofa	10031492	KRINGLE SOFA 2S LIGHT BROWN#CU203-5	5	35310900	5.593.626.500
6	Sofa	10031493	KRINGLE SOFA 3S LIGHT BROWN#CU203-5	3	27902749	3.776.762.000
7	Sofa	10032365	MACLAINE SOFA 3S RC BROWN-803(S/2)	4	40580584	4.878.319.500
8	Sofa	10041971	SPAREPART OTHERS LIVING UP	4	2522522	4.878.319.500
9	Sofa	10060984	KAMMA ARM CHAIR GREY-HOLLY7	3	16213514	3.776.762.000
10	Sofa	10069031	MADISON SOFA 2S RC BROWN#2527	1	11621081	1795242.3

Pada Tabel 4 merupakan hasil prediksi nilai penjualan produk menggunakan model regresi XGBoost. Setiap baris merepresentasikan sebuah produk dengan estimasi nilai penjualan yang diprediksi oleh model. Produk-produk yang menduduki peringkat teratas dalam tabel ini dapat dikategorikan sebagai *Best Seller* dan sebaiknya diprioritaskan dalam proses produksi, distribusi, serta promosi. Informasi dari tabel ini dapat dijadikan dasar objektif dalam pengambilan keputusan manajemen inventory dan strategi penjualan.

PROKASDADIK: Prosiding Kecerdasan Artifisial, Sains Data, dan Pendidikan Masa Depan
Vol. 3, No. 2, (2025) E-ISSN: 3063-5845

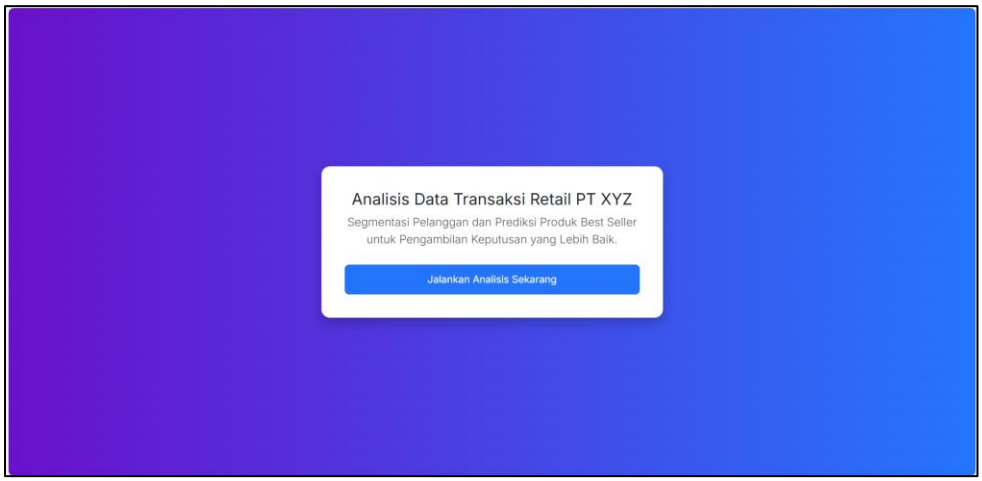


Gambar 4 Top 10 Produk Prediksi Penjualan

Pada Gambar 4 menampilkan grafik batang yang memvisualisasikan sepuluh produk teratas berdasarkan hasil prediksi nilai penjualan tertinggi. Grafik ini menyajikan informasi kuantitatif yang mudah diinterpretasikan oleh pengguna non-teknis, dan membantu dalam mengidentifikasi produk dengan performa penjualan tertinggi secara visual. Penggunaan visualisasi interaktif dari pustaka *Plotly* memungkinkan eksplorasi data lebih mendalam oleh pengguna aplikasi.

D. Hasil Implementasi Aplikasi

Sistem aplikasi berhasil diimplementasikan dalam bentuk *web* interaktif menggunakan *framework Flask*. Pengguna dapat mengunggah *Dataset* dalam format CSV, menjalankan proses segmentasi dan prediksi secara otomatis, serta melihat hasil visualisasi dalam bentuk grafik interaktif. Hasil analisis dapat diunduh dalam format laporan HTML. Visualisasi hasil yang ditampilkan meliputi *scatter plot* segmentasi pelanggan, histogram distribusi klaster, dan grafik batang prediksi nilai penjualan. Implementasi antarmuka ini bertujuan untuk memudahkan pengguna non-teknis dalam memahami dan memanfaatkan hasil analitik untuk pengambilan keputusan bisnis.



Gambar 5 Hasil Implementasi Aplikasi Berbasis *Web*site

Pada Gambar 5 menunjukkan antarmuka pengguna dari aplikasi *web* berbasis *Flask* yang dikembangkan dalam penelitian. Fitur utama aplikasi meliputi unggah *Dataset* transaksi dalam format CSV, pemrosesan segmentasi pelanggan dan prediksi produk, serta penyajian hasil analisis dalam bentuk grafik interaktif. Antarmuka dirancang agar mudah digunakan oleh pengguna non-teknis, dengan tampilan yang responsif dan informatif. Implementasi ini menjadi bukti keberhasilan integrasi teknologi data mining dengan sistem informasi berbasis *web* untuk mendukung pengambilan keputusan bisnis secara efisien.

PROKASDADIK: Prosiding Kecerdasan Artifisial, Sains Data, dan Pendidikan Masa Depan
Vol. 3, No. 2, (2025) E-ISSN: 3063-5845

E. Hasil Evaluasi Model HDBSCAN

Evaluasi segmentasi menggunakan metrik *Silhouette Score* menunjukkan nilai sebesar **-0.0427**, yang mengindikasikan bahwa hasil pemisahan klaster belum optimal. Hal ini kemungkinan disebabkan oleh keberadaan data *noise* yang cukup banyak dalam struktur transaksi pelanggan yang tidak homogen. Nilai negatif mengindikasikan adanya tumpang tindih antar klaster, meskipun dari visualisasi *scatter plot*, sebagian kelompok pelanggan masih dapat teridentifikasi dengan baik.

F. Hasil Evaluasi Model XGBoost

Evaluasi regresi menggunakan XGBoost untuk prediksi nilai penjualan produk menghasilkan nilai:

a. ***Mean Absolute Error* (MAE):** 8.032.732,50
b. ***Mean Squared Error* (MSE):** 421.645.562.937.344
c. ***R-squared* (R^2):** 0.5745

Nilai R^2 sebesar 0.5745 menunjukkan bahwa model mampu menjelaskan sekitar 57,45% variabilitas pada data target. Meskipun belum mendekati 1 (model sempurna), model ini sudah menunjukkan kinerja yang cukup baik untuk konteks data *retail* bersifat kompleks dan *heterogen*.

IV. KESIMPULAN

Berdasarkan hasil implementasi dan evaluasi, sistem analitik berbasis HDBSCAN dan XGBoost yang dikembangkan mampu membantu dalam proses segmentasi pelanggan dan prediksi nilai penjualan produk. Meskipun hasil evaluasi HDBSCAN menunjukkan skor *Silhouette* yang negatif, sistem tetap berhasil mengidentifikasi kelompok pelanggan yang relevan untuk strategi pemasaran. Di sisi lain, model XGBoost mampu memberikan prediksi penjualan dengan tingkat akurasi yang memadai, terbukti dari nilai R^2 yang mendekati 0.6. Sistem ini berhasil diintegrasikan ke dalam aplikasi berbasis *web* menggunakan *Flask* dan *Plotly* untuk visualisasi interaktif, sehingga dapat dioperasikan oleh pengguna non-teknis. Secara keseluruhan, pendekatan ini memberikan solusi berbasis data yang aplikatif bagi perusahaan *retail* dalam pengambilan keputusan bisnis yang lebih terarah.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada program studi magister teknik informatika, universitas pamulang, atas dukungan yang telah diberikan dalam pelaksanaan penelitian ini. Ucapan terima kasih secara khusus disampaikan kepada Dr. Tukiyat, M.si selaku dosen pembimbing yang telah membimbing dan memberikan arahan konstruktif selama proses penyusunan dan penyelesaian penelitian ini. Penulis juga menyampaikan apresiasi kepada PT. XYZ selaku penyedia data transaksi yang menjadi objek kajian, serta seluruh pihak yang telah memberikan kontribusi dalam bentuk masukan, dukungan teknis, maupun motivasi selama proses penelitian berlangsung. Semoga hasil dari penelitian ini dapat memberikan kontribusi nyata dalam pengembangan ilmu pengetahuan dan penerapan teknologi analitik di bidang *retail*.

DAFTAR PUSTAKA

[1] D. Ashari, M. S. Ladaina, And T. Hartini, “Peran Big Data Dalam Pengambilan Keputusan Strategis Perusahaan,” Pp. 401–422.
[2] J. M. Polgan *Et Al.*, “Evolusi Sistem Informasi Akuntansi Dalam Era Digital : Tinjauan Literatur Tentang Tren ,” Vol. 14, Pp. 77–85, 2025.
[3] A. A. Zahra, A. K. Putri, And Z. N. Cahyani, “Business Plan Sebagai Cermin Kognisi Dan Perilaku Wirausaha : Pendekatan Psikologis Dalam Perencanaan Bisnis,” Vol. 02, No. June, Pp. 773–782, 2025.
[4] S. Jurnal, “Jurnal Sistem Informasi Dan Teknologi (S I N T E K),” Vol. V, No. 01, Pp. 90–99.
[5] U. Stikubank And S. W. Pratama, “Segmentasi Pelanggan Berdasarkan Analisis Rfm (Recency ,” 2018.
[6] M. Syam Al Ghifari, M. Martanto, And U. Hayati, “Pengelompokan Transaksi Penjualan Aksesoris Hp Dan Pulsa Dengan Metode *K-Means* Untuk Meningkatkan Strategi Pemasaran Di Toko Bagus Celluler,” *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 8, No. 3, Pp. 2838–2849, 2024, Doi: 10.36040/Jati.V8i3.9559.
[7] F. A. Maresti, W. I. Rahayu, M. B. B. C. Lustin, T. H. Pakpahan, And K. Bandung, “Implementasi K- Means Untuk Melakukan,” Vol. 9, No. 1, Pp. 20–32, 2025.
[8] F. A. Setiawan, “Pola Pembelian Produk Parfum Menggunakan Algoritma Apriori Berdasarkan Data Mining *Rule* Asosiasi,” Vol. 4, No. 3, Pp. 51–63, 2024.
[9] A. U. Haspriyanti And P. W. Prasetyaningrum, “Penerapan Data Mining Untuk Prediksi Layanan Produk *Indihome* Menggunakan Metode K-Nearst Neighbor Arwa,” *Inf. Syst. Artif. Intell.*, vol. 20, no. 2, pp. 100–107, 2021.