Saturnus: Jurnal Teknologi dan Sistem Informasi
Volume. 4, Nomor. 1, Januari 2026
E-ISSN : 3031-9943; P-ISSN : 3031-9935; Hal. 94-102
DOI: https://doi.org/10.61132/saturnus.v4i1.1409
Tersedia: https://journal.arteii.or.id/index.php/Saturnus

Analisis Prediksi Penjualan Bisnis Retail Menggunakan Metode *Decision Tree* dan *Random Forest*

Agung Narayana Adhi Putra[1*], I Wayan Sudiarsa[2], I Kadek Adi Gunawan[3], Kadek Bagus Karunia Dwi Dharmayasa[4], I Wayan Eka Saputra[5]
[1-5] S1 Rekayasa Sistem Komputer, Institut Bisnis dan Teknologi Indonesia, Indonesia
**Penulis Korespondensi: sudiarsa@instiki.ac.id*

Abstract. *The retail industry generates an extremely large and continuously growing volume of transactional data along with the advancement of digital technology, thereby requiring sophisticated and systematic data analysis approaches to support effective and evidence-based business decision-making. This study aims to analyze retail sales data by utilizing the Retail Sales Dataset obtained from the Kaggle platform, which consists of 100,000 transaction records and broadly represents the characteristics of retail transactions. The main focus of this study is to classify product categories and predict customer segments, including the identification of high-spending customers (high spenders), based on demographic attributes such as age and gender, as well as various transaction-related features. The research methodology includes data preprocessing, label encoding, and feature engineering to generate additional variables, including Age_Group, Is_Holiday, and Spender_Group, which are expected to enhance the predictive capability of the models. Several machine learning algorithms, namely Decision Tree, Random Forest, and XGBoost, were implemented and evaluated to compare their respective performance. The experimental results indicate that multiclass product category classification achieves relatively low accuracy, ranging from 27% to 34%. These findings suggest the high complexity of retail data and highlight the need for further model optimization, class balancing techniques, and feature refinement to improve predictive performance in future studies.*

Keywords: *Feature Engineering; High Spender Prediction; Hyperparameter Tuning; Random Forest; Retail Sales Prediction.*

Abstrak. Industri ritel menghasilkan volume data transaksi yang sangat besar dan terus meningkat seiring dengan perkembangan teknologi digital, sehingga memerlukan pendekatan analisis data yang canggih dan sistematis untuk mendukung pengambilan keputusan bisnis yang efektif dan berbasis bukti. Penelitian ini bertujuan untuk menganalisis data penjualan ritel dengan memanfaatkan Retail Sales Dataset yang diperoleh dari platform Kaggle, yang terdiri dari 100.000 catatan transaksi dan merepresentasikan karakteristik transaksi ritel secara luas. Fokus utama penelitian ini adalah melakukan klasifikasi kategori produk serta memprediksi segmen pelanggan, termasuk mengidentifikasi pelanggan dengan tingkat pengeluaran tinggi (high spender), berdasarkan atribut demografis seperti usia dan jenis kelamin, serta berbagai fitur yang berkaitan dengan transaksi pembelian. Metodologi penelitian meliputi tahapan pra-pemrosesan data, pengkodean label, serta rekayasa fitur (feature engineering) untuk menghasilkan variabel tambahan, antara lain Age_Group, Is_Holiday, dan Spender_Group, yang diharapkan dapat meningkatkan kemampuan prediksi model. Beberapa algoritma pembelajaran mesin, yaitu Decision Tree, Random Forest, dan XGBoost, diimplementasikan dan dievaluasi untuk membandingkan kinerja masing-masing model. Hasil eksperimen menunjukkan bahwa klasifikasi kategori produk secara multikelas menghasilkan tingkat akurasi yang relatif rendah, yaitu berkisar antara 27% hingga 34%. Temuan ini mengindikasikan tingginya kompleksitas data ritel serta perlunya optimasi model, penyeimbangan kelas data, dan penyempurnaan fitur lebih lanjut guna meningkatkan performa prediksi di masa mendatang.

Kata Kunci: Penyetelan Hyperparameter; Prediksi High Spender; Prediksi Penjualan Ritel; Random Forest; Rekayasa Fitur.

1. LATAR BELAKANG

Perkembangan teknologi informasi mendorong meningkatnya jumlah data yang dihasilkan dari berbagai aktivitas bisnis, termasuk pada sektor ritel. Data transaksi penjualan yang terus bertambah setiap harinya mengandung informasi penting yang dapat dimanfaatkan untuk memahami perilaku konsumen serta mendukung pengambilan keputusan strategis.

Naskah Masuk: 28 November 2025; Revisi: 26 Desember 2025; Diterima: 23 Januari 2026; Tersedia: 28 Januari 2026

Namun, tanpa proses pengolahan dan analisis yang tepat, data tersebut hanya akan menjadi arsip yang tidak memberikan nilai tambah bagi perusahaan.

Data Science hadir sebagai disiplin ilmu yang berfokus pada pengolahan dan analisis data untuk menghasilkan informasi yang bermakna. Dalam praktiknya, *Data Science* tidak dapat dipisahkan dari *Data Engineering* yang berperan dalam memastikan data tersimpan secara terstruktur, bersih, dan siap digunakan untuk keperluan analitik maupun pembelajaran mesin. Proses seperti ekstraksi, transformasi, dan pemuatan data menjadi fondasi utama agar data dapat dimanfaatkan secara optimal.

Penjualan merupakan indikator penting dalam menilai kinerja suatu bisnis, khususnya pada sektor usaha ritel. Aktivitas penjualan mencerminkan tingkat permintaan pasar serta kemampuan perusahaan dalam memenuhi kebutuhan konsumen. Bisnis ritel sendiri merupakan kegiatan penjualan barang secara langsung kepada konsumen akhir, baik melalui toko fisik maupun *platform* digital, yang terus berkembang seiring perubahan gaya hidup masyarakat dan kemajuan teknologi.

Dalam analisis penjualan ritel, karakteristik demografis pelanggan seperti usia dan jenis kelamin, serta atribut transaksi seperti jumlah pembelian dan harga produk, dapat digunakan untuk mengidentifikasi pola penjualan. Salah satu pendekatan yang umum digunakan untuk menganalisis pola tersebut adalah metode klasifikasi menggunakan algoritma *machine learning*. *Decision Tree* menjadi salah satu algoritma yang banyak digunakan karena mampu menyajikan proses pengambilan keputusan secara sistematis dan mudah dipahami.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis data penjualan ritel dengan menerapkan metode *Decision Tree* dan *Random Forest*. Analisis dilakukan untuk mengidentifikasi pola penjualan dan segmentasi pelanggan berdasarkan karakteristik transaksi. Diharapkan hasil penelitian ini dapat memberikan gambaran yang lebih jelas mengenai perilaku konsumen serta menjadi referensi dalam pengambilan keputusan bisnis berbasis data.

2. KAJIAN TEORITIS

Data Engineering

Data Engineering merupakan bidang yang berfokus pada pengelolaan infrastruktur data agar data dapat digunakan secara efektif dalam proses analisis dan pengambilan keputusan. Aktivitas dalam *Data Engineering* mencakup perancangan sistem penyimpanan data, pemrosesan data mentah, serta penyediaan data yang siap digunakan untuk analitik dan

machine learning. Peran *Data Engineer* menjadi krusial karena bertanggung jawab memastikan kualitas, konsistensi, dan ketersediaan data bagi pengguna akhir.

Pohon Keputusan

Decision Tree adalah metode klasifikasi yang menyusun proses pengambilan keputusan dalam bentuk struktur bercabang. Setiap cabang merepresentasikan kondisi tertentu yang mengarah pada keputusan atau kelas tertentu. Algoritma ini banyak digunakan karena hasil klasifikasinya mudah diinterpretasikan dan dapat menjelaskan alasan di balik suatu keputusan. Dalam konteks analisis penjualan, *Decision Tree* dapat membantu mengidentifikasi faktor-faktor yang paling berpengaruh terhadap perilaku pembelian pelanggan.

Extract, Transform, Loading

Extract, Transform, Load (ETL) merupakan proses penting dalam pengelolaan data yang bertujuan untuk memindahkan data dari berbagai sumber ke dalam sistem penyimpanan terpusat seperti data *warehouse*. Tahap *extract* berfungsi untuk mengambil data dari sumber, tahap *transform* digunakan untuk membersihkan dan menyesuaikan data sesuai kebutuhan, sedangkan tahap *load* bertujuan untuk menyimpan data yang telah diolah ke dalam sistem tujuan. Proses *ETL* memastikan data berada dalam kondisi yang siap digunakan untuk analisis lebih lanjut.

Dataset

Dataset adalah kumpulan data terstruktur yang disusun dalam bentuk baris dan kolom, di mana setiap baris merepresentasikan satu entri data dan setiap kolom menunjukkan atribut tertentu. Dataset menjadi komponen utama dalam penelitian berbasis data karena kualitas dataset sangat memengaruhi hasil analisis dan performa model yang digunakan. Dataset dapat diperoleh dari berbagai sumber, seperti transaksi bisnis, pengumpulan manual, maupun platform penyedia data daring.

3. METODE PENELITIAN

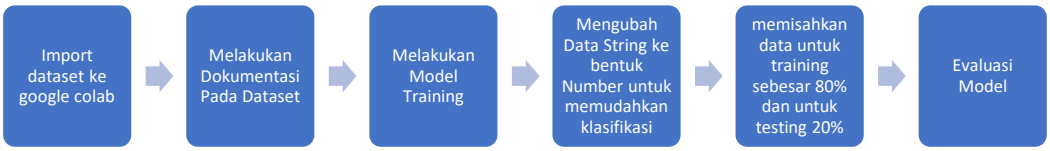
Dataset yang digunakan berjumlah 1000 data dimana terdiri atas 9 kolom yaitu :

id, date, customer id, gender, age, product category, quantity, price per unit, total amount.

kemudian kita tentukan fitur dan label dari dataset tersebut, didapat:

Label : *product category*

Fitur : *age, gender, quantity, price per unit*



Gambar 1. Flowchart Metode Pengujian.

Adapun dari 100000 data tersebut, 80% data digunakan sebagai model untuk di *training* dan 20% data digunakan sebagai *testing*. Dari 20% *data testing* didapat hasil sebagai berikut

Tabel 1. Tabel Hasil Evaluasi Data Testing.

Akurasi Decision Tree: 84.00%
Akurasi XGBoost: 79.00%
Detailed Report:

	Precision	**Recall**	**F1-score**	**Support**
Non High Spender	0.87	0.80	0.83	130
High Spender	0.68	0.77	0.72	70
Accuracy			0.79	200
Macro avg	0.77	0.79	0.78	200
Weighted avg	0.80	0.79	0.79	200

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset penjualan retail yang diperoleh dari *Kaggle* dengan jumlah 1000 data transaksi dan terdiri dari 9 atribut utama, yaitu *Transaction ID, Date, Customer ID, Gender, Age, Product Category, Quantity, Price per Unit,* dan *Total Amount.* Dataset ini dianalisis menggunakan metode *Decision Tree* dan *XGBoost* untuk melakukan klasifikasi dan segmentasi pelanggan berdasarkan pola transaksi.

Tahap awal yang dilakukan adalah analisis kualitas data. Berdasarkan hasil pengecekan, tidak ditemukan data kosong (*missing value*) maupun data ganda (*duplicate data*). Hal ini menunjukkan bahwa dataset berada dalam kondisi yang baik dan dapat langsung digunakan untuk proses pemodelan tanpa perlu proses pembersihan data lanjutan.

Tabel 2. Tabel Kolom dan Tipe Data.

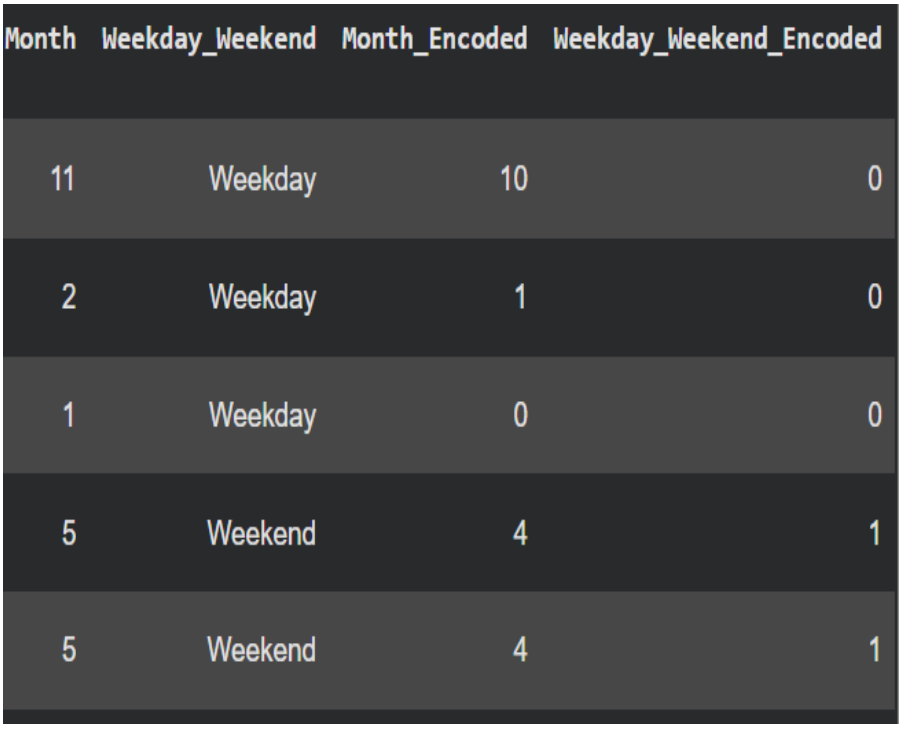
#	**Column**	**Non-null count**	**Dtype**
0	Transaction ID	1000 non-null	Int64
1	Date	1000 non-null	Object
2	Customer ID	1000 non-null	Object
3	Gender	1000 non-null	Object
4	Age	1000 non-null	Int64
5	Product Category	1000 non-null	Object
6	Quantity	1000 non-null	Int64
7	Price per Unit	1000 non-null	Int64
8	Total Amount	1000 non-null	Int64

Setelah melakukan pengecekan struktur dataset, tahap selanjutnya adalah memeriksa keberadaan nilai kosong (*missing values*). Pengecekan ini dilakukan dengan mengonversi dataset ke dalam bentuk *DataFrame* dan menghitung jumlah nilai kosong pada setiap kolom. Berdasarkan hasil pemeriksaan tersebut, seluruh kolom pada dataset, yaitu *Transaction ID*, *Date*, *Customer ID*, *Gender*, *Age*, *Product Category*, *Quantity*, *Price per Unit*, dan *Total Amount*, tidak memiliki nilai kosong. Dengan demikian, dataset dinyatakan lengkap dan dapat digunakan untuk tahap analisis selanjutnya. Setelah memastikan tidak terdapat data kosong, dilakukan pengujian lanjutan untuk memeriksa keberadaan data ganda (*duplicate values*).

Selanjutnya dilakukan pengecekan terhadap keberadaan data ganda (*duplicate values*) dengan menghitung jumlah baris yang terduplikasi dalam dataset. Hasil pemeriksaan menunjukkan bahwa tidak ditemukan data ganda, yang ditunjukkan oleh nilai hasil pengecekan sebesar nol. Dengan demikian, dataset dinyatakan bebas dari duplikasi dan layak digunakan untuk proses analisis dan pemodelan selanjutnya.

Melakukan Dokumentasi Dataset

Selanjutnya dilakukan pra-proses data, khususnya pada atribut yang masih berbentuk teks seperti *Gender* dan *Product Category*. Atribut tersebut diubah ke dalam bentuk numerik menggunakan teknik Label *Encoding* agar dapat diproses oleh algoritma *machine learning*. Selain itu, atribut *Date* dikonversi ke format *datetime* untuk mengekstraksi informasi tambahan seperti bulan transaksi, hari kerja atau akhir pekan, kuartal, serta status hari libur.



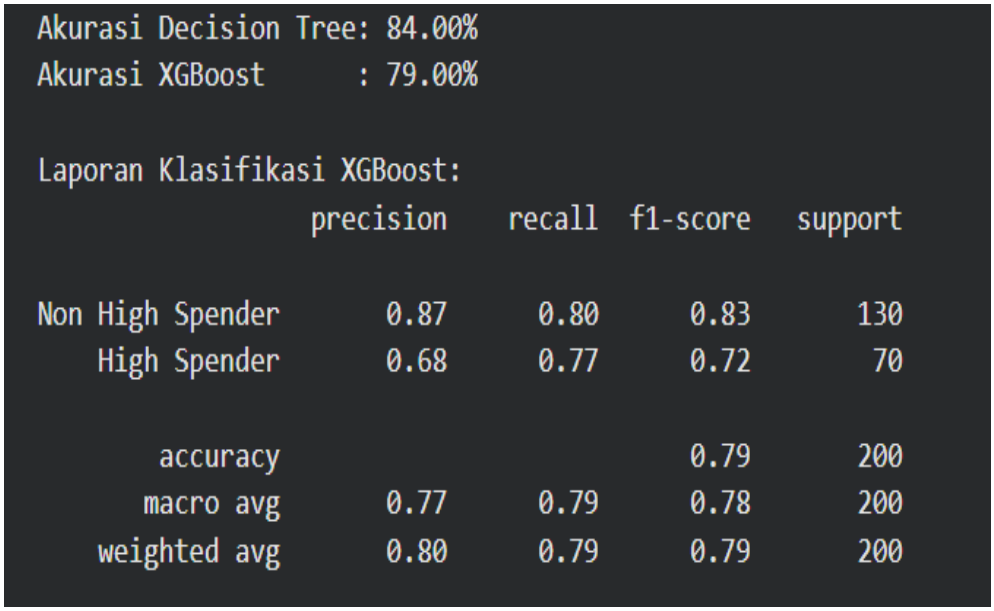
Month	Weekday_Weekend	Month_Encoded	Weekday_Weekend_Encoded
11	Weekday	10	0
2	Weekday	1	0
1	Weekday	0	0
5	Weekend	4	1
5	Weekend	4	1

Gambar 2. Dokumentasi Dataset.

Untuk meningkatkan kemampuan model dalam mengenali pola penjualan, dilakukan rekayasa fitur (*feature engineering*). Beberapa fitur baru yang ditambahkan antara lain *Age Group*, *Is Holiday*, *Price per Item*, serta *Spender Group* yang mengelompokkan pelanggan berdasarkan tingkat pengeluaran. Penambahan fitur-fitur ini bertujuan untuk merepresentasikan perilaku belanja pelanggan secara lebih jelas dibandingkan hanya menggunakan data demografis.

Setelah proses pra-proses dan rekayasa fitur selesai, data dibagi menjadi 80% data *training* (800 data) dan 20% data *testing* (200 data). Pada tahap awal pengujian dengan target klasifikasi kategori produk secara *multiclass*, model menghasilkan akurasi yang relatif rendah, yaitu sekitar 27–33%. Rendahnya akurasi ini disebabkan oleh keterbatasan fitur dalam merepresentasikan hubungan langsung antara karakteristik pelanggan dan kategori produk.

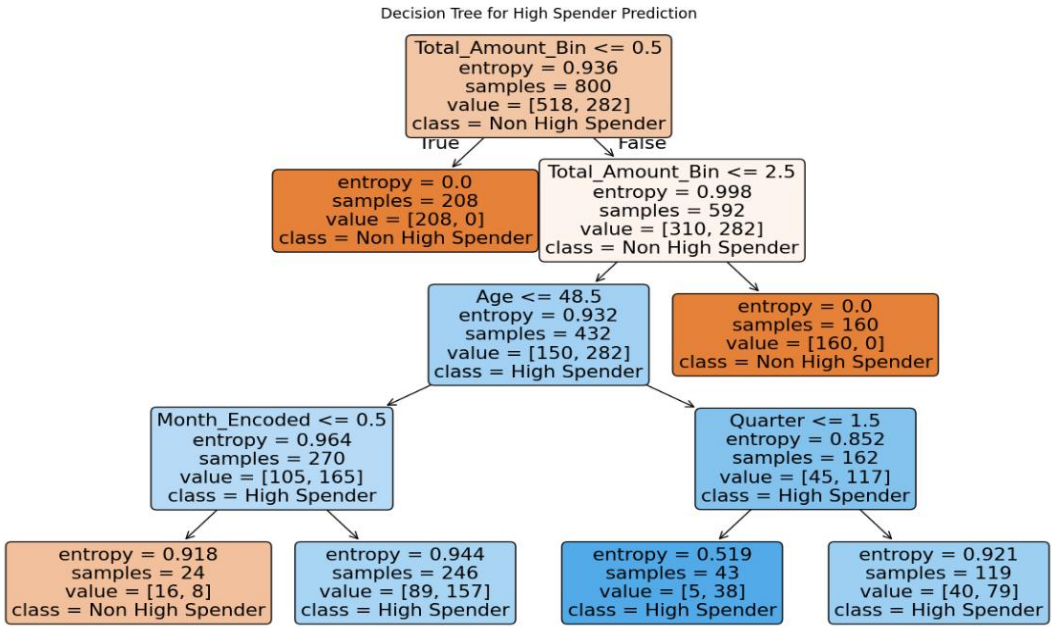
Untuk mengatasi permasalahan tersebut, target klasifikasi disederhanakan menjadi klasifikasi biner, yaitu pelanggan *High Spender* dan *Non-High Spender*. Dengan penyederhanaan target serta penambahan fitur transaksi yang lebih relevan, performa model meningkat secara signifikan. Model *Decision Tree* berhasil mencapai akurasi sebesar 84%, sedangkan model *XGBoost* memperoleh akurasi sebesar 79%.



```
Akurasi Decision Tree: 84.00%
Akurasi XGBoost      : 79.00%

Laporan Klasifikasi XGBoost:
                precision    recall  f1-score   support

Non High Spender       0.87      0.80      0.83       130
    High Spender       0.68      0.77      0.72        70

        accuracy                           0.79       200
       macro avg       0.77      0.79      0.78       200
    weighted avg       0.80      0.79      0.79       200
```

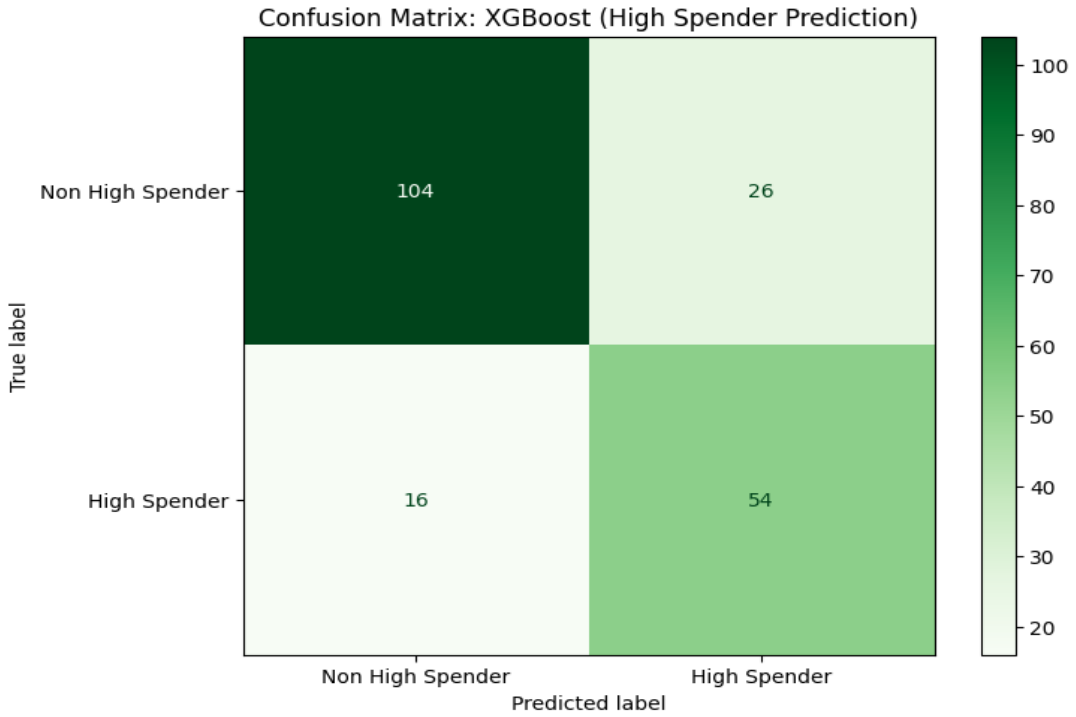
Gambar 3. Hasil Evaluasi.

Hasil evaluasi menunjukkan bahwa *Decision Tree* memiliki performa terbaik dalam penelitian ini. Model ini mampu membentuk aturan keputusan yang jelas berdasarkan fitur-fitur transaksi, sehingga pola pengeluaran pelanggan dapat dipelajari dengan lebih efektif. Selain itu, struktur pohon keputusan memudahkan interpretasi hasil karena setiap keputusan dapat ditelusuri secara visual.



Gambar 4. Struktur Pohon.

Sementara itu, meskipun akurasi *XGBoost* sedikit lebih rendah, model ini tetap menunjukkan performa yang cukup baik, terutama dalam menjaga keseimbangan antara *precision* dan *recall* pada klasifikasi pelanggan *High Spender.* Perbedaan kinerja antara kedua model dipengaruhi oleh karakteristik dataset serta kesesuaian metode terhadap fitur yang telah didiskretisasi.



Gambar 5. *XGBoost.*

5. KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil analisis dan implementasi model *Decision Tree* pada dataset penjualan retail, dapat disimpulkan bahwa Dataset yang digunakan terdiri dari 1.000 entri dengan 9 kolom, tanpa ditemukan data ganda maupun data kosong. Data dibagi menjadi 80% untuk pelatihan (800 data) dan 20% untuk pengujian (200 data). Pada tahap awal, model menghasilkan akurasi pengujian sebesar 27%. Setelah dilakukan optimasi fitur dan penyederhanaan target menjadi segmentasi pelanggan (*High Spender*), performa model meningkat signifikan, di mana *Decision Tree* mencapai akurasi tertinggi sebesar 84% dan *XGBoost* sebesar 79%.

Saran

Adapun saran yang dapat kami sampaikan untuk penelitian dikemudian hari yaitu Penelitian selanjutnya disarankan melakukan eksperimen lanjutan pada parameter *Decision Tree*, seperti pengaturan *max_depth* atau penggunaan kriteria selain *entropy*, untuk meningkatkan kemampuan generalisasi model. Selain itu, dapat dicoba algoritma klasifikasi lain seperti *Random Forest* (Pal, 2005) atau *Gradient Boosting* (Bentéjac et al., 2020) yang umumnya memberikan akurasi lebih baik pada data retail yang kompleks. Penelitian juga perlu memperhatikan keseimbangan distribusi antar kategori produk, karena rendahnya nilai *recall* pada kelas tertentu dapat disebabkan oleh ketidakseimbangan jumlah sampel.

DAFTAR REFERENSI

Aditya, M. A., Mulyana, R. D., Eka, I. P., & Widianto, S. R. (2020). Penggabungan teknologi untuk analisa data berbasis data science. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, *1*(1), 51–56.

Agustina, A., Tukino, T., Huda, B., & Novalia, E. (2025). Prediksi volume penjualan gadget berdasarkan promo dan channel penjualan menggunakan random forest. *JUSIFOR: Jurnal Sistem Informasi dan Informatika*, *4*(1), 85–91. https://doi.org/10.70609/jusifor.v4i1.6962

Apriliyani, E., & Salim, Y. (2022). Analisis performa metode klasifikasi Naive Bayes classifier pada unbalanced dataset. *Indonesian Journal of Data and Science*, *3*(2), 47–54.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Kurniawan, M. A., Syauqi, G. Z., Safriyanti, M., Azmie, F. U., & Setiawan, A. (2025). Prediksi pendapatan penjualan di Indomaret menggunakan algoritma random forest regression. *JSI (Jurnal Sistem Informasi) Universitas Suryadarma*, *12*(2), 93–99. https://doi.org/10.35968/jsi.v12i2.1478

Kurniawan, R. D., Sukarman, D. N. D., Rumaropen, K. W., & Allo, C. B. G. (2025). Analisis komparatif algoritma decision tree dan random forest untuk klasifikasi penjualan produk pada dataset superstore. *STATMAT: Jurnal Statistika dan Matematika, 7*(2), 94–103. https://doi.org/10.32493/sm.v7i2.48856

Martinus, H. (2011). Analisis industri ritel nasional. *Humaniora*, *2*(2), 1309–1321. https://doi.org/10.21512/humaniora.v2i2.3193

Nahda, Z., Rahma, A., Al Fath, L. H., & Suhairi, S. (2022). Konsep pohon keputusan. *VISA: Journal of Vision and Ideas*, *2*(1), 135–142. https://doi.org/10.47467/visa.v2i1.961

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Riza, N., Aulia, M. Z., Kolin, P. B., & Mustaqim, K. (2025). Analisis faktor pengaruh terhadap penghasilan profesi data engineer menggunakan metode regresi linear berganda. *Jurnal Informatika dan Teknik Elektro Terapan*, *13*(1), 2830–7062. https://doi.org/10.23960/jitet.v13i1.5740

Selay, A., Andgha, G. D., Alfarizi, M. A., Izdhihar, M., Wahyudi, B., Falah, M. N., & Khaira, M. (2023). Sistem informasi penjualan. *Karimah Tauhid*, *2*(1), 232–237. https://doi.org/10.30997/karimahtauhid.v2i1.7746

Soemarso, S. R. (1983). *Akuntansi: Suatu pengantar*. Lembaga Penerbit Fakultas Ekonomi Universitas Indonesia. https://books.google.co.id/books?id=JbZaAQAACAAJ

Verdiyanto, R., Hartanti, D., & Purwanto, E. (2025). Pengembangan aplikasi point of sales untuk prediksi penjualan harian usaha minuman menggunakan algoritma random forest regression. *Infotek: Jurnal Informatika dan Teknologi*, *8*(1), Article 28386. https://doi.org/10.29408/jit.v8i1.28386

Warnars, S. (2009). Desain ETL dengan contoh kasus perguruan tinggi. *Jurnal Informatika*, *10*(2), 86–93.

Yao, B. (2023). *Walmart sales prediction based on decision tree, random forest, and k neighbors regressor*. *Highlights in Business, Economics and Management, 5*, 330–335. https://doi.org/10.54097/hbem.v5i.5100

Zulfia, A., Ilfa, T. N., Damia, Z., Sukiman, T. S. A., & Karima, A. (2025). AI decision support for demand forecasting and retail stock using random forest. *Brilliance: Research of Artificial Intelligence, 5*(2), Article 5901. https://doi.org/10.47709/brilliance.v5i2.5901