StatMat
Jurnal Statistika dan Matematika

STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

ANALISIS KOMPARATIF ALGORITMA *DECISION TREE* DAN *RANDOM FOREST* UNTUK KLASIFIKASI PENJUALAN PRODUK PADA DATASET SUPERSTORE

Rezza Debby Kurniawan[1*], Dian Natasia Dewanti Sukarman[2], Kartini Wika Rumaropen[3], dan Caecilia Bintang Girik Allo[4]

[1,2,3,4]Universitas Cenderawasih
[*]Email Korespondensi: caecilia.bintang@fmipa.uncen.ac.id

ABSTRACT

This study aims to compare the performance of two machine learning classification algorithms, namely Decision Tree and Random Forest, in classifying product sales based on the Sample Superstore dataset. The classification objective is to categorize products into two groups: "popular" and "less popular," determined by comparing the product's sold quantity with the dataset's average. The research begins with preprocessing steps such as duplicate removal, handling categorical variables using label encoding, standardizing numerical features, and constructing a binary label variable. The classification models are then trained and evaluated using a confusion matrix and performance metrics including accuracy, precision, recall, and F1-score. Hyperparameter tuning is applied to both models using GridSearchCV with five-fold cross-validation. The results indicate that the Decision Tree algorithm achieved the highest accuracy of 73.1%, slightly outperforming Random Forest, which reached 72.7% after tuning. Despite its reputation as a more robust ensemble method, Random Forest did not significantly outperform Decision Tree in this context, likely due to the relatively simple structure and balance of the dataset. This finding supports the notion that algorithm selection should consider data characteristics and preprocessing effectiveness.
Keywords: Decision Tree, Random Forest, Classification, Product Sales, Machine Learning.

ABSTRAK

Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma klasifikasi *machine learning*, yaitu *Decision Tree* dan *Random Forest*, dalam mengklasifikasikan produk berdasarkan data penjualan pada dataset *Sample* Superstore. Tujuan klasifikasi adalah untuk mengelompokkan produk ke dalam dua kategori, yaitu "laris" dan "tidak laris", berdasarkan perbandingan antara nilai *Quantity* dengan nilai rata-rata dataset. Penelitian diawali dengan tahapan pra-pemrosesan seperti penghapusan data duplikat, pengolahan variabel kategorikal dengan label encoding, standardisasi fitur numerik, dan pembentukan variabel target biner. Model klasifikasi dilatih dan dievaluasi menggunakan *confusion matrix* serta metrik performa seperti akurasi, *precision*, *recall*, dan *F1-score*. *Tuning* parameter dilakukan pada kedua model menggunakan *GridSearchCV* dengan validasi silang lima lipat. Hasil evaluasi menunjukkan bahwa algoritma *Decision Tree* memperoleh akurasi tertinggi sebesar 73,1%, sedikit lebih baik dibandingkan *Random Forest* yang mencapai akurasi 72,7% setelah dilakukan *tuning*. Meskipun *Random Forest*

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

dikenal sebagai algoritma *ensemble* yang lebih stabil, pada kasus ini performanya tidak menunjukkan peningkatan signifikan. Temuan ini memperkuat pentingnya pemilihan algoritma yang disesuaikan dengan karakteristik dan struktur data.

Kata kunci: *Decision Tree*, *Random Forest*, Klasifikasi, Penjualan Produk, *Machine Learning.*

ARTICLE INFO	
Submission received: 03 May 2025	Accepted: 31 August 2025
Revised: 27 May 2025	Published: 31 August 2025
Available on: https://doi.org/10.32493/sm.v7i2.xxxxx	

1. PENDAHULUAN

Perkembangan teknologi informasi telah mengubah lanskap bisnis modern, termasuk dalam sektor ritel. Perusahaan *retail* saat ini tidak hanya dituntut untuk menjual produk secara efisien, tetapi juga mampu memanfaatkan data transaksi sebagai aset strategis untuk pengambilan keputusan. Salah satu pendekatan yang banyak digunakan dalam mengolah data transaksi adalah data mining, khususnya metode klasifikasi yang berfungsi untuk memetakan entitas data ke dalam kelompok-kelompok tertentu berdasarkan atribut-atribut yang dimiliki.

Dalam konteks penjualan, klasifikasi dapat digunakan untuk mengelompokkan produk berdasarkan tingkat kelarisannya. Informasi ini sangat penting dalam perencanaan stok, promosi, dan manajemen inventori. Dengan mengetahui produk mana yang tergolong laris dan tidak laris, perusahaan dapat mengambil keputusan lebih tepat terkait pengadaan, distribusi, hingga penetapan harga.

Dua algoritma klasifikasi yang umum digunakan dalam analisis data penjualan adalah *Decision Tree* dan *Random Forest*. *Decision Tree* adalah bentuk sederhana dari metode klasifikasi untuk sejumlah kelas yang terbatas, di mana simpul akar dan simpul internal diidentifikasi dengan nama atribut, sementara cabang-cabangnya dilabeli berdasarkan kemungkinan nilai atribut, dan simpul daunnya menunjukkan kelas-kelas yang berbeda (Eska, 2016). Di sisi lain, *Random Forest* merupakan algoritma *Ensemble Learning* yang menggunakan pendekatan *Bagging* untuk membangun banyak pohon keputusan dari sampel acak yang dihasilkan, kemudian melatih masing-masing pohon tersebut (Sun et al., 2020).

Penelitian terdahulu telah membuktikan efektivitas kedua algoritma ini. Firnanda et al. (2023) melakukan studi komparatif pada klasifikasi produk supermarket menggunakan algoritma *Decision Tree* dan *Random Forest*, dan menemukan bahwa *Random Forest* memiliki akurasi tertinggi sebesar 98% pada dataset berskala besar.

Namun demikian, tidak semua dataset menunjukkan pola yang sama. Keunggulan algoritma tertentu sangat bergantung pada karakteristik data, seperti jumlah atribut, keseimbangan kelas, dan kompleksitas relasi antar fitur. Oleh karena itu, penelitian ini

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

bertujuan untuk membandingkan kembali performa *Decision Tree* dan *Random Forest* dalam konteks klasifikasi produk berdasarkan data penjualan pada dataset *Sample* Superstore, yang memiliki struktur data yang relatif sederhana namun representatif terhadap data penjualan *retail* secara umum.

Tujuan dari penelitian ini adalah untuk (1) mengimplementasikan algoritma *Decision Tree* dan *Random Forest* dalam klasifikasi kelarisan produk menggunakan dataset *Sample* Superstore, (2) membandingkan performa kedua algoritma berdasarkan metrik evaluasi klasifikasi seperti akurasi, *precision, recall,* dan *F1-score*, dan (3) menganalisis pengaruh *preprocessing* terhadap hasil klasifikasi, khususnya pada proses transformasi data kategorikal dan numerik. Diharapkan hasil dari penelitian ini dapat memberikan gambaran empiris mengenai efektivitas kedua algoritma pada skenario data *retail*, serta menjadi referensi dalam pemilihan model klasifikasi yang sesuai dengan karakteristik dataset.

2. METODOLOGI

Penelitian ini menggunakan metode eksperimen komparatif dengan pendekatan kuantitatif. Proses penelitian dilakukan melalui tahapan: eksplorasi dan pemrosesan data, pembentukan variabel target, pemodelan dengan dua algoritma klasifikasi, evaluasi performa model, dan *tuning hyperparameter*. Evaluasi performa model dilakukan menggunakan metrik akurasi, *precision, recall,* dan *F1-score*.

2.1. Dataset dan Variabel

Data yang digunakan dalam penelitian ini berasal dari dataset publik yaitu Superstore yang diperoleh dari platform Kaggle. Dataset ini terdiri dari 9.994 baris dan 13 kolom, yang memuat informasi penjualan produk *retail* di berbagai kota di Amerika Serikat. Atribut yang tersedia antara lain: *Sales, Quantity, Profit, Discount, Category, Sub-Category, Region, Segment,* serta informasi lokasi (*City, State, Postal Code*).

Variabel target (Label) dibentuk dengan membandingkan nilai *Quantity* masing-masing produk terhadap nilai rata-rata *Quantity* dari seluruh dataset. Produk diklasifikasikan sebagai "laris" (Label = 1) jika *Quantity* ≥ rata-rata, dan "tidak laris" (Label = 0) jika sebaliknya. Pembentukan target seperti ini dimaksudkan untuk menghasilkan dua kelas yang relatif seimbang.

2.2. Algoritma *Decision Tree*

Decision Tree adalah algoritma klasifikasi yang membangun struktur pohon keputusan berdasarkan pemilihan atribut yang memberikan informasi paling tinggi dalam membagi data ke dalam kelas target. Setiap node pada pohon mewakili atribut, cabang mewakili keputusan atas nilai atribut, dan daun mewakili kelas akhir. Langkah-langkah algoritma *Decision Tree* sebagai berikut:

1. Hitung *Entropy* seluruh dataset.
2. Hitung *Entropy* masing-masing atribut terhadap kelas target.
3. Hitung Information Gain dari masing-masing atribut.
4. Pilih atribut dengan *Gain* tertinggi sebagai *node* cabang.
5. Bagi dataset berdasarkan nilai atribut tersebut.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

6. Ulangi proses secara rekursif pada subset data sampai semua data diklasifikasikan atau memenuhi kriteria berhenti.

Rumus *entropy* digunakan untuk menghitung tingkat homogenitas atribut *A* pada suatu sampel data *S* (Sinambela, 2022). Rumus *entropy* dapat dituliskan sebagai:

$$Entropy(S) = -\sum_{i=1}^{c} pi\ log_2(pi) \quad (1)$$

di mana:
- S = himpunan data,
- c = jumlah kelas,
- p_i = proporsi data pada kelas ke-i

Rumus *Information Gain*:

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

di mana:
- A = atribut,
- v = nilai dari atribut A,
- S_v = subset data yang memiliki nilai v pada atribut A.

Atribut dengan *Gain* tertinggi akan dipilih sebagai akar (*root*) dari pohon, dan proses dilanjutkan hingga seluruh data terkategori atau kondisi berhenti terpenuhi (misalnya: kedalaman maksimum tercapai atau entropi nol).

2.3. Algoritma *Random Forest*

Random Forest merupakan algoritma klasifikasi berbasis *ensemble learning* yang menggabungkan banyak *Decision Tree* untuk meningkatkan performa klasifikasi. Algoritma ini membangun sejumlah pohon keputusan secara paralel menggunakan teknik *bootstrap sampling* (pengambilan sampel acak dengan pengembalian) dan pemilihan subset fitur secara acak. Langkah-langkah algoritma *Random Forest* sebagai berikut:

1. Buat *n* buah sampel *bootstrap* dari data pelatihan.
2. Untuk setiap sampel, bangun *Decision Tree* berdasarkan subset acak dari fitur.
3. Lakukan klasifikasi pada setiap *Decision Tree* terhadap data uji.
4. Hasil akhir ditentukan berdasarkan mayoritas suara (*voting*) dari semua pohon.

Rumus *Voting* Klasifikasi:

$$Prediction = arg\ \max_k \sum_{i=1}^{n} 1(Tree_i(x) = k) \quad (3)$$

di mana:
- n = jumlah pohon,
- $Tree_i(x)$ = hasil prediksi dari pohon ke-i,
- k = kelas target (0 atau 1),

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

- $\mathbb{1}$ = fungsi indikator.

Kelebihan *Random Forest* terletak pada kemampuannya mengurangi *overfitting* dan meningkatkan generalisasi model karena variasi antar pohon dan fitur yang digunakan.

2.4. Evaluasi Model

Kinerja model machine learning pada penelitian ini dinilai dengan menggunakan *confusion matrix*. *Confusion matrix* menyajikan ringkasan jumlah prediksi yang tepat maupun tidak tepat, yang diklasifikasikan menurut masing-masing kelas. Evaluasi hasil prediksi dilakukan melalui beberapa metrik, yaitu akurasi, *precision*, *recall,* dan *F1-Score* (Fadli & Saputra, 2023).
Akurasi:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall:

$$Recall = \frac{TP}{TP + FP} \tag{6}$$

F1-Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

Keterangan:
- *TP = True Positive*, *FP = False Positive*,
- *TN = True Negative*, *FN = False Negative*.

Agar hasil lebih optimal, *tuning hyperparameter* dilakukan menggunakan teknik pencarian *GridSearchCV* dengan validasi silang (*cross-validation*) sebanyak 5 kali lipat. Parameter utama yang disesuaikan meliputi jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), serta jumlah minimum sampel pada daun (min_samples_leaf).

3. HASIL DAN PEMBAHASAN

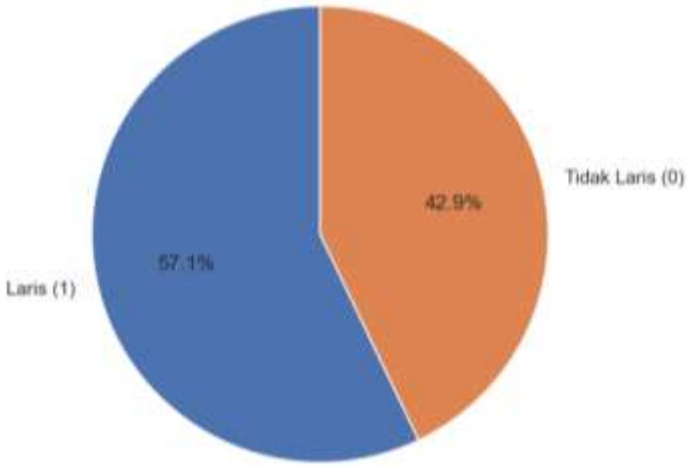
Penelitian ini dimulai dengan eksplorasi dan pembersihan data dari dataset *Sample* Superstore. Dataset ini berisi informasi transaksi penjualan *retail* di Amerika Serikat, terdiri dari 9.994 baris dan 13 kolom fitur. Beberapa atribut yang relevan untuk analisis klasifikasi antara lain *Quantity, Sales, Profit, Discount, Category, Sub-Category, Region,* dan *Segment*.

3.1. Pra-pemrosesan Data

Langkah awal dalam pra-pemrosesan adalah penghapusan data duplikat yang teridentifikasi sebanyak 17 baris. Tidak ditemukan adanya nilai kosong (*missing values*) dalam dataset.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

Selanjutnya, dilakukan pembentukan variabel target (Label) untuk klasifikasi produk laris dan tidak laris. Rata-rata *Quantity* dari seluruh data dihitung sebagai ambang batas. Produk dengan *Quantity* lebih besar atau sama dengan rata-rata dikategorikan sebagai "laris" (Label = 1), sedangkan sisanya dikategorikan sebagai "tidak laris" (Label = 0). Hasilnya, distribusi kelas target cukup seimbang, dengan proporsi produk laris sekitar 57%.



Gambar 1. *Pie Chart* – Distribusi Label Produk Laris *vs* Tidak Laris

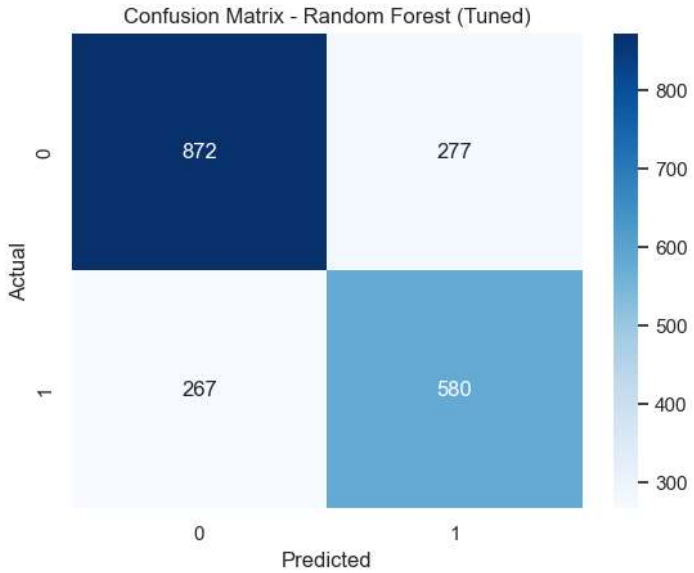
Proses berikutnya adalah transformasi fitur kategorikal. Dataset memiliki beberapa fitur *non-numeric* seperti *Category, Sub-Category, Region,* dan *Segment*. Agar fitur-fitur ini dapat digunakan dalam algoritma klasifikasi, dilakukan proses transformasi menggunakan teknik Label *Encoding*. Teknik ini mengubah setiap kategori menjadi angka unik tanpa menambah jumlah dimensi. Misalnya, nilai "*Consumer*", "*Corporate*", dan "*Home Office*" pada fitur Segment dikonversi menjadi 0, 1, dan 2.

Fitur numerik seperti *Sales, Profit,* dan *Discount* kemudian dilakukan standardisasi menggunakan metode *Standard Scaler*. Proses ini bertujuan untuk menyamakan skala antar fitur agar tidak terjadi dominasi nilai besar (misalnya *Profit* yang bernilai ribuan) terhadap algoritma klasifikasi. Data kemudian dibagi menjadi dua bagian: 80% untuk pelatihan dan 20% untuk pengujian.

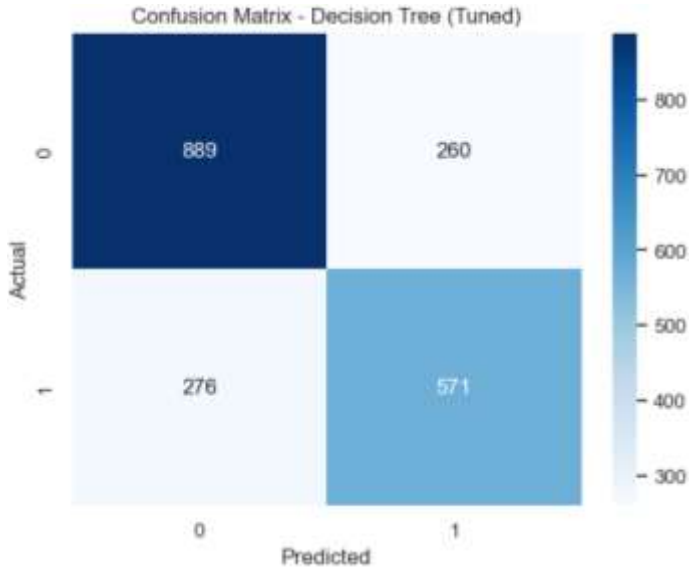
3.2. Hasil Klasifikasi Awal

Decision Tree dan *Random Forest* masing-masing dibangun menggunakan parameter *default* terlebih dahulu. Hasil evaluasi pada data uji menunjukkan bahwa *Decision Tree* mencapai akurasi awal sebesar 73,1%, sementara *Random Forest* memperoleh akurasi 71,7%. *Decision Tree* juga menunjukkan keunggulan pada metrik *precision, recall,* dan *F1-score* dibandingkan *Random Forest*.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx



Gambar 2. *Confusion Matrix – Random Forest* (*Tuned*)



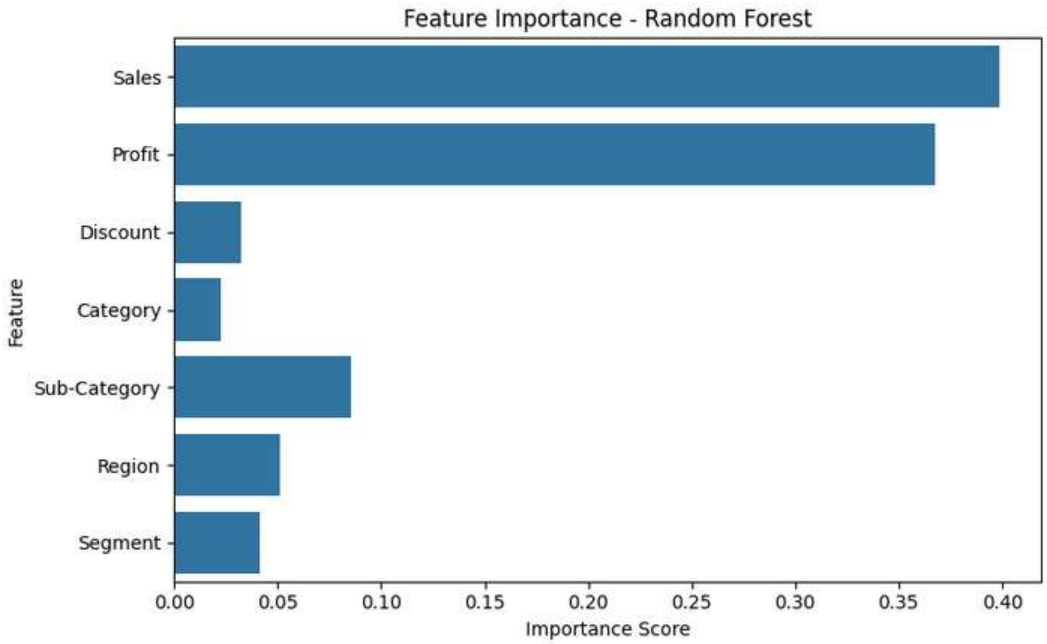
Gambar 3. *Confusion Matrix – Decision Tree* (*Tuned*)

3.3. *Tuning Hyperparameter* dan Evaluasi Lanjutan

Agar performa model dapat ditingkatkan, dilakukan *tuning hyperparameter* menggunakan teknik *GridSearchCV*. Untuk Random Forest, parameter yang disesuaikan meliputi jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan minimum sampel pada daun (*min_samples_leaf*). Untuk *Decision Tree*, dilakukan penyesuaian pada *max_depth, min_samples_split*, dan *min_samples_leaf*.

Setelah proses tuning, *Random Forest* mengalami peningkatan akurasi menjadi 72,7%, mendekati performa *Decision Tree*. Namun, hasil ini tetap menunjukkan bahwa *Decision Tree* secara keseluruhan lebih unggul dalam menangani struktur data pada dataset *Sample* Superstore.

STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx



Gambar 4. Visualisasi *feature importance* (*Random Forest*)

Visualisasi *feature importance* dari *Random Forest* menunjukkan bahwa fitur *Quantity, Sales,* dan *Profit* adalah tiga atribut paling dominan dalam menentukan hasil klasifikasi. Hal ini selaras dengan logika bisnis bahwa produk dengan nilai *Quantity* tinggi dan *Profit* stabil cenderung tergolong produk laris.

3.4. Perbandingan Performa Model

Tabel berikut merangkum hasil evaluasi akhir dari kedua algoritma klasifikasi:

Tabel 1. Perbandingan Akurasi dan Metrik Evaluasi Model

Algoritma	Akurasi	*Precision*	*Recall*	*F1-Score*
Decision Tree	0,731	0,73	0,73	0,73
Random Forest	0,717	0,72	0,72	0,72
Random Forest (*Tuned*)	0,727	0,73	0,73	0,73

Hasil ini menunjukkan bahwa pada struktur data yang relatif sederhana dan seimbang seperti pada *Sample* Superstore, algoritma *Decision Tree* memberikan performa lebih baik.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

4. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan performa algoritma *Decision Tree* dan *Random Forest* dalam mengklasifikasikan kelarisan produk berdasarkan data penjualan pada dataset *Sample* Superstore. Melalui proses pra-pemrosesan yang mencakup penghapusan duplikat, pembentukan label target, *encoding* variabel kategorikal, serta standardisasi fitur numerik, data diolah menjadi siap digunakan dalam pemodelan klasifikasi.

Hasil evaluasi menunjukkan bahwa algoritma *Decision Tree* memiliki akurasi yang lebih tinggi, yaitu sebesar 73,1%, dibandingkan *Random Forest* yang hanya mencapai 71,7% sebelum tuning dan meningkat menjadi 72,7% setelah tuning. Keunggulan *Decision Tree* ini menunjukkan bahwa pada struktur data yang relatif sederhana, model pohon tunggal dapat bekerja secara efisien dan akurat.

Visualisasi matriks kebingungan dan analisis *feature importance* memperkuat hasil tersebut, dengan fitur *Quantity, Sales,* dan *Profit* menjadi kontributor utama dalam prediksi kelarisan produk. Meskipun *Random Forest* umumnya memiliki reputasi sebagai model yang lebih stabil dan tahan *overfitting*, pada penelitian ini performanya tidak melampaui *Decision Tree*, yang justru menghasilkan hasil klasifikasi lebih optimal.

Temuan ini menegaskan bahwa pemilihan algoritma klasifikasi sebaiknya disesuaikan dengan karakteristik data yang digunakan. Model sederhana seperti *Decision Tree* masih sangat relevan dan efektif, khususnya untuk dataset *retail* yang tidak terlalu kompleks. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi algoritma lain seperti *Gradient Boosting* atau *XGBoost*, serta memperluas ruang lingkup data ke sektor ritel lain yang memiliki struktur fitur dan kelas yang lebih bervariasi.

StatMat
Jurnal Statistika dan Matematika
STATMAT (Jurnal Statistika dan Matematika)
P-ISSN: 2655-3724 E-ISSN: 2720-9881
Vol. 7, No. 2, August 2025
1st article, Pages: 94 – 103
DOI: https://doi.org/10.32493/sm.v7i2.xxxxx

5. DAFTAR PUSTAKA

Calistus, R. 2022. Superstore [Dataset]. Kaggle. https://www.kaggle.com/datasets/roopacalistus/superstore/code

Eska, J. 2016. Penerapan Data Mining untuk Prediksi Penjualan Wallpaper Menggunakan Algoritma C4.5. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 2(2), 9–13.

Fadli, M., & Saputra, R. A. 2023. Klasifikasi dan Evaluasi Performa Model Random Forest untuk Prediksi Stroke. *Jurnal Teknik*, 12(2), 72–80.

Firnanda, P. A., Shofwatillah, L., Rahma, F., & Fauzi, F. 2023. Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket. *Emerging Statistics and Data Science Journal*, 3(1), 45–54.

Sinambela, D. P., Naparin, H., Zulfadhilah, M., & Hidayah, N. 2022. Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin. *Jurnal Informasi dan Teknologi*, 5(3), 58–64.

Sun, Y., Zhang, H., Zhao, T., Zou, Z., Shen, B., & Yang, L. 2020. A New Convolutional Neural Network With Random Forest Method for Hydrogen Sensor Fault Diagnosis. *INFOTEK Mesin*, 16(1), 127–134.