Sainteks
ISSN: 2686-0546
Volume 23 No 1, April 2026
DOI: 10.30595/sainteks.v23i1.29930 (79-92)

Klasifikasi Kinerja Penjualan Produk Nike Menggunakan Algoritma Random Forest dengan Pendekatan Hold-Out dan K-Fold Cross Validation

Classification of Nike Product Sales Performance Using the Random Forest Algorithm with Hold-Out and K-Fold Cross Validation Approaches

Divta Khoirun Nisa[1*], Fida Maisa Hana[2], Saiful Ulya[3]

[1,2,3]*Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kudus, Indonesia*

[*]corr-author: 42022100016@std.umku.ac.id

ABSTRAK

Dinamika industri ritel menuntut pemanfaatan data transaksi besar untuk pengambilan keputusan strategis. Penelitian ini bertujuan mengklasifikasi kinerja penjualan produk Nike ke dalam kategori rendah, sedang, dan tinggi menggunakan algoritma *Random Forest.* Penelitian ini memberikan kontribusi melalui pengujian model menggunakan dua pendekatan, yaitu *Hold-Out* dan *K-Fold Cross Validation,* untuk menjamin stabilitas performa. *Dataset* yang digunakan merupakan data sekunder dari *Kaggle* sebanyak 9.360 baris data transaksi Nike di Amerika Serikat periode 2020-2021. Tahapan penelitian meliputi *preprocessing* data melalui *label encoding*, pembagian data, pemodelan *Random Forest*, serta evaluasi menggunakan *confusion matrix*, hasil pengujian menunjukkan bahwa model memiliki performa yang sangat tinggi, dengan tingkat akurasi pada metode *Hold-Out* mencapai 98,13%. Sementara itu, pengujian menggunakan *10-Fold Cross Validation* menghasilkan akurasi tertinggi mencapai 94,39% pada *fold* ke-4. Secara keseluruhan, nilai *weighted average precision*, *recall*, dan *F1-score* mencapai 0,98 yang membuktikan efektivitas algoritma *Random Forest* dalam memberikan klasifikasi yang akurat. Temuan ini diharapkan dapat mendukung manajemen dalam pengambilan keputusan berbasis data di sektor ritel.

Kata Kunci: ***Random Forest*, Klasifikasi Penjualan, Nike, *Hold-Out*, *10-Fold Cross Validation***

ABSTRACT

The dynamics of the retail industry demand the utilization of large transaction data for strategic decision-making. This study aims to classify the sales performance of Nike products into low, medium, and high categories using the Random Forest algorithm. In contrast to previous studies that utilized the KNN method, this research contributes by testing the model using two approaches, namely Hold-Out and 10-Fold Cross Validation, to ensure performance stability. The dataset used is secondary data from Kaggle, consisting of 9,360 rows of Nike transaction data in the United States for the 2020–2021 period. The research stages include data preprocessing through label encoding, data partitioning, Random Forest modeling, and evaluation using a confusion matrix. The test 98.13% using the Hold-Out method. Meanwhile, testing with 10-Fold Cross Validation

with the highest accuracy reaching 94.39% in the 4th fold. Overall, the weighted average precision, recall, and f1-score reached 0.98, proving the effectiveness of the Random Forest algorithm in providing accurate classifications. These findings are expected to support management in data-driven decision-making within the retail sector.

Keywords: Random Forest, Sales Classification, Nike, Hold-Out, 10-Fold Cross Validation

PENDAHULUAN

Perkembangan teknologi informasi yang cepat telah mendorong perubahan signifikan di berbagai bidang, termasuk sektor ritel. Penggunaan teknologi menjadi instrumen krusial dalam meningkatkan efisiensi operasional dan daya saing perusahaan di tengah persaingan pasar global yang kompetitif (Indrawan, 2025; Hafizh, 2023). Industri ritel bersifat sangat dinamis karena dipengaruhi oleh fluktuasi perilaku konsumen, efektivitas strategi promosi, dan inovasi produk yang berkelanjutan. Sebagai salah satu pemimpin pasar global, Nike menghasilkan data transaksi dalam volume besar yang mencerminkan pola konsumsi lintas wilayah dan waktu. Namun, tanpa pengelolaan yang tepat, data tersebut hanya menjadi arsip digital yang tidak memberikan nilai tambah bagi perusahaan.

Pemanfaatan *data mining* melalui teknik klasifikasi menjadi solusi strategis untuk mengekstrak pengetahuan dari sekumpulan data besar tersebut (Cahyono, 2010). Salah satu algoritma yang memiliki performa unggul dalam tugas klasifikasi adalah *Random Forest*. Algoritma berbasis *ensemble learning* ini mampu menangani data berdimensi tinggi, meningkatkan konsistensi prediksi, serta meminimalisir risiko *overfitting*. Penelitian Rahmat Hidayat dkk. (2025) menunjukan bahwa penerapan Random Forest pada data transaksi ritel mampu menghasilkan Tingkat akurasi yang sangat tinggi, sehingga relevan untuk digunakan dalam analisis penjualan.

Efektivitasnya telah dibuktikan dalam penelitian Posangi dkk. (2023) pada klasifikasi indeks pembangunan manusia. Selain itu, penelitian Hayya' (2025) menunjukkan bahwa *Random Forest* memiliki tingkat akurasi yang lebih tinggi dibandingkan algoritma *K-Nearest Neighbor* (KNN) dalam konteks prediksi harga.

Penelitian ini dibangun melalui sintesis terhadap keterbatasan penelitian terdahulu mengenai analisis penjualan produk Nike. Sebelumnya, Danestiara dkk. (2024) telah menerapkan metode KNN untuk mengidentifikasi produk terlaris dengan akurasi 87%. Namun, metode KNN memiliki keterbatasan dalam hal stabilitas model dan sensitivitas terhadap variasi data yang kompleks. Meskipun penelitian tersebut memberikan fondasi awal, terdapat celah dalam hal optimalisasi akurasi dan ketahanan model terhadap bias data. Oleh karena itu, penelitian ini mentransformasikan pendekatan klasifikasi dari algoritma berbasis jarak (KNN) ke algoritma berbasis *decision tree* (*Random Forest*) untuk meningkatkan presisi klasifikasi pada kategori kinerja penjualan rendah, sedang, dan tinggi.

Selain pemilihan algoritma, penelitian ini juga menerapkan dua teknik validasi sekaligus, yaitu *Hold-Out* dan *K-Fold Cross Validation* pada dataset penjualan Nike. Validasi hold-out memberikan kesempatan bagi setiap data untuk digunakan sebagai data latih dan data uji, sedangkan *K-Fold Cross Validation* membagi data sampel menjadi K bagian kecil yang berbeda. Dalam setiap kali *dilakukan K-fold cross validation*, bagian-bagian data bergantian digunakan sebagai data uji dan data latih (Rian Oktafiani, 2023)

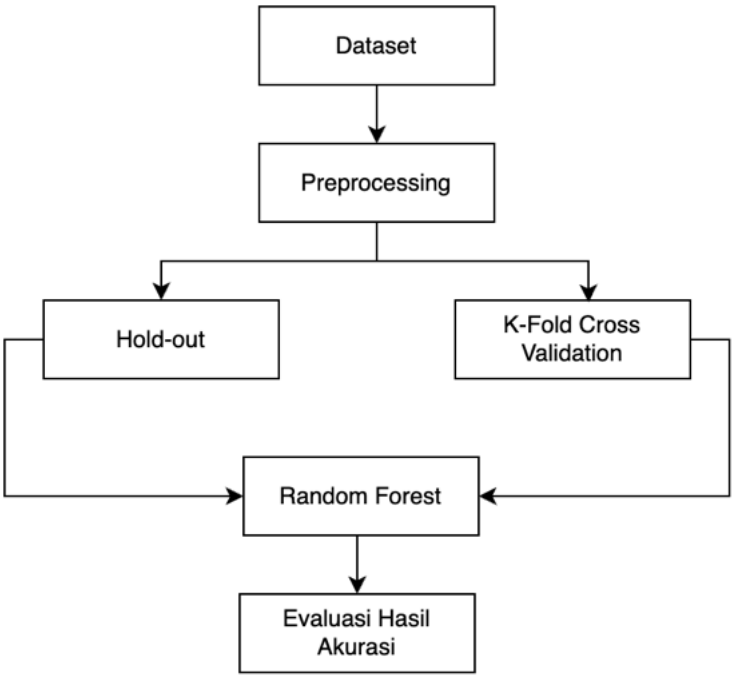
Berbeda dengan penelitian sebelumnya yang cenderung hanya menggunakan satu skema pengujian, penggunaan dua metode validasi ini bertujuan untuk membuktikan

stabilitas dan reliabilitas model secara objektif. Kontribusi utama dari penelitian ini adalah dihasilkannya model klasifikasi yang tidak hanya akurat secara statistik, tetapi juga stabil secara performa, sehingga dapat menjadi instrumen pengambilan keputusan yang valid bagi manajemen ritel dalam memetakan performa produk dan mengoptimalkan strategi distribusi berbasis data.

Tujuan penelitian ini adalah untuk memperoleh model klasifikasi yang akurat dan stabil serta memberikan pemahaman mendalam mengenai faktor-faktor yang memengaruhi performa penjualan produk Nike sebagai dasar pengambilan keputusan strategis di industri ritel.

METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan yang disusun secara sistematis, yaitu pengumpulan data, *preprocessing* data, pemodelan, implementasi, evaluasi model, serta analisis hasil dan penarikan kesimpulan. Penyusunan tahapan tersebut bertujuan untuk memastikan proses analisis berjalan secara terstruktur sehingga model yang dihasilkan memiliki tingkat akurasi dan keandalan yang optimal. Penelitian ini dilakukan dengan tahapan sesuai pada Gambar 1.



Gambar 1. Tahapan Penelitian

1. Dataset

Dataset diperoleh dari *platform Kaggle* dengan total 9.360 baris data transaksi Nike periode 2020-2021. *Dataset* ini mencakup sembilan atribut, yaitu tanggal transaksi, produk, wilayah, *retailer*, metode penjualan, negara bagian, harga per unit, total penjualan, dan jumlah unit terjual. *Dataset* tersebut dipilih karena memiliki berbagai data yang luas, struktur yang lengkap, serta mampu mencerminkan aktivitas penjualan selama dua tahun, sehingga sesuai untuk kebutuhan analisis *data mining*.

2. *Preprocessing* Data

Pengolahan data merupakan tahap penting dalam penelitian ini karena data mentah biasanya belum bisa digunakan langsung untuk proses pemodelan (Putra et al., 2024; Suryanegara, 2021). Tahap ini bertujuan untuk meningkatkan kualitas data melalui beberapa proses seperti pembersihan, penyesuaian jenis data, serta transformasi agar data sesuai dengan kebutuhan. Berikut ini tahapan pengolahan data yang dilakukan:

a. Pemeriksaan nilai yang hilang dilakukan pada setiap atribut untuk memastikan tidak ada data yang kosong atau terlewat. Jika ditemukan nilai yang hilang, penanganannya dapat dilakukan dengan cara menghapus data tersebut, mengisi nilai dengan angka tertentu, atau melakukan pengecekan kembali terhadap sumber data.
b. Penyesuaian tipe data dilakukan agar setiap atribut dapat dibaca dan diproses dengan tepat oleh algoritma *random forest.* Atribut kategorikal seperti *Product, Region, Sales Method,* dan *State* dikonversi ke dalam format *category*, sedangkan atribut *Units Sold* dipastikan berada dalam format numerik (bilangan bulat).
c. Tahap transformasi data mencakup penerapan *Label Encoding* pada variabel kategorikal, mengubah format kategori menjadi format numerik, serta melakukan normalisasi data apabila diperlukan. Setelah itu, *Dataset* diperiksa kembali untuk mendeteksi adanya data yang duplikat.

3. *Splitting* data

Selanjutnya dilakukan proses pra-pengolahan data berupa *splitting data* dan *data formatting* agar data hasil pembagian pada setiap skenario dapat digunakan dalam proses pelatihan model. Skenario pembagian data yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Skenario Pelatihan

Skenario	Split
Hold-out	80 : 20
K-Fold Cross Validation	10

Pembagian data dilakukan dengan dua skema:

a. *Hold-out* (80:20)

Metode *Hold-Out* digunakan dengan membagi *dataset* menjadi dua bagian, yaitu data pelatihan (*training data*) dan data pengujian (*testing data*). Pada penelitian ini, data dibagi dengan perbandingan 80% sebagai data pelatihan dan 20% sebagai data pengujian. Skema ini bertujuan untuk memberikan gambaran performa model *Random Forest* pada satu set data uji yang bersifat *independen*. Dari total 9.360 data, sebanyak 7.488 data digunakan sebagai data pelatihan dan 1.872 data digunakan sebagai data pengujian.

b. K-Fold Cross Validation (K = 10)

Metode K-Fold Cross Validation diterapkan menggunakan data pelatihan sebesar 80% (7.488 data). Data tersebut dibagi ke dalam 10 bagian *(fold)* dengan ukuran yang relatif sama. Proses pelatihan dan pengujian model *Random Forest* dilakukan sebanyak 10 kali, di mana pada setiap iterasi satu *fold* digunakan sebagai data validasi dan sembilan *fold* lainnya digunakan sebagai data pelatihan. Pendekatan ini bertujuan untuk memperoleh hasil evaluasi model yang lebih stabil serta mengurangi ketergantungan terhadap satu pembagian data tertentu.

4. Implementasi Algoritma *Random Forest*

Penelitian ini menggunakan algoritma *Random Forest* karena kemampuannya dalam menangani keterkaitan *antar-fitur* yang *non-linear Parameter* model dikonfigurasi sebagai berikut untuk mencapai performa optimal:

a. *n_estimators:* (Jumlah pohon keputusan yang dibangun untuk meningkatkan akurasi).
b. *max_features: 'sqrt'* (Membatasi jumlah fitur pada setiap *split* untuk mencegah dominasi fitur tertentu).
c. *Criterion:* 'Gini' atau 'Entropy' (Digunakan sebagai fungsi pengukur kualitas pemisahan/ *splitting criteria*).

Dengan menerapkan algoritma *random forest* serta strategi pemecahan data tersebut, model yang dibangun diharapkan mampu memberikan klasifikasi kinerja penjualan yang akurat, stabil, serta dapat digunakan sebagai dasar dalam analisis dan pengambilan keputusan. Menurut (Ega Sri Lestari, 2022) Pembentukan akar pohon akan diambil berdasarkan nilai terkecil *splitting criteria* pada masing-masing *fitur* yang digunakan.

5. Evaluasi Model

Evaluasi model merupakan tahap terakhir dalam proses pengembangan sistem pembelajaran mesin yang bertujuan untuk menilai kemampuan model berdasarkan data uji yang tidak digunakan selama proses pelatihan. Tahap ini memiliki peran penting dalam mengukur sejauh mana model mampu melakukan *generalisasi* terhadap data baru serta memastikan bahwa hasil prediksi yang dihasilkan akurat dan konsisten, Jika suatu model menunjukkan performa yang baik pada data pelatihan namun buruk pada data pengujian, maka model tersebut dapat dikategorikan mengalami *overfitting.* Oleh karena itu, proses evaluasi menjadi langkah penting dalam menilai kualitas model secara menyeluruh.

Dalam penelitian ini, evaluasi model dilakukan dengan menggunakan beberapa metrik yang umum digunakan pada algoritma klasifikasi, khususnya *random forest.* Metrik evaluasi yang digunakan mencakup *confusion matrix* dan *classification report*, yang terdiri dari nilai *precision*, *recall*, serta *F1-score.* Penggunaan metrik tersebut bertujuan untuk memberikan gambaran yang lebih lengkap mengenai tingkat akurasi prediksi model.

Selain itu, evaluasi menggunakan *10-Fold Cross Validation* dilakukan dengan membagi data latih menjadi sepuluh bagian, di mana setiap bagian secara bergantian digunakan sebagai data uji. Pendekatan ini bertujuan untuk meminimalkan bias akibat pembagian data secara acak serta memastikan bahwa performa model tidak hanya bergantung pada satu skema pembagian data tertentu. Dengan demikian, hasil evaluasi yang diperoleh mencerminkan kemampuan generalisasi model *Random Forest* secara lebih objektif dan stabil.

HASIL DAN PEMBAHASAN

1. *Dataset*

Dataset dalam penelitian ini diperoleh dari platform Kaggle.com sebagai sumber data sekunder yang telah melalui tahap awal pengumpulan. Meskipun dataset tersedia secara terbuka, tetap diperlukan proses seleksi dan pemilihan data dari himpunan data operasional sebelum tahapan *knowledge discovery in database* dapat dilakukan (Ginantra et al., 2021). Setelah proses seleksi, dataset transaksi kemudian diproses lebih lanjut melalui tahap *preprocessing* untuk memastikan kualitas dan kesiapan data sebelum

pemodelan. Adapun atribut-atribut yang digunakan dalam penelitian ini disajikan pada Tabel 2.

Tabel 2. Atribut *Dataset*

Data atribut	Deskripsi
Invoice_date	Tanggal terjadinya transaksi penjualan
Month	Bulan terjadinya transaksi penjualan (hasil ekstraksi dari *invoice_Date)*
Product	Jenis atau nama produk nike yang dijual
Region	Wilayah geografis tempat penjualan dilakukan
State	Negara bagian atau lokasi administratif tempat transaksi berlangsung.
Sales_Method	Metode penjualan yang digunakan,seperti *In-Store* atau *Online*
Price_per_units	Harga satuan produk yang dijual.
Units_Sold	Jumlah unit produk yang berhasil terjual dalam satu transaksi
Total_Sales	Total nilai penjualan yang diperoleh dari transaksi.
Category_Units_Sold	Kategori tingkat penjualan berdasarkan jumlah *Units sold*: (rendah, sedang, tinggi).

Dataset ini dipilih karena memiliki atribut yang lengkap, jumlah data yang cukup besar, serta mencakup aktivitas penjualan selama dua tahun, sehingga sesuai untuk digunakan dalam analisis prediksi serta klasifikasi.

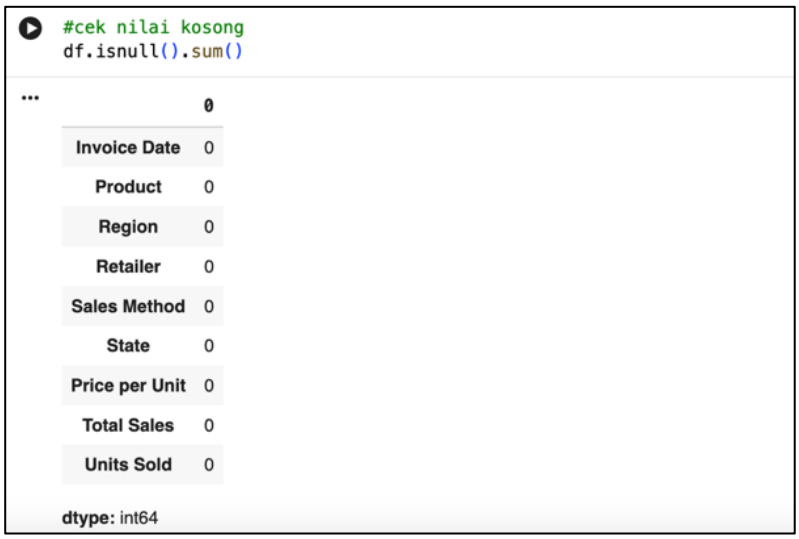
2. *Preprocessing* Data

Pada tahap awal pengolahan data, *preprocessing* memiliki peran yang sangat penting untuk memastikan hasil analisis menjadi akurat dan dapat diandalkan. Sehingga hal ini bertujuan untuk mempersiapkan *dataset* agar siap digunakan sebelum diterapkan pada model dan analisis lanjutan. *preprocessing* yang dilakukan mencakup pemeriksaan data yang berulang, pengecekan nilai yang kosong, serta penyesuaian format data agar konsisten berikut penjelasnya :

a. Memastikan Tidak Ada Nilai yang Hilang

Setelah data dipisahkan, setiap kolom diperiksa untuk memastikan tidak terdapat nilai kosong (*missing value*) yang dapat memengaruhi proses analisis. Apabila ditemukan nilai yang hilang, maka dilakukan penanganan seperti imputasi nilai yang sesuai, penghapusan baris yang tidak lengkap, atau verifikasi kembali terhadap sumber data. Berdasarkan hasil pemeriksaan, seluruh atribut pada dataset tidak memiliki nilai yang hilang sehingga data dapat langsung digunakan pada tahap analisis selanjutnya, sebagaimana ditunjukkan pada Gambar 2.

Sainteks
ISSN: 2686-0546
Volume 23 No 1, April 2026
DOI: 10.30595/sainteks.v23i1.29930 (79-92)



Gambar 2. Hasil cek nilai kosong

b. Menyesuaikan Tipe Data pada Setiap Atribut

Menyesuaikan jenis data pada setiap atribut Kolom kategorikal seperti *Product*, *Region*, *Sales Method*, dan *State* dikonversi ke format *category*. Kolom *Units Sold* dipastikan berupa format numerik (bilangan bulat). Penyesuaian jenis data ini disajikan pada Tabel 3.

Tabel 3. Menyesuaikan Tipe Data

Atribut	Format
Product	*category*
Region	*category*
Sales Method	*Category*
State	*Category*
Units Sold	*int64*

c. *Label Encoding*

Pada tahap pra-pemrosesan data, dilakukan proses *encoding* terhadap beberapa fitur kategorikal dalam *dataset*. Kolom *Product, Region, Sales Method*, dan *State* yang semula berupa data kategori diubah menjadi format numerik. Setiap kategori dalam kolom-kolom tersebut diberi angka agar memudahkan pemrosesan lebih lanjut. Kolom *Product* menunjukkan jenis produk yang telah diberi label berupa angka, seperti 0, 1, 2, dan seterusnya. Selanjutnya Kolom *region* menunjukkan wilayah penjualan yang juga diubah menjadi angka, tabel kolom *sales method* mewakili metode penjualan, diberi nilai numerik. Serta kolom *state* menunjukkan status wilayah atau area tertentu, dengan angka sebagai pengganti nama wilayah tersebut.

Sementara itu, kolom *Units Sold* tetap dalam bentuk numerik karena sudah berupa data kuantitatif dan menggambarkan jumlah unit yang terjual.Variabel *Units Sold* yang semula dalam bentuk numerik diubah menjadi variabel kategorikal untuk mengubah permasalahan regresi menjadi permasalahan klasifikasi. Kategorisasi dilakukan dengan membagi *Units Sold* ke dalam tiga kategori, yaitu: Rendah, Sedang, dan Tinggi. Tujuan dari proses kategorisasi ini adalah untuk mempermudah

Sainteks
ISSN: 2686-0546
Volume 23 No 1, April 2026
DOI: 10.30595/sainteks.v23i1.29930 (79-92)

interpretasi hasil klasifikasi serta meningkatkan konsistensi kinerja model *random forest* disajikan dalam Gambar 3.

```
hasil_encoding = df[['Product', 'Region', 'Sales Method', 'State', 'Units Sold']].head(5)
hasil_encoding
```

	Product	Region	Sales Method	State	Units Sold
0	2	1	0	29	120
1	1	1	0	29	100
2	5	1	0	29	100
3	4	1	0	29	85
4	0	1	0	29	90

Gambar 3. Hasil *encoding*

d. Transformasi Data

Sebelum dilakukan transformasi data, diperlukan peninjauan terhadap *dataset* dalam kondisi aslinya untuk memahami struktur dan karakteristik data yang masih terdiri dari atribut numerik dan kategorikal. Tahap ini penting sebagai dasar penentuan metode transformasi yang tepat agar data dapat diproses secara optimal oleh algoritma *random forest.* Dataset awal yang digunakan dalam penelitian ini masih terdiri atas atribut numerik dan kategorikal. Struktur data sebelum dilakukan proses transformasi dapat dilihat pada Tabel 4.

Tabel 4. Data Sebelum Transformasi

Invoice Date	*Product*	*Region*	*Retailer*	*Sales Method*	*State*	*Price per Unit*	*Total Sales*	*Units Sold*
01-01-2020	*Men's Street Footwear*	Northeast	Foot Locker	*In-store*	New York	50	6000	120
02-01-2020	*Men's 88 Footwear*	Northeast	Foot Locker	*In-store*	New York	50	5000	100
03-01-2020	*Women's Street Footwear*	Northeast	Foot Locker	*In-store*	New York	40	4000	100
04-01-2020	*Women's Athletic Footwear*	Northeast	Foot Locker	*In-store*	New York	45	3825	85
05-01-2020	*Men's Apparel*	Northeast	Foot Locker	*In-store*	New York	60	5400	90
06-01-2020	*Women's Apparel*	Northeast	Foot Locker	*In-store*	New York	50	5000	100

Setelah proses encoding dilakukan pada variabel kategorikal, struktur dataset mengalami perubahan menjadi format numerik. Data hasil transformasi tersebut dapat dilihat pada Tabel 5.

Tabel 5. Data Setelah Transformasi

Invoice Date	*Product*	*Region*	*Retailer*	*Sales Method*	*State*	*Price per Unit*	*Total Sales*	*Units Sold*
01/01/20	2	1	1	0	29	50	6000	120
02/01/20	1	1	1	0	29	50	5000	100
03/01/20	5	1	1	0	29	40	4000	100
04/01/20	4	1	1	0	29	45	3825	85
05/01/20	0	1	1	0	29	60	5400	90
06/01/20	3	1	1	0	29	50	5000	100

Selanjutnya, *dataset* diperiksa untuk mendeteksi adanya baris yang muncul lebih dari satu kali. Baris yang berulang dapat memberikan bias dalam analisis, sehingga perlu dihilangkan. Pemeriksaan dilakukan terhadap seluruh kolom, Berdasarkan hasil pemeriksaan, tidak ditemukan nilai kosong pada semua atribut yang digunakan dalam penelitian ini.

3. *Spliting* data

a. *Hold-out*

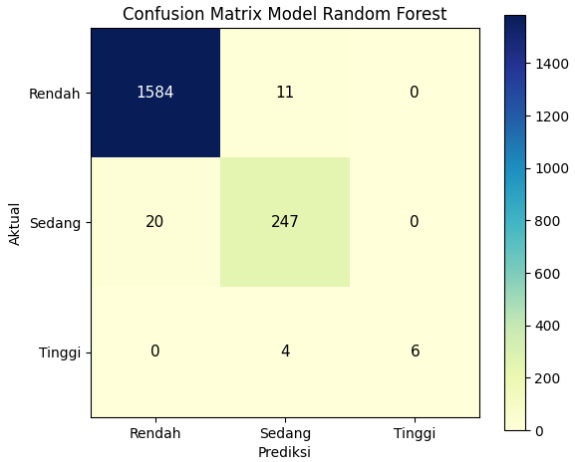
Pada tahap ini, evaluasi kinerja model dilakukan menggunakan metode *Hold-Out* dengan pembagian data sebesar 80% sebagai data latih dan 20% sebagai data uji. Pendekatan ini digunakan untuk mengukur kemampuan model *Random Forest* dalam melakukan klasifikasi terhadap data yang belum pernah dilihat sebelumnya. Dengan menggunakan data uji yang bersifat independen, metode *Hold-Out* dapat memberikan gambaran awal mengenai tingkat akurasi dan keandalan model dalam melakukan prediksi kinerja penjualan produk Nike.

Evaluasi model pada skema *Hold-Out* dilakukan dengan menggunakan beberapa metrik evaluasi klasifikasi yang bertujuan untuk menilai performa model secara menyeluruh. Salah satu metrik utama yang digunakan adalah *confusion matrix*, yang mampu menggambarkan jumlah prediksi benar dan salah pada setiap kelas kinerja penjualan, yaitu Rendah, Sedang, dan Tinggi.

- *Confusion matrix*

  Confusion matrix digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data uji ke dalam kelas *Rendah*, *Sedang*, dan *Tinggi*. Hasil *confusion matrix* ditunjukkan pada Gambar 4.

Sainteks
ISSN: 2686-0546
Volume 23 No 1, April 2026
DOI: 10.30595/sainteks.v23i1.29930 (79-92)



Gambar 4. ***Confusion matrix Random Forest***

Berdasarkan hasil tersebut, dari total 1.872 data uji, jumlah prediksi yang benar (diagonal utama *confusion matrix*) mencapai 1.837 data. Nilai akurasi model diperoleh sebesar 98,13%**,** yang menunjukkan bahwa algoritma *random forest* memiliki tingkat ketepatan klasifikasi yang sangat tinggi dan mampu meminimalkan kesalahan prediksi pada sebagian besar kelas. Adapun nilai akurasi digunakan untuk mengukur tingkat ketepatan model secara keseluruhan dalam mengklasifikasikan data uji. Nilai akurasi menunjukkan perbandingan antara jumlah prediksi yang benar terhadap keseluruhan data yang diuji. Semakin tinggi nilai akurasi yang diperoleh, maka semakin baik kemampuan model dalam melakukan klasifikasi, hasil confusion matrix yang diubah ke bentuk tabel disajikan dalam Tabel 6.

Tabel 6. ***Confusion matrix*** **Model** ***Random Forest***

Aktual \ Prediksi	Rendah	Sedang	Tinggi
Rendah	1584	11	0
Sedang	20	247	0
Tinggi	0	4	6

$$\text{Akurasi} = \frac{\text{Jumlah prediksi benar}}{Total\ Data}$$

$$\text{Akurasi} = \frac{1584 + 247 + 6}{1872} = \frac{1837}{1872} = 0{,}9813 = \ 98{,}13\%$$

- *Classification Report*

  Evaluasi kinerja model tidak hanya dilakukan menggunakan *confusion matrix*, tetapi juga dianalisis melalui *classification report*. *Classification report* menyajikan metrik *precision*, *recall*, *f1-score*, dan *support* untuk masing-masing kelas. Hasil evaluasi *classification report* pada penelitian ini ditunjukkan pada Gambar 5.

```
Classification Report:
              precision    recall  f1-score   support

      Rendah       0.99      0.99      0.99      1595
      Sedang       0.94      0.93      0.93       267
      Tinggi       1.00      0.60      0.75        10

    accuracy                           0.98      1872
   macro avg       0.98      0.84      0.89      1872
weighted avg       0.98      0.98      0.98      1872
```

Gambar 5. Hasil klasifikasi

Selain *confusion matrix*, evaluasi model juga dilakukan menggunakan *classification report* yang menyajikan metrik *precision*, *recall*, dan *f1-score* untuk setiap kelas, sebagaimana ditampilkan pada Gambar 5 Hasil evaluasi menunjukkan bahwa kelas Rendah memiliki nilai *precision*, *recall*, dan *f1-score* masing-masing sebesar 0,99, yang menandakan performa klasifikasi yang sangat optimal. Pada kelas Sedang, nilai *precision* sebesar 0,94, *recall* 0,93, dan *f1-score* 0,93 menunjukkan bahwa model mampu mengklasifikasikan kelas ini dengan baik dan seimbang.

Meskipun model *Random Forest* menghasilkan akurasi global yang sangat tinggi sebesr 98,13% analisis mendalam melalui *classification report* menunjukkan adanya perbedaan performa pada tiap kategori. Berdasarkan hasil pengujian, kategori "Tinggi" memiliki nilai *recall* yang lebih rendah yaitu 0,60 dibandingkan kategori "Rendah" dan "Sedang" yang mencapai 0.99 dan 0,94. Hal ini disebabkan oleh ketidakseimbangan jumlah data (*support*). Dalam *dataset* ini, dengan kategori "Tinggi" berjumlah jauh lebih sedikit daripada kategori lainya.

Secara teknis, algoritma *Random Forest* bekerja dengan memaksimalkan akurasi keseluruhan. Akibatnya, model cenderung lebih "mengenal" pola pada kelas mayoritas (Rendah dan Sedang) dan kurang sensitif terhadap kelas minoritas (Rendah). Rendahnya *recall* pada kelas "Rendah" menunjukkan adanya sejumlah data penjualan tinggi yang salah diprediksi sebagai "Sedang" atau "Rendah" (*False Negative*).

Hal ini menyebabkan risiko kehilangan peluang *(Oppertunity Loss)*. Jika produk yang seharusnya masuk kategori "Rendah" (sangat laku) diprediksi sebagai "Sedang", manajemen mungkin tidak menyediakan stok yang cukup atau mengurangi intensitas pemasaran. Hal ini berpotensi menyebabkan kehabisan stok pada saat permintaan sedang memuncak.

b. *K-Fold Cross Validation*

Hasil dari penelitian didapatan dari pelatihan berdasarkan skenario uji coba yang telah direncanakan melalui serangkaian uji coba, Hasil pengujian akurasi pada setiap fold dapat dilihat pada Tabel 7.

Berdasarkan hasil pengujian pada Tabel 7, dapat diamati bahwa penerapan teknik *10-Fold Cross Validation* memberikan gambaran yang cukup stabil terhadap kinerja algoritma *Random Forest* dalam melakukan klasifikasi data. Nilai akurasi pada setiap *fold* menunjukkan variasi, dengan rentang akurasi berada di sekitar 90%–94%, yang mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik terhadap data yang digunakan.

Pada beberapa *fold* awal khususnya pada dan ke-4, algoritma *Random Forest* mencapai nilai akurasi tertinggi, yaitu 94,39%. Hal ini menunjukkan bahwa kombinasi data latih dan data uji pada *fold* tersebut mampu merepresentasikan pola data secara

optimal, sehingga model dapat melakukan klasifikasi dengan tingkat ketepatan yang tinggi. Kondisi ini sejalan dengan karakteristik *Random Forest* sebagai metode *ensemble* yang menggabungkan banyak *decision tree* untuk menghasilkan prediksi yang lebih stabil dan akurat.

Tabel 7. Akurasi *K-Fold Cross Validation*

K-Fold	Total Akurasi
1	92,92%
2	93,59%
3	92,52%
4	94,39%
5	91,85%
6	92,79%
7	91,32%
8	92,25%
9	92,78%
10	90,90%

Namun demikian, hasil pengujian juga menunjukkan bahwa setelah *fold* tertentu, khususnya mulai dari *fold* ke-7 hingga *fold* ke-10, terjadi penurunan nilai akurasi meskipun tidak bersifat drastis. Nilai akurasi terendah tercatat pada *fold* ke-10 sebesar 90,91%. Penurunan ini dapat disebabkan oleh perbedaan distribusi data pada masing-masing *fold*, di mana sebagian data uji memiliki karakteristik yang lebih kompleks atau mengandung *noise*, sehingga memengaruhi performa model.

Fenomena ini menunjukkan bahwa meskipun *Random Forest* dikenal memiliki ketahanan yang baik terhadap *overfitting*, variasi komposisi data latih dan data uji pada setiap fold tetap dapat memengaruhi hasil klasifikasi. Oleh karena itu, penggunaan *10-Fold Cross Validation* dalam penelitian ini terbukti efektif untuk memberikan evaluasi kinerja model yang lebih menyeluruh dan tidak bergantung pada satu pembagian data saja. Secara keseluruhan, hasil ini menegaskan bahwa algoritma *Random Forest* mampu menghasilkan performa klasifikasi yang konsisten dan andal pada *dataset* yang digunakan

Walaupun terdapat kendala ketidakseimbangan kelas, model ini tetap dianggap sangat layak. Nilai *weighted average, f1-score* sebesar 0.98 menunjukkan bahwa secara keseluruhan , model masih mampu menjaga keseimbangan antara *precison* dan *recall*. Penggunaan *10-Fold Cross Validation* dengan rata-rata akurasi 92,53% juga membuktikan bahwa model tidak mengalami *overfitting*. Yang parah dan mampu melakukan generalisasi dengan cukup baik pada data yang bervariasi.

Jika dibandingkan dengan penelitian terdahulu oleh Danestiara dkk (2024) yang menggunakan algoritma *K-Nearest Neighbor (KNN)* pada *dataset* yang sama dengan akurasi 87%, hasil penelitian ini membuktikan bahwa *Random Forest* lebih unggul dalam menangani variabilitas data ritel. Keunggulan ini sejalan dengan temuan Hayya' (2025) yang menyatakan bahwa *Random forest* lebih stabil terhadap *outliers* dibandingkan KNN, karena keputusan akhir diambil berdasarkan pemungutan suara terbanyak *(majority voting)* dari sekumpulan pohon, bukan berdasarkan jarak antar titik dan tunggal.

KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan, dapat disimpulkan bahwa penerapan algoritma Random Forest efektif dalam mengklasifikasikan kinerja penjualan produk Nike

ke dalam kategori rendah, sedang, dan tinggi dengan tingkat akurasi yang sangat tinggi, yaitu 98,13% pada metode *Hold-Out* serta pada metode *10-Fold Cross Validation* dengan capaian akurasi tertinggi 94,39%, sehingga menunjukkan bahwa model memiliki performa yang akurat, stabil, dan mampu melakukan generalisasi dengan baik. Meskipun demikian, penelitian ini masih memiliki keterbatasan pada ketidakseimbangan jumlah data antar kelas dan terbatasnya variabel yang digunakan, sehingga penelitian selanjutnya disarankan untuk menerapkan teknik penyeimbangan data seperti SMOTE, melakukan optimasi *hyperparameter* menggunakan *Grid Search*, serta menguji algoritma lain seperti *XGBoost* atau *LightGBM* agar diperoleh model yang lebih adaptif dan sensitif terhadap kelas minoritas pada dataset ritel yang lebih kompleks.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada penyedia dataset melalui platform Kaggle yang telah menyediakan data transaksi penjualan produk Nike yang digunakan dalam penelitian ini. Penulis juga menyampaikan terima kasih kepada prodi Sistem Informasi Fakultas Sains dan Teknologi Universitas Muhammadiyah Kudus atas dukungan akademik yang diberikan selama pelaksanaan penelitian ini.

DAFTAR PUSTAKA

Cahyono, E.T. (2010) 'Data mining: Solusi pengembangan pengetahuan berdasarkan basis data', Teknomatika, 2(1), pp. 24–34. Available at: https://ejournal.unjaya.ac.id/index.php/teknomatika/article/view/355.

Danestiara, V.R. and Maulana, M.A. (2024) 'Penerapan metode K-Nearest Neighbor untuk identifikasi produk Nike paling laris terjual', In Search, 23(2), pp. 88–97. https://doi.org/10.37278/insearch.v23i2.1129.

Ginantra, N.L.W.S.R. et al. (2021) Data mining dan penerapan algoritma. Medan: Yayasan Kita Menulis.

Hafizh, M. and Nurdin, T. (2023) 'Implementasi data mining menggunakan algoritma FP-Growth untuk menganalisis transaksi penjualan ekspor online', Jurnal Teknologi dan Sistem Informasi Bisnis, 5(3), pp. 242–249. https://doi.org/10.47233/jteksis.v5i3.847.

Hayya', M. (2025) 'Perbandingan metode K-Nearest Neighbor (KNN) dan Random Forest dalam memprediksi harga rumah', Sainteks, 22(1), pp. 99–108. https://doi.org/10.30595/sainteks.v22i1.25998.

Hidayat, R.A. (2025) 'Implementasi algoritma Random Forest Regression untuk memprediksi penjualan produksi di supermarket', Jurnal Sistem Informasi dan Sistem Komputer, 10(1), pp. 101–109. https://doi.org/10.51717/simkom.v10i1.703.

Indrawan, Y.W. (2025) Manajemen pemasaran ritel. Jakarta: Yayasan Putra Adi Dharma. Available at: https://journal.yayasanpad.org/index.php/ypadbook/article/view/328/221.

Lestari, E.S. (2022) 'Penerapan Random Forest Regression untuk memprediksi harga jual rumah dan cosine similarity untuk rekomendasi rumah pada Provinsi Jawa Barat', Jurnal Ilmiah FIFO, 14(2), pp. 131–146. https://doi.org/10.22441/fifo.2022.v14i2.003.

Oktafiani, R. and H.A. (2023) 'Pengaruh komposisi split data terhadap performa klasifikasi penyakit kanker payudara menggunakan algoritma machine learning', Jurnal Sains dan Informatika, 9(1), pp. 19–28. https://doi.org/10.34128/jsi.v9i1.622.

Posangi, T., Yasin, L. and Lumintang, V. (2023) 'Implementasi algoritma Random Forest dengan forward selection untuk klasifikasi Indeks Pembangunan Manusia', Jambura Journal of Probability and Statistics, 4(2), pp. 85–91. https://doi.org/10.37905/jjps.v4i2.18460.

Putra, Y.S. and Kurniawan, R. (2024) 'Penerapan data mining menggunakan algoritma FP-Growth pada data penjualan sembako', JATI (Jurnal Mahasiswa Teknik Informatika), 8(1), pp. 561–567. https://doi.org/10.36040/jati.v8i1.8391.

Suryanegara, G.A.B. and Mahendra, A. (2021) 'Peningkatan hasil klasifikasi pada algoritma Random Forest untuk deteksi pasien penderita diabetes menggunakan metode normalisasi', Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(1), pp. 114–122. Available at: https://jurnal.iaii.or.id/.