TUGAS MANDIRI
FUNDAMENTALS OF DATA MINING

IMPLEMENTASI ALGORITMA C4.5 UNTUK IDENTIFIKASI PRODUK PALSU BERDASARKAN DATA PENJUALAN ONLINE

Nama : Amma Natasya
NPM : 231510043
Dosen : Erlin Elisa, S.Kom., M.Kom.

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN KOMPUTER
UNIVERSITAS PUTERA BATAM

2026

1.1 Deskripsi Dataset

- Sumber dataset : Kaggle.com
  https://www.kaggle.com/datasets/2ceac1b213331b93f3c999da491c2f9e922597502c180e57f90bfed60ed9ad46
- Jumlah record : 5.000 data
- Jumlah atribut : 27 atribut (23 fitur, 1 target, atribut ID, dan tanggal)
- Tipe data : Numerik (int dan float), kategorikal, boolean
- Target/label : is_counterfeit (true → produk palsu, false → produk asli)
- Target Permasalahan: Melakukan klasifikasi untuk mendeteksi apakah suatu produk termasuk palsu atau asli berdasarkan karakteristik produk dan penjual.

1.2 Persiapan Data dan Preprocessing

- Data Cleaning

  Data cleaning merupakan tahap awal preprocessing yang bertujuan untuk memastikan kualitas data sebelum dilakukan pemodelan. Pada dataset penjualan produk palsu yang digunakan, tidak ditemukan nilai kosong (missing value) pada seluruh atribut, sehingga tidak diperlukan proses penghapusan atau imputasi data. Selain itu, pengecekan terhadap outlier dilakukan secara deskriptif dan tidak ditemukan nilai ekstrem yang mengganggu proses analisis. Oleh karena itu, seluruh data dapat digunakan tanpa perlu dilakukan penghapusan data tambahan.

- Encoding Data Kategorikal (LabelEncoder / OneHotEncoder)

  Dataset memiliki beberapa atribut bertipe kategorikal, seperti kategori produk, merek, negara penjual, dan asal pengiriman. Karena algoritma data mining tidak dapat memproses data dalam bentuk teks, maka diperlukan proses encoding untuk mengubah data kategorikal menjadi data numerik. Pada penelitian ini digunakan metode Label Encoding, di mana setiap kategori diubah menjadi nilai numerik yang unik. Metode ini dipilih karena sederhana, efisien, dan sesuai untuk dataset dengan jumlah kategori yang tidak terlalu besar.

- Scaling / Normalization (StandardScaler)

Proses normalisasi dilakukan untuk menyamakan skala antar fitur, karena dataset memiliki rentang nilai yang berbeda-beda, seperti harga produk, rating penjual, dan jumlah ulasan. Pada penelitian ini digunakan metode **StandardScaler**, yang menstandarisasi data dengan mengubah nilai fitur sehingga memiliki rata-rata nol dan standar deviasi satu. Normalisasi ini bertujuan untuk mencegah fitur dengan skala besar mendominasi proses pembelajaran model.

- Feature Selection / Feature Engineering

  Feature selection dilakukan untuk memilih atribut yang relevan dan menghilangkan atribut yang tidak berkontribusi terhadap proses klasifikasi. Pada dataset ini, atribut seperti *product_id*, *seller_id*, dan *listing_date* dihapus karena hanya berfungsi sebagai identitas dan tidak memiliki pengaruh langsung terhadap penentuan keaslian produk. Dengan melakukan feature selection, model menjadi lebih efisien dan mampu belajar dari fitur-fitur yang benar-benar berpengaruh.

- Split Data Train & Test

  Tahap pembagian data dilakukan untuk mengevaluasi kinerja model secara objektif. Dataset dibagi menjadi dua bagian, yaitu data latih (training data) sebesar 80% dan data uji (testing data) sebesar 20%. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengukur performa model pada data yang belum pernah dilihat sebelumnya. Pembagian ini bertujuan untuk menghindari overfitting dan memastikan model memiliki kemampuan generalisasi yang baik.

Tabel 1. 1 Ringkasan Prepocessing Data

Aspek	**Sebelum Preprocessing**	**Sesudah Preprocessing**
Jumlah Data	5.000 data	5.000 data
Jumlah Fitur	27 atribut	23 fitur
Missing Value	Tidak ada	Tidak ada
Data Kategorikal	Ada (category, brand, seller_country, shipping_origin)	Sudah di-encode menjadi numerik
Data Boolean	Ada (True / False)	Dikonversi menjadi 0 dan 1
Skala Data	Beragam (harga, rating, jumlah)	Sudah dinormalisasi
Kesiapan Data	Belum siap untuk modeling	Siap digunakan untuk pemodelan

Tabel 1. 2 Distribusi Data Training dan Testing

Jenis Data	Jumlah Data	Persentase
Data Latih (Training)	4.000	80%
Data Uji (Testing)	1.000	20%
Total	**5.000**	**100%**

1.3 Analisis Statistik dan Visualisasi

- Statistik Deskriptif Dataset

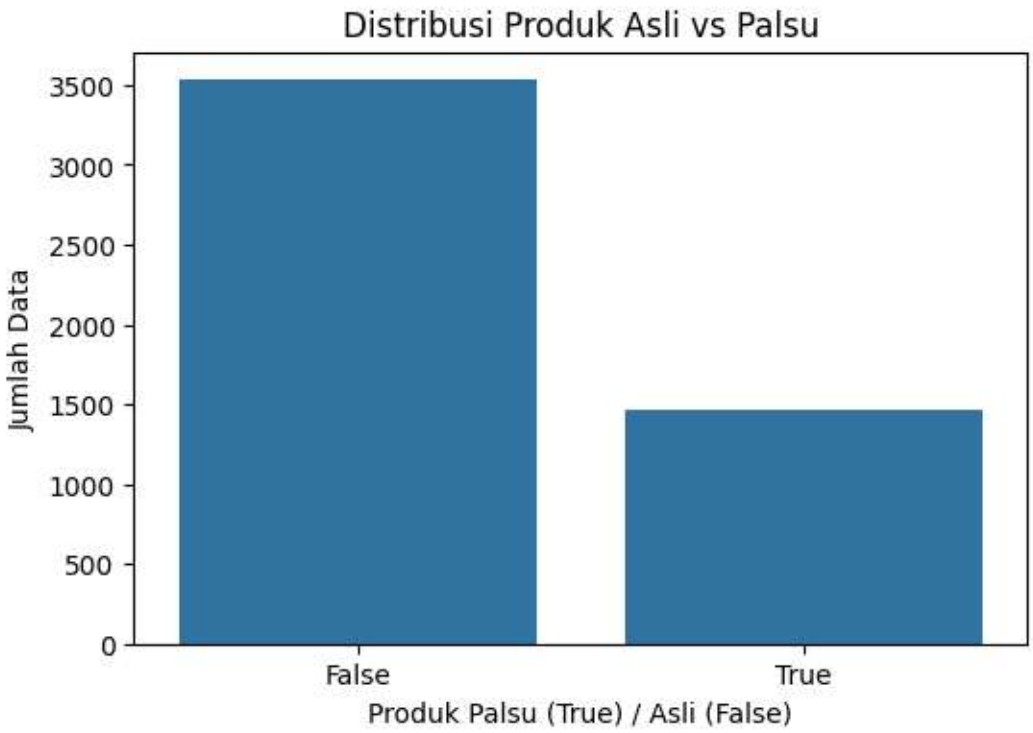
  Analisis statistik deskriptif dilakukan untuk memahami karakteristik dasar dataset sebelum dilakukan pemodelan. Dataset penjualan produk palsu yang digunakan dalam penelitian ini terdiri dari **5.000 data** dengan **23 fitur numerik** setelah tahap preprocessing. Fitur-fitur numerik tersebut mencakup karakteristik produk, reputasi penjual, pola transaksi, serta informasi pengiriman.

  Nilai statistik seperti rata-rata, nilai minimum, dan nilai maksimum menunjukkan adanya variasi yang cukup signifikan antar fitur, terutama pada atribut harga produk, jumlah ulasan penjual, jumlah tampilan produk, dan waktu pengiriman. Variasi ini menunjukkan bahwa dataset memiliki informasi yang kaya dan berpotensi kuat untuk membedakan antara produk palsu dan produk asli.

- Distribusi Target / Label

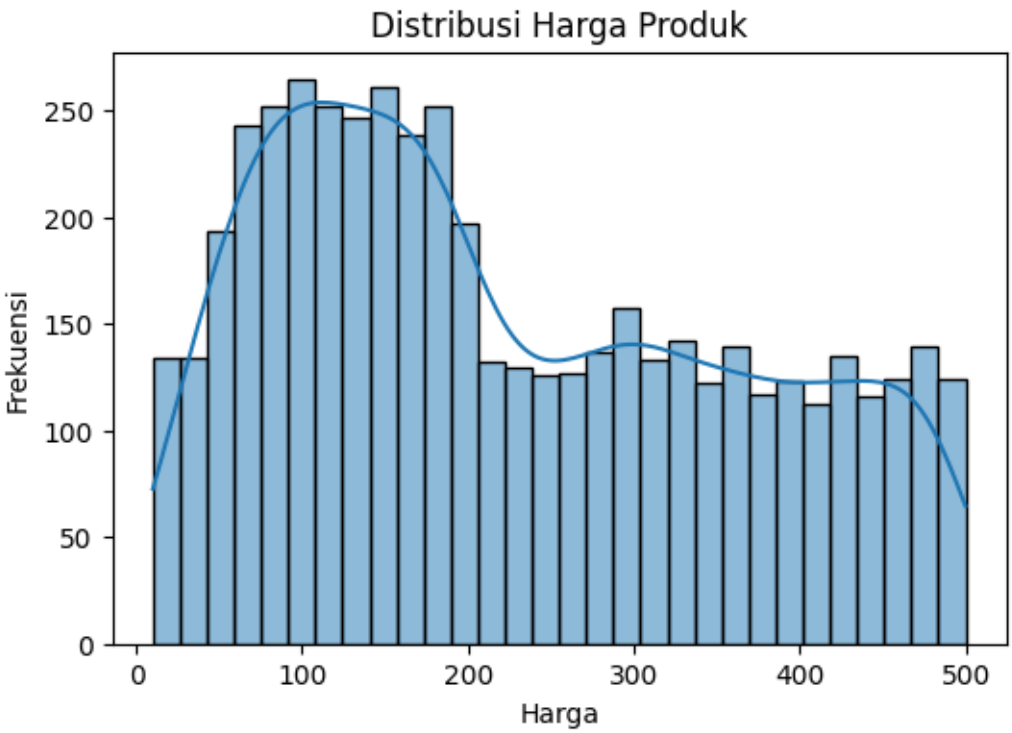
  Distribusi target `is_counterfeit` menunjukkan bahwa dataset terdiri dari dua kelas, yaitu produk palsu dan produk asli. Berdasarkan hasil visualisasi distribusi label, jumlah produk asli lebih banyak dibandingkan produk palsu, namun perbedaannya tidak terlalu signifikan.

  Grafik distribusi produk asli dan produk palsu menunjukkan bahwa jumlah produk asli lebih banyak dibandingkan produk palsu, namun selisihnya tidak terlalu besar. Hal ini menandakan bahwa dataset memiliki distribusi kelas yang relatif seimbang. Kondisi tersebut menguntungkan dalam proses pelatihan model klasifikasi karena mengurangi risiko bias terhadap salah satu kelas.



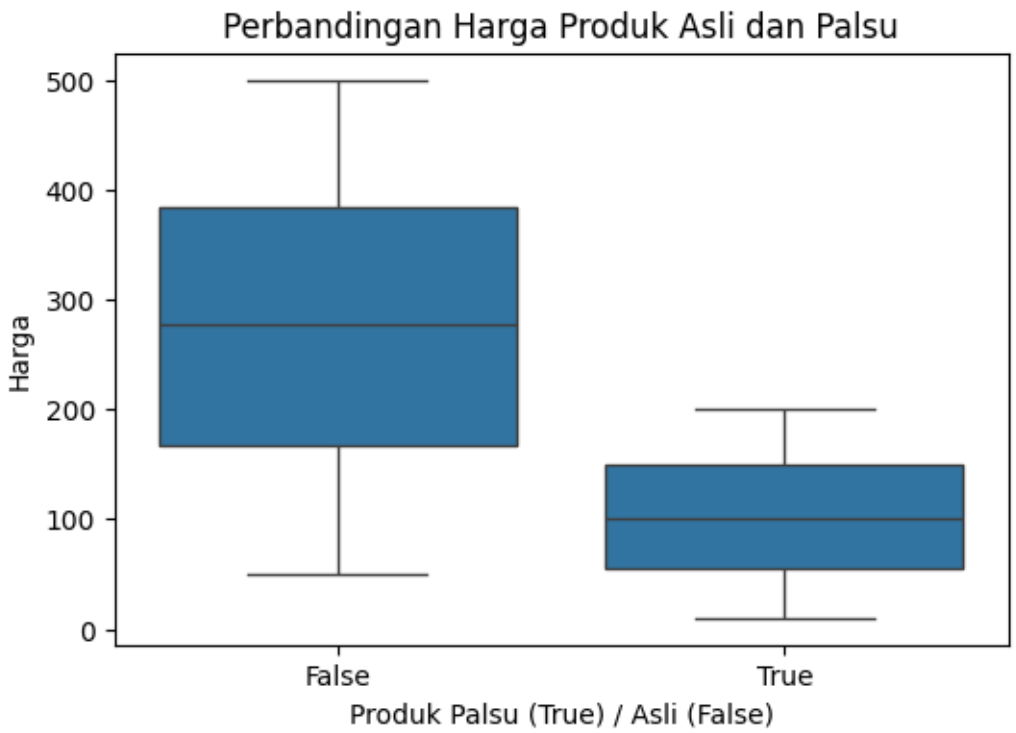
Gambar 1. 1 Grafik Distribusi Produk Asli vs Produk Palsu

Grafik distribusi harga produk menunjukkan bahwa sebagian besar produk berada pada rentang harga tertentu dengan sebaran yang cukup bervariasi. Pola distribusi ini mengindikasikan bahwa harga produk tidak terdistribusi secara merata dan memiliki perbedaan nilai yang signifikan antar produk. Oleh karena itu, fitur harga berpotensi menjadi salah satu indikator penting dalam membedakan produk asli dan produk palsu.



Gambar 1. 2 Grafik Distribusi Harga Produk

Grafik perbandingan harga menunjukkan bahwa produk palsu cenderung memiliki harga yang lebih rendah dibandingkan produk asli. Selain itu, variasi harga pada produk palsu terlihat lebih besar, yang mengindikasikan adanya pola harga yang tidak konsisten. Hal ini menunjukkan bahwa fitur harga memiliki kontribusi yang kuat dalam proses identifikasi produk palsu.

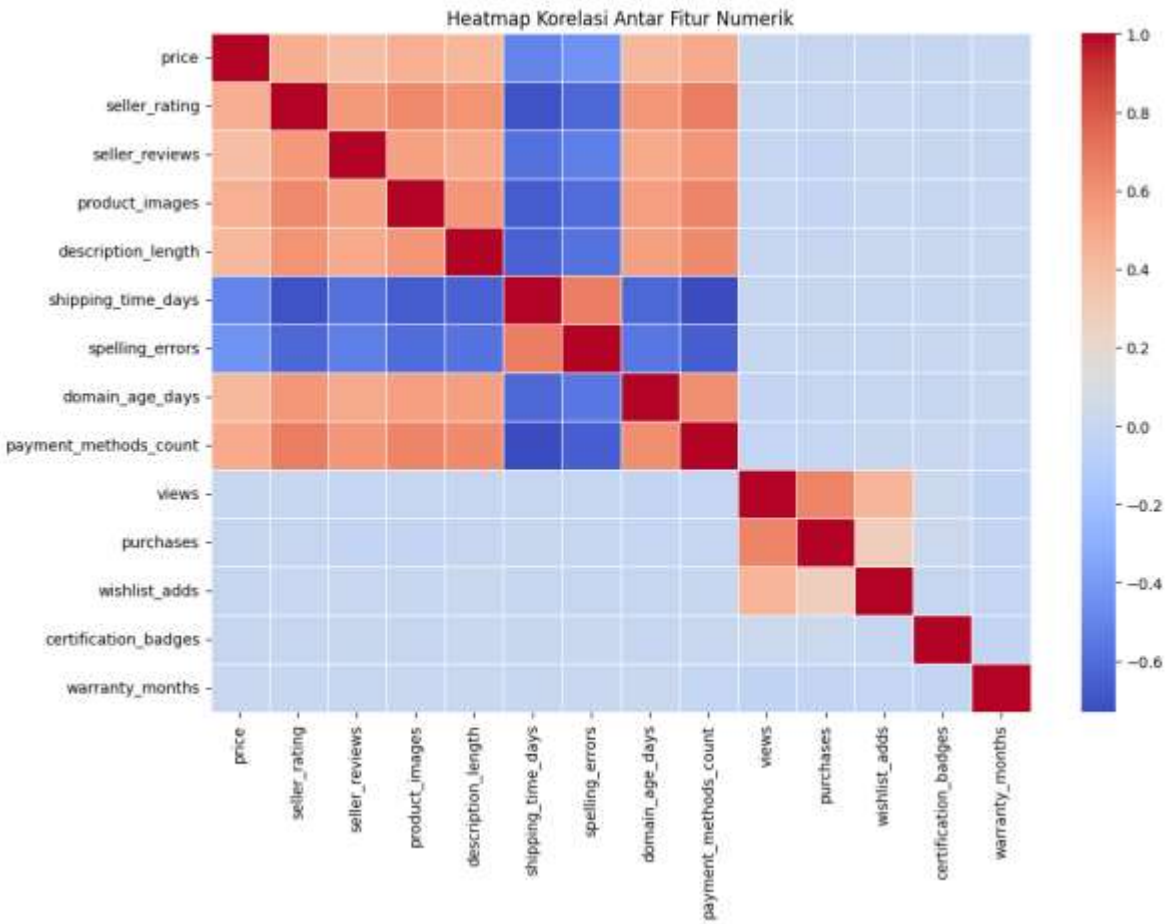


Gambar 1. 3 Grafik Perbandingan Harga

- Korelasi Antar Fitur

  Analisis korelasi antar fitur dilakukan menggunakan visualisasi **heatmap korelasi** untuk melihat hubungan antar atribut numerik. Hasil analisis menunjukkan bahwa sebagian besar fitur memiliki korelasi rendah hingga sedang, sehingga risiko multikolinearitas relatif kecil. Beberapa fitur yang berkaitan dengan reputasi penjual dan aktivitas transaksi, seperti jumlah ulasan penjual, jumlah pembelian, dan jumlah tampilan produk, menunjukkan korelasi positif satu sama lain. Hal ini menunjukkan bahwa fitur-fitur tersebut saling berkaitan dalam merepresentasikan tingkat kepercayaan terhadap penjual.

  Tidak ditemukan hubungan yang sangat kuat antar fitur, sehingga risiko multikolinearitas relatif kecil. Dengan demikian, fitur-fitur numerik yang digunakan masih relevan dan dapat dipertahankan dalam proses pemodelan klasifikasi.



Gambar 1. 4 Grafik Heatmap Korelasi

1.4 Pemilihan dan Penerapan Algoritma

Algoritma utama yang digunakan dalam penelitian ini adalah **C4.5 (Decision Tree)**. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang mampu menangani data numerik dan kategorikal serta memiliki kemampuan dalam membangun model klasifikasi berbasis pohon keputusan.

C4.5 dipilih karena sesuai untuk permasalahan **klasifikasi**, khususnya dalam mengidentifikasi produk palsu dan produk asli berdasarkan berbagai atribut produk dan penjual. Selain itu, algoritma ini mampu menangani hubungan data yang bersifat **non-linear** serta menghasilkan model yang mudah dipahami dan diinterpretasikan.

Pemilihan algoritma C4.5 didasarkan pada beberapa pertimbangan berikut:

- Algoritma C4.5 sangat sesuai untuk permasalahan klasifikasi biner, seperti klasifikasi produk palsu dan produk asli.

- Mampu menangani data dengan atribut numerik maupun kategorikal secara efektif.
- Memiliki kemampuan untuk memodelkan hubungan non-linear antar fitur.
- Hasil model berupa pohon keputusan yang mudah dipahami, sehingga memudahkan analisis dan interpretasi hasil klasifikasi.

Dengan karakteristik tersebut, algoritma C4.5 dinilai tepat untuk digunakan dalam penelitian ini.

Dalam implementasinya menggunakan Python, algoritma C4.5 direpresentasikan dengan **Decision Tree Classifier** pada library `scikit-learn`. Parameter utama yang digunakan dalam proses pemodelan antara lain:

- **criterion**: digunakan untuk menentukan ukuran kualitas pemisahan data, yaitu `entropy`, yang sesuai dengan pendekatan C4.5.
- **max_depth**: membatasi kedalaman pohon keputusan untuk menghindari overfitting.
- **min_samples_split**: menentukan jumlah minimum sampel yang diperlukan untuk membagi node.
- **random_state**: digunakan untuk memastikan hasil pemodelan dapat direproduksi.

Parameter-parameter tersebut membantu menghasilkan model yang stabil dan memiliki performa klasifikasi yang optimal.

Selain algoritma utama C4.5, beberapa algoritma lain juga diuji sebagai pembanding untuk mengevaluasi performa model klasifikasi.

Tabel 1. 3 Algoritma yang di ujikan

Algoritma	Library Python	Tujuan
K-Nearest Neighbors (KNN)	sklearn.neighbors	Klasifikasi
Random Forest	sklearn.ensemble	Klasifikasi & feature importance
Support Vector Machine (SVM)	sklearn.svm	Klasifikasi data non-linear

Pengujian beberapa algoritma dilakukan untuk membandingkan performa dan memastikan bahwa algoritma yang dipilih memberikan hasil yang paling optimal dalam mengklasifikasikan produk palsu dan produk asli.

1.5 Pengujian dan Evaluasi Model

- Metode Evaluasi Model

  Penelitian ini merupakan permasalahan **klasifikasi biner**, yaitu mengklasifikasikan produk menjadi produk palsu dan produk asli. Oleh karena itu, metode evaluasi yang digunakan dalam penelitian ini meliputi **accuracy, precision, recall, F1-score**, serta **confusion matrix**.

  Accuracy digunakan untuk mengukur tingkat ketepatan model secara keseluruhan, sedangkan precision dan recall digunakan untuk melihat kemampuan model dalam mengklasifikasikan masing-masing kelas dengan benar. F1-score digunakan sebagai ukuran keseimbangan antara precision dan recall, sehingga memberikan gambaran performa model secara lebih komprehensif.

- Hasil Evaluasi Model

  Berdasarkan hasil pengujian menggunakan data uji, algoritma **C4.5 (Decision Tree)** menunjukkan performa klasifikasi yang sangat baik. Tabel berikut menyajikan hasil evaluasi model berdasarkan metrik klasifikasi yang digunakan.

Tabel 1. 4 Hasil Evaluasi Model

Kelas	**Precision**	**Recall**	**F1-Score**	**Support**
Produk Asli	1.00	1.00	1.00	719
Produk Palsu	1.00	1.00	1.00	281
Accuracy			**1.00**	1000

1.6 Analisis dan Interprestasi Hasil

Berdasarkan hasil pengujian dan evaluasi model, algoritma **C4.5 (Decision Tree)** menunjukkan performa yang paling optimal dalam mengklasifikasikan produk palsu dan produk asli. Hal ini ditunjukkan oleh nilai accuracy, precision, recall, dan F1-score yang mencapai nilai maksimal. Kemampuan C4.5 dalam menangani hubungan data

yang bersifat non-linear serta memanfaatkan atribut numerik dan kategorikal secara efektif menjadikannya sesuai untuk karakteristik dataset yang digunakan dalam penelitian ini.

Hasil analisis juga menunjukkan bahwa beberapa fitur memiliki pengaruh yang signifikan terhadap proses klasifikasi. Fitur harga produk memperlihatkan perbedaan yang jelas antara produk palsu dan produk asli, di mana produk palsu cenderung memiliki harga yang lebih rendah dan variasi harga yang lebih besar. Selain itu, fitur yang berkaitan dengan reputasi penjual, seperti rating penjual dan jumlah ulasan, turut berperan penting dalam membedakan keaslian produk. Pola transaksi, termasuk jumlah pembelian dan tampilan produk, juga memberikan kontribusi dalam mengidentifikasi perilaku penjual yang mencurigakan.

Secara umum, model klasifikasi yang dibangun dapat dikatakan memiliki kualitas yang sangat baik berdasarkan hasil evaluasi yang diperoleh. Model mampu mengklasifikasikan data uji secara konsisten tanpa kesalahan yang signifikan. Namun demikian, performa yang sangat tinggi ini perlu diinterpretasikan secara hati-hati karena berpotensi mengindikasikan terjadinya overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data yang digunakan. Meskipun tidak ditemukan indikasi underfitting, pengujian tambahan menggunakan data yang lebih beragam atau metode validasi silang disarankan untuk memastikan kemampuan generalisasi model.

Dari sisi domain permasalahan, hasil analisis menunjukkan bahwa produk palsu umumnya memiliki karakteristik tertentu, seperti harga yang lebih rendah, reputasi penjual yang kurang baik, serta pola transaksi yang tidak konsisten. Insight ini menunjukkan bahwa pendekatan data mining dapat dimanfaatkan secara efektif untuk mendukung proses deteksi produk palsu pada platform penjualan online. Dengan demikian, hasil penelitian ini berpotensi menjadi dasar pengembangan sistem pendukung keputusan yang dapat meningkatkan keamanan dan kepercayaan konsumen dalam transaksi daring.

1.7 Kesimpulan dan Rekomendasi

- Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa tujuan penelitian untuk mengklasifikasikan produk palsu dan produk asli menggunakan

pendekatan data mining telah berhasil dicapai. Proses data mining yang meliputi tahap preprocessing data, pemodelan, serta evaluasi model mampu menghasilkan model klasifikasi dengan performa yang sangat baik.

Algoritma **C4.5 (Decision Tree)** terbukti menjadi model terbaik dalam penelitian ini. Hal ini ditunjukkan oleh nilai accuracy, precision, recall, dan F1-score yang mencapai nilai maksimal pada data uji. Kemampuan algoritma C4.5 dalam menangani data numerik dan kategorikal serta memodelkan hubungan non-linear menjadikannya sesuai dengan karakteristik dataset penjualan produk palsu yang digunakan.

Selain itu, hasil analisis menunjukkan bahwa fitur-fitur seperti harga produk, reputasi penjual, dan pola transaksi memiliki pengaruh yang signifikan dalam membedakan produk palsu dan produk asli. Dengan demikian, kombinasi fitur-fitur tersebut mampu mendukung proses klasifikasi secara efektif.

- Rekomendasi

Meskipun hasil penelitian menunjukkan performa model yang sangat baik, beberapa pengembangan dapat dilakukan pada penelitian selanjutnya untuk meningkatkan kualitas dan keandalan model. Pertama, disarankan untuk menggunakan dataset dengan jumlah data yang lebih besar dan lebih beragam agar model memiliki kemampuan generalisasi yang lebih baik terhadap data di dunia nyata.

Kedua, pengaturan parameter model (*hyperparameter tuning*) dapat dilakukan untuk memperoleh struktur pohon keputusan yang lebih optimal dan mengurangi potensi overfitting. Selain itu, apabila pada penelitian selanjutnya ditemukan ketidakseimbangan kelas, maka disarankan untuk menerapkan teknik penyeimbangan data seperti oversampling atau undersampling.

Terakhir, penelitian selanjutnya juga dapat mempertimbangkan penggunaan algoritma lain sebagai pembanding, seperti Random Forest, Support Vector Machine, atau metode ensemble lainnya, guna memperoleh performa klasifikasi yang lebih stabil dan robust.

https://colab.research.google.com/drive/1e-accSYtBgrJQMfyvDWBjT9p237A5QEJ?usp=sharing