

Profile Penjualan Hoodie H&M Menggunakan Ekstraksi Fitur dan Kombinasi Pendekatan Kluster

Rosyid Amrullah¹, Tjokroda Agung Budi Wirayuda²

Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

rosyidamrullah@students.telkomuniversity.ac.id¹, cokagung@telkomuniversity.ac.id²

Abstrak

Perkembangan teknologi digital telah mengubah cara konsumen berbelanja, termasuk dalam industri fesyen. H&M, sebagai retailer fesyen global, menghadapi tantangan dalam memahami preferensi konsumen yang terus berubah. Penelitian ini bertujuan untuk mengidentifikasi karakteristik visual hoodie H&M menggunakan CNN (Inception V3) dan mengelompokkan produk berdasarkan fitur tertentu melalui analisis clustering. Data gambar dan penjualan hoodie dari tahun 2018-2020 yang diambil dari Kaggle.com digunakan untuk menganalisis karakteristik visual dan mengelompokkan produk menggunakan K-means. Selain itu, model Random Forest digunakan untuk menganalisis variabel yang mempengaruhi sales profiling hoodie H&M, mencakup fitur visual yang diekstraksi oleh CNN serta kombinasi cluster. Hasil penelitian menunjukkan bahwa model memberikan performa terbaik dengan nilai MAPE sebesar 12.66%, RMSE sebesar 53.867, dan MAE sebesar 43.943. Jumlah cluster terbukti berpengaruh signifikan dalam membangun model profiling penjualan yang akurat. Penelitian ini berfokus pada hubungan antara karakteristik visual produk dan data penjualan, tanpa mempertimbangkan faktor eksternal seperti strategi pemasaran atau tren fesyen global, dengan tujuan mengembangkan sistem profiling penjualan yang efektif bagi H&M.

Kata Kunci : Fashion Retail, Computer Vision, Random Forest, Sales Profiling, Machine learning in Fashion, Inception V3, Visual Feature Extraction

Abstract

The digital technology revolution has transformed consumer shopping behavior, including in the fashion industry. H&M, as a global fashion retailer, faces challenges in understanding and adapting to rapidly changing consumer preferences. In this dynamic fashion industry, predicting hoodie sales performance has become increasingly complex due to these fast-evolving preferences. This study aims to identify the visual characteristics of H&M hoodies using CNN (Inception V3) and cluster the products based on specific features through clustering analysis. Hoodie image and sales data from 2018-2020, sourced from Kaggle.com, are used to analyze visual characteristics and cluster the products using K-means. Additionally, a Random Forest model is employed to analyze various variables affecting H&M hoodie sales profiling, including visual features extracted by CNN, along with cluster combinations, leaf parameters, and depth in the Random Forest. The experimental results indicate that the model performs relatively better than other K values, as indicated by the lowest MAPE of 12.66%, RMSE of 53.867, and MAE of 43.943. This study focuses on the relationship between the product's visual characteristics and sales data, without considering external factors such as marketing strategies or global fashion trends, aiming to develop an accurate sales profiling system for H&M.

Keywords : Fashion Retail, Computer Vision, Random Forest, Sales Profiling, Machine learning in Fashion, Inception V3, Visual Feature Extraction

1. Pendahuluan

Di tengah persaingan yang ketat, kemampuan untuk memahami pasar dan mengoptimalkan pengelolaan stok menjadi faktor kunci dalam menjaga daya saing perusahaan. Frekuensi kedatangan produk baru di pasaran semakin tinggi, terutama dalam industri fesyen yang sangat dinamis. Setiap produk baru harus mampu bersaing secara langsung dalam hal penjualan dan permintaan. Persaingan ini tidak hanya terjadi di antara berbagai merek, tetapi juga di dalam satu merek yang sama[3]. Produk-produk baru harus menarik perhatian konsumen yang sudah terbiasa dengan produk sebelumnya. Tantangan ini semakin kompleks bagi para pengambil keputusan dalam industri fesyen, yang harus mampu membuat keputusan cepat dan tepat untuk memastikan produk yang tepat tersedia di pasar pada waktu yang tepat.

Konsumen modern tidak lagi hanya mencari produk yang berkualitas, tetapi juga menginginkan pengalaman belanja yang cepat, mudah, dan personal. Dalam konteks produk hoodie, tantangan ini semakin jelas terlihat. Perubahan preferensi konsumen yang cepat, dikombinasikan dengan ketidakmampuan untuk mengidentifikasi tren dalam jangka waktu tertentu, dapat menyebabkan kesulitan dalam memprediksi kinerja penjualan produk.

Produk fast fashion, khususnya hoodie, memiliki siklus hidup yang relatif singkat, di mana perubahan preferensi konsumen dapat terjadi dengan cepat[4]. Konsumen mungkin tiba-tiba beralih dari satu gaya atau desain tertentu ke gaya atau desain lainnya, sehingga menciptakan ketidakpastian dalam memprediksi kinerja penjualan produk baru. Memperkirakan jumlah permintaan produk yang optimal selalu menjadi perhatian utama dalam industri retail. Estimasi yang akurat biasanya didasarkan pada analisis pola historis penjualan. Namun, dalam industri yang sangat dinamis seperti fesyen, mengandalkan pola historis saja seringkali tidak cukup. Ketidakmampuan untuk mengidentifikasi jumlah permintaan produk dengan benar dapat menyebabkan dua skenario yang merugikan yaitu kekurangan stok yang dapat mengakibatkan hilangnya potensi penjualan, atau kelebihan stok yang dapat menyebabkan biaya inventori yang berlebihan dan penurunan keuntungan. Kekurangan stok dapat membuat perusahaan kehilangan kesempatan untuk memenuhi permintaan konsumen, sementara kelebihan stok dapat menyebabkan penumpukan barang yang tidak terjual, yang akhirnya menurunkan nilai produk dan memengaruhi profitabilitas.

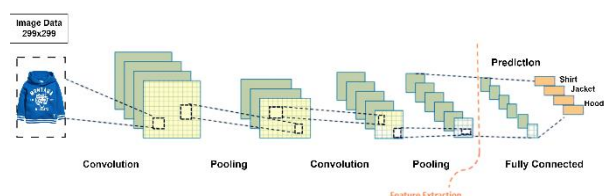
Oleh karena itu, penting bagi perusahaan seperti H&M untuk menerapkan pendekatan yang lebih canggih dalam mengelola stok dan memprediksi permintaan. Salah satu pendekatan yang dapat digunakan adalah melalui analisis data yang lebih mendalam, seperti ekstraksi fitur visual dari gambar produk dan analisis clustering. Dengan memanfaatkan teknologi seperti Convolutional Neural Network (CNN), penelitian ini diharapkan dapat membantu perusahaan memberikan pemahaman lebih dalam mengenai bagaimana karakteristik visual produk—seperti hoodie dari H&M—dapat memengaruhi pola penjualan, menjadi semakin relevan. Memanfaatkan pendekatan ekstraksi fitur visual menggunakan CNN dan analisis clustering diharapkan memberikan wawasan dan informasi bagi perusahaan dalam menghadapi tantangan ini.

Selain pendekatan berbasis CNN dan clustering, model Random Forest juga digunakan sebagai alat yang sangat bermanfaat dalam mengatasi tantangan pengelolaan stok dan prediksi permintaan yang dihadapi oleh H&M. Random Forest, sebagai metode ensemble learning yang menggabungkan multiple decision trees, memiliki kemampuan yang kuat dalam menangani data kompleks[5]. Random Forest dapat digunakan untuk menganalisis berbagai variabel yang mempengaruhi penjualan hoodie H&M, tidak hanya terbatas pada fitur visual yang diekstraksi oleh CNN, tetapi juga mencakup faktor-faktor lain seperti data penjualan. Keunggulan Random Forest terletak pada kemampuannya untuk menangani interaksi antar variabel dan resistensinya terhadap overfitting[6].

2. Kajian Pustaka dan Metodologi Penelitian

2.1. Penelitian Terkait

Penelitian ini menggunakan deep learning untuk mengekstraksi fitur visual dari gambar produk hoodie H&M melalui Convolutional Neural Network (CNN), khususnya menggunakan model Inception v3. Ekstraksi fitur ini bertujuan untuk mengubah gambar produk menjadi vektor fitur yang merepresentasikan karakteristik visualnya, memungkinkan analisis lebih lanjut. Inception v3 dipilih karena telah dilatih pada ribuan objek, termasuk pakaian, sehingga mampu menghasilkan akurasi tinggi dalam pengenalan gambar.



Setelah ekstraksi fitur, jumlah fitur yang besar dapat menimbulkan tantangan komputasi. Untuk mengatasi ini, diterapkan metode K-Means clustering yang bertujuan untuk mereduksi dimensi data dan mengelompokkan artikel hoodie berdasarkan kemiripan fitur. Dengan cara ini, kompleksitas data berkurang tanpa mengorbankan informasi penting yang diperlukan untuk analisis lebih lanjut.

Selanjutnya, metode Random Forest digunakan untuk menganalisis variabel yang memengaruhi penjualan hoodie, dengan memanfaatkan fitur hasil clustering sebagai input. Beberapa parameter penting dalam Random Forest, seperti jumlah pohon ($n_estimators$), jumlah daun ($leaf$), dan kedalaman pohon ($depth$), dioptimalkan untuk meningkatkan kinerja model. Penentuan jumlah pohon yang lebih besar dapat menghasilkan model yang lebih akurat, namun perlu diimbangi dengan kontrol terhadap jumlah daun dan kedalaman pohon agar menghindari overfitting, di mana model terlalu sesuai dengan data pelatihan sehingga sulit untuk melakukan generalisasi terhadap data baru. Dengan kombinasi pendekatan CNN, K-Means, dan Random Forest, penelitian ini memberikan wawasan mengenai bagaimana karakteristik visual hoodie dapat dihubungkan dengan performa penjualannya, membantu dalam pengambilan keputusan yang lebih akurat terkait pengelolaan stok dan prediksi permintaan.

2.2 Metodologi Penelitian

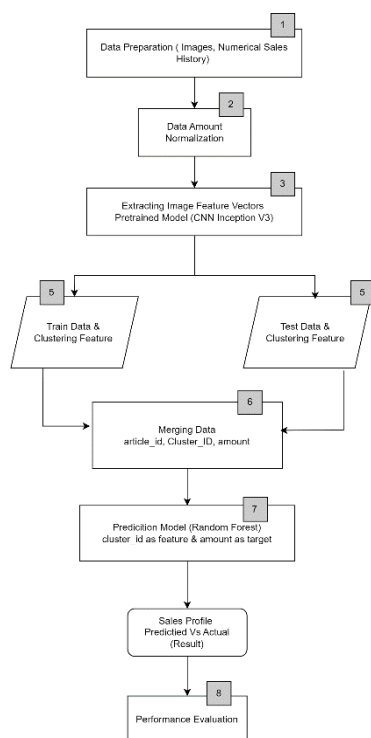


Diagram diatas merupakan ilustrasi dari proses pembangunan model yang diusulkan, mencakup langkah-langkah seperti persiapan dataset, normalisasi data, ekstraksi fitur, pemisahan data, klustering, dan evaluasi model. Setiap langkah ini sangat penting untuk keseluruhan proses. Rincian lebih lanjut tentang setiap langkah disajikan dalam subbagian berikut.

2.3 Dataset Preparation

Dataset yang digunakan dalam penelitian ini terdiri dari data artikel H&M Global yang mencakup periode dari tahun 2018 hingga 2020. Data ini diperoleh dari Kaggle.com dan memiliki cakupan yang luas, meliputi 2.311 gambar artikel hoodie serta data transaksi penjualan historis yang terkait dengan setiap produk tersebut. Namun, perlu diperhatikan bahwa data penjualan historis dalam dataset ini mencakup semua transaksi untuk berbagai artikel H&M, bukan hanya hoodie. Oleh karena itu, langkah penting yang perlu dilakukan adalah menyortir data lebih lanjut agar fokus pada transaksi yang relevan dengan penelitian, yakni penjualan artikel hoodie.

Data Atribut	Deskripsi
Item no	Article_id
Image	299 x 299 x 3 pixel
Amount	Jumlah Hoodie Terjual

Pensortiran data menghasilkan 144 artikel hoodie dengan jumlah penjualan dari tahun ke tahun, di mana setiap artikel memiliki total penjualan (amount) pada range 300 hingga 500 item. Hal ini bertujuan untuk meminimalkan gap yang terlalu jauh dan untuk mencapai tujuan studi yang lebih komprehensif. Contoh data gambar hoodie terdapat pada Gambar 3 di bawah ini. Item no diidentifikasi sebagai nomor unik dari setiap article_id, sedangkan image adalah ukuran pixel yang dipakai untuk pembangunan model, yaitu 299x299x3, yang artinya panjang dan lebar senilai 299 dan 3 adalah channel warna yang digunakan, yaitu RGB. Yang terakhir adalah amount, yaitu total penjualan hoodie.

2.4 Data Amount Normalization

Dalam konteks normalisasi penjualan, data penjualan dengan rentang angka yang besar, seperti dari 300 hingga 500, dapat menyebabkan perbedaan skala yang signifikan di antara fitur-fitur yang digunakan dalam model. Perbedaan skala ini dapat memengaruhi kinerja model, terutama pada algoritma yang sensitif terhadap skala data, seperti regresi linier dan Random Forest, yang akan digunakan dalam model profiling. Untuk mengatasi masalah ini, normalisasi data dilakukan dengan mengubah rentang nilai penjualan menjadi rentang 0 hingga 1.

2.5 Ekstraksi Fitur

Model InceptionV3 digunakan tanpa lapisan klasifikasi (include_top=False) dan dengan global average pooling (pooling='avg'). Pooling ini mereduksi output dari layer terakhir, yang berbentuk (8, 8, 2048), menjadi vektor dengan panjang tetap 2048 untuk setiap gambar. Oleh karena itu, setiap gambar

yang diproses oleh model akan menghasilkan vektor fitur berukuran 2048.

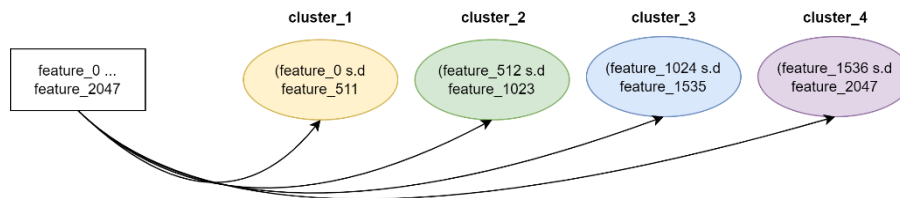
Selanjutnya, dataset gambar disusun dalam bentuk daftar path gambar, yang kemudian diubah menjadi generator dataset menggunakan fungsi `tf.data.Dataset`. Setiap gambar dalam dataset diproses menggunakan fungsi `load_and_preprocess_image` yang memuat dan mempersiapkan gambar untuk digunakan oleh model. Dataset ini di-batch dalam ukuran tertentu (dalam hal ini 25 gambar per batch) untuk menghindari kehabisan memori selama proses ekstraksi fitur.

2.6 Data Splitting & Clustering

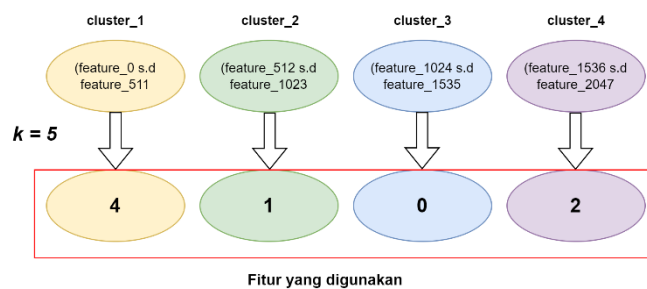
Hasil ekstraksi fitur dari 144 artikel hoodie menghasilkan 2048 fitur per artikel, sehingga total keseluruhan hasil ekstraksi fitur mencapai 294.912 fitur. Jumlah yang sangat besar ini dapat menimbulkan tantangan signifikan dalam hal komputasi. Proses pengolahan data dengan jumlah fitur sebesar ini berpotensi menjadi sangat kompleks dan memerlukan waktu yang cukup lama untuk diproses. Oleh karena itu, untuk mengatasi tantangan ini dan meningkatkan efisiensi analisis, diterapkan metode clustering sebagai langkah untuk mereduksi dimensi data tanpa mengorbankan informasi penting yang terkandung di dalamnya.

Langkah pertama adalah memisahkan fitur gambar dari kolom `article_id`. Fitur-fitur ini mewakili karakteristik gambar yang telah diproses sebelumnya, sementara `article_id` berfungsi sebagai identitas unik untuk setiap titik data. Data kemudian dibagi menjadi set pelatihan dan pengujian dengan pembagian 80-20.

Langkah selanjutnya adalah membagi data yang sangat besar ini menjadi subset-subset yang lebih kecil dan lebih mudah dikelola. Sebagai contoh, data tersebut dibagi menjadi 4 kluster, di mana setiap kluster terdiri dari 512 fitur. Proses ini digambarkan pada Gambar di bawah ini.

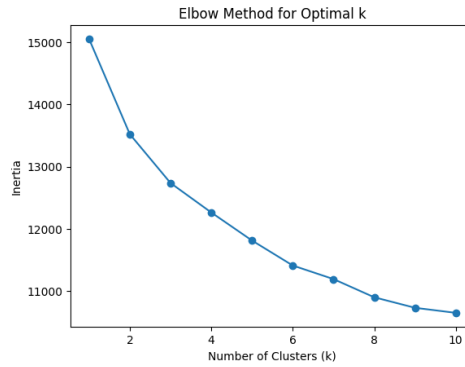


Setelah grup cluster (`Cluster_ID`) diidentifikasi dalam beberapa grup, langkah berikutnya adalah menentukan jumlah cluster dalam grup (Nilai `K`) yang optimal. Setiap subset dievaluasi dengan variasi jumlah cluster `K` yang diuji, yaitu $k=5$ dan $k=10$. Performansi diuji berdasarkan nilai error yang paling optimal, dengan tujuan untuk mengidentifikasi fitur yang akan dimasukkan dalam model regressor nantinya. Ilustrasi untuk menentukan jumlah cluster dalam grup (Nilai `K`) dapat dilihat pada Gambar.



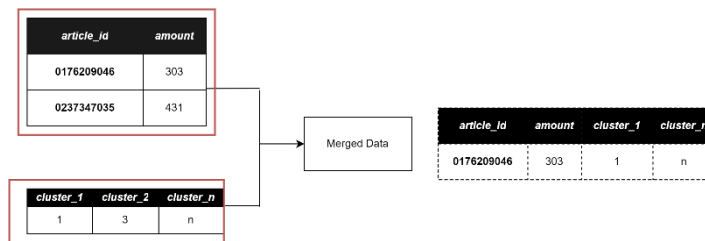
Sehingga, 512 fitur yang ada pada setiap grup cluster (`Cluster_ID`) akan diubah dan disesuaikan dengan grup dalam kluster (Nilai `K`) yang paling sesuai. Dengan demikian, grup cluster (`Cluster_ID`) dan grup dalam kluster (Nilai `K`) akan digunakan sebagai fitur untuk model regressor.

Untuk menentukan (Nilai `K`) yang optimal, dapat digunakan Elbow Method, yang diilustrasikan pada Gambar 8. Metode ini populer untuk menentukan titik optimal jumlah cluster, memberikan keseimbangan terbaik antara keakuratan pengelompokan dan kompleksitas komputasi. Penjelasan mengenai Elbow Method dapat dilihat pada Gambar di bawah ini.



2.7 Merging Data

Proses penggabungan ini dilakukan untuk menghasilkan sebuah dataset gabungan yang mencakup setiap article_id dengan amount yang terkait, serta nilai-nilai cluster yang merepresentasikan karakteristik atau kelompok dari artikel tersebut. Tujuan dari penggabungan ini adalah agar data cluster dapat digunakan sebagai fitur (independent variable), sedangkan amount akan digunakan sebagai target (dependent variable). Dataset gabungan ini dapat digunakan untuk keperluan analisis lebih lanjut atau dalam model pembelajaran mesin untuk melakukan prediksi berdasarkan cluster.



2.8 Sales Profiling Model

Sales profiling model adalah proses yang digunakan untuk memprediksi jumlah penjualan hoodie berdasarkan data fitur gambar yang telah melalui proses klusterisasi. Dalam pendekatan ini, model pembelajaran mesin Random Forest digunakan dengan data cluster sebagai fitur dan penjualan yang dinormalisasi sebagai target. Langkah pertama melibatkan pemisahan fitur (X) dan target (y) dari dataset, di mana fitur merepresentasikan data kluster, sementara article_id dan amount_normalized digunakan sebagai target. Selanjutnya, model Random Forest Regressor diinisialisasi dengan n_estimators = 100, yang berarti akan membangun 100 pohon keputusan, dengan random_state diatur untuk menjaga konsistensi hasil. Setelah inisialisasi, model dilatih menggunakan data latih untuk mempelajari pola hubungan antara fitur dan jumlah penjualan hoodie. Setelah proses pelatihan selesai, model digunakan untuk memprediksi data uji yang belum pernah dilihat sebelumnya, dengan hasil prediksi disimpan dalam variabel y_pred. Evaluasi model dilakukan dengan membandingkan hasil prediksi dengan nilai aktual untuk mengukur akurasi model dalam memprediksi penjualan hoodie berdasarkan data kluster.

2.9 Matriks Evaluasi

Dalam penelitian ini, digunakan beberapa matriks evaluasi untuk mengukur performa model Random Forest dalam memprediksi jumlah penjualan hoodie. Matriks yang digunakan meliputi Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R²), dan Mean Absolute Percentage Error (MAPE). Masing-masing matriks ini memberikan perspektif yang berbeda tentang akurasi dan efektivitas model prediksi.

Root Mean Square Error (RMSE) adalah akar kuadrat dari MSE dan memberikan interpretasi yang lebih intuitif karena hasilnya dalam satuan yang sama dengan data asli. RMSE memungkinkan kita untuk melihat seberapa jauh prediksi model dari nilai aktual secara rata-rata. Nilai RMSE yang lebih rendah menunjukkan bahwa model memiliki kinerja yang lebih baik dalam memprediksi penjualan hoodie. RMSE sangat berguna untuk memahami besarnya kesalahan prediksi model regresi. Rumus RMSE dituliskan sebagai berikut:

$$RMSE = \sqrt{MSE}$$

Mean Absolute Error (MAE) adalah metrik evaluasi yang digunakan untuk mengukur rata-rata

absolut dari kesalahan prediksi. Dalam konteks prediksi penjualan hoodie, MAE memberikan gambaran yang jelas tentang seberapa jauh rata-rata prediksi penjualan model menyimpang dari penjualan aktual. MAE lebih mudah diinterpretasikan karena berada dalam satuan yang sama dengan penjualan hoodie, sehingga membantu dalam memahami kesalahan prediksi secara langsung. Rumus MAE dituliskan sebagai berikut:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Dengan menggunakan MAE sebagai metrik evaluasi, kita dapat menilai seberapa dekat prediksi model dengan penjualan hoodie yang sebenarnya.

R-squared (R^2) atau koefisien determinasi mengukur proporsi variabilitas dalam data target yang dapat dijelaskan oleh model prediksi. Nilai R^2 berkisar antara 0 hingga 1, di mana nilai yang lebih mendekati 1 menunjukkan bahwa model mampu menjelaskan sebagian besar variabilitas dalam data. Rumus R^2 adalah:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Mean Absolute Percentage Error (MAPE) adalah matriks yang mengukur kesalahan prediksi dalam bentuk persentase, yang memungkinkan untuk memahami seberapa besar kesalahan relatif terhadap nilai aktual. Rumus MAPE adalah:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$$

3. Hasil dan Pembahasan

Dalam analisis hasil ini, pembahasan akan dibagi menjadi 2 skenario. Setiap skenario eksperimen akan dijelaskan secara spesifik pada bagian berikut:

Skenario Satu – secara eksplisit menjawab dari rumusan masalah satu yaitu ‘apakah kombinasi jumlah cluster mempengaruhi dalam membangun model profiling? Proses eksperimen skenario ini adalah memprediksi penjualan hoodie dengan menjumlahkan semua transaksi artikel hoodie, artinya model tidak mempertimbangkan aspek waktu seperti jumlah penjualan setiap bulan. Pendekatan ini bertujuan untuk melihat seberapa efektif model dapat mengidentifikasi hubungan antara kombinasi grup cluster (Cluster ID) dan kluster dalam grup (Nilai K) mempengaruhi model dalam memprediksi jumlah penjualan.

Skenario Dua - secara eksplisit menjawab dari rumusan masalah dua yaitu ‘Apakah kombinasi jumlah max_depth dan jumlah leaf pada Random Forest mempengaruhi performansi model profiling? Dengan menggunakan data eksperimen satu, yaitu jumlah transaksi semua artikel hoodie, beserta kombinasi grup cluster (Cluster Id) dan kluster dalam grup (Nilai K) yang paling optimal, lalu bereksperimen dengan kombinasi parameter lain pada model regresi yaitu jumlah leaf dan max depth pada Random Forest. Pendekatan ini bertujuan untuk melihat pengaruh nilai eror jika jumlah leaf dan max depth diubah dalam jumlah tertentu.

A. Skenario Satu

Pada eksperimen ini, terdapat 2 parameter yang diuji, yaitu jumlah grup cluster (ID cluster) dan jumlah kluster dalam grup cluster (Nilai K) untuk mengamati performa prediksi. Dalam skenario ini, kita ingin mencoba kombinasi grup cluster yang paling optimal. Dengan membagi jumlah fitur per artikel yang terdiri dari 2048 baris menjadi beberapa ID cluster. Pendekatan ini bertujuan untuk memahami apakah pengelompokan fitur ke dalam grup cluster dapat mempengaruhi kontribusi terhadap performa model prediksi.

Tabel 2 Eksperimen Skenario Satu Berdasarkan ID Cluster

ID Cluster	RMSE	MAE	R-squared (R2)	MAPE (%)
4	54.659	46.815	-0.209	13.34
8	53.867	43.943	-0.174	12.66
16	58.558	48.954	-0.387	14.28
32	53.454	44.409	-0.156	12.94
64	53.129	45.615	-0.142	13.13

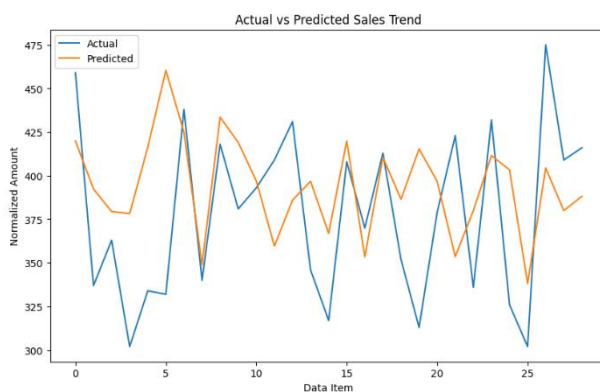
Berdasarkan analisis dari Tabel 2, terlihat bahwa eksperimen yang dilakukan pada berbagai ID cluster menunjukkan hasil yang cukup beragam. Menariknya, untuk ID cluster = 8 dengan jumlah grup dalam cluster (Nilai K) = 5, model prediksi menghasilkan nilai Mean Absolute Percentage Error (MAPE) yang paling rendah, yaitu sebesar 12.66%. Hal ini menunjukkan bahwa model dengan konfigurasi tersebut mampu memberikan prediksi yang relatif lebih akurat dibandingkan dengan konfigurasi lainnya. Namun, untuk mendapatkan pemahaman yang lebih menyeluruh, kita perlu mempertimbangkan metrik-metrik lain dan implikasinya terhadap penjualan hoodie. Misalnya, nilai Root Mean Squared Error (RMSE) sebesar 54.659 menunjukkan bahwa dalam beberapa kasus, perbedaan antara nilai prediksi dan nilai penjualan aktual bisa mencapai sekitar 55 unit hoodie. Hasil percobaan ini menunjukkan bahwa prediksi jumlah penjualan hoodie berdasarkan kombinasi jumlah grup cluster memberikan hasil yang bervariasi, sangat bergantung pada jumlah grup cluster yang digunakan.

Untuk pengujian nilai kluster dalam grup (nilai K) terhadap akurasi prediksi, dilakukan serangkaian pengujian dengan variasi jumlah K, yaitu K = 5, K = 10, dan K = 15, dan seterusnya. Sementara itu, jumlah grup cluster tetap dijaga sebanyak 8 grup Cluster ID. Pengujian ini bertujuan untuk mengevaluasi sejauh mana variasi dalam jumlah cluster dalam grup dapat mempengaruhi keakuratan prediksi penjualan, serta mengidentifikasi konfigurasi cluster dalam grup yang paling efektif dalam konteks ini. Dengan mengeksplorasi berbagai nilai K.

Table 3 Eksperimen Skenario Satu Berdasarkan Jumlah Cluster Dalam Grup (Nilai K)

ID Cluster	k	RMSE	MAE	R-squared (R2)	MAPE (%)
8	5	53.867	43.943	-0.174	12.66
8	10	57.842	47.259	-0.354	13.47
8	15	56.801	47.689	-0.305	13.63
8	20	57.757	46.248	-0.350	13.48
8	25	53.165	44.208	-0.144	12.79

Dari hasil ini, terlihat bahwa model dengan jumlah cluster K = 5 memberikan performa yang relatif lebih baik dibandingkan dengan nilai K lainnya. Hal ini ditunjukkan oleh nilai MAPE yang paling rendah, yaitu sebesar 12.66%, serta nilai RMSE sebesar 53.867, dan MAE sebesar 43.943. Hasil dari evaluasi ini menunjukkan bahwa dengan konfigurasi grup cluster (Cluster ID) dan cluster dalam grup (Nilai K) yang tepat, model sudah mulai memberikan prediksi yang hampir sesuai dengan data aktual, sebagaimana dapat dilihat dari Gambar di bawah ini. Dengan demikian, penting untuk terus mempertimbangkan berbagai metrik evaluasi dan melakukan pengujian lebih lanjut untuk mengoptimalkan hasil.



B. Skenario Dua

Eksperimen Variasi Leaf Skenario Dua

Uji coba ini dirancang untuk mengevaluasi apakah penyesuaian pada parameter leaf dapat lebih lanjut mengurangi nilai error dan meningkatkan akurasi prediksi model, seperti yang ditunjukkan pada Tabel.

Tabel 4 Eksperimen Skenario Dua Berdasarkan Jumlah Leaf

Leaf	RMSE	MAE	R-squared (R2)	MAPE (%)
10	53.489	44.033	-0.158	12.83
50	54.374	44.233	-0.197	12.75
100	54.260	44.135	-0.192	12.72
150	54.260	44.135	-0.192	12.72
200	54.260	44.135	-0.192	12.72

Dari hasil uji coba yang telah dilakukan, terlihat bahwa konfigurasi leaf pada nilai 100 menghasilkan performa terbaik, dengan Root Mean Squared Error (RMSE) menunjukkan rata-rata kesalahan prediksi sekitar 44 unit dari nilai aktual, dan Mean Absolute Percentage Error (MAPE) sebesar 12.72%. Angka-angka ini menunjukkan bahwa model dengan leaf 100 mampu memberikan sedikit penurunan error dibandingkan dengan variasi leaf lainnya. Meskipun terdapat peningkatan kecil dalam beberapa metrik error, hasil ini mengindikasikan bahwa variasi pada parameter leaf tidak memberikan peningkatan yang signifikan terhadap kinerja keseluruhan model.

Dalam upaya lain untuk mengoptimalkan kinerja model Random Forest, pengujian dilakukan dengan variasi kedalaman pohon, dimulai dari max depth 5 hingga 1000. Tujuan dari pengujian ini adalah untuk memahami bagaimana perubahan dalam kedalaman pohon mempengaruhi kemampuan model dalam memprediksi nilai target, serta untuk mengidentifikasi kedalaman optimal yang memberikan keseimbangan terbaik antara bias dan variansi. Pengujian ini dilakukan dengan mengatur parameter max depth pada berbagai nilai yang telah ditentukan, yaitu 5, 50, 100, 500, dan 1000. Hasil dari pengujian ini dirangkum dalam Tabel di bawah

Tabel 5 Eksperimen Skenario Dua Berdasarkan Variasi max depth

Max Depth	RMSE	MAE	R-squared (R2)	MAPE (%)
5	53.907	44.743	-0.163	12.70
50	53.867	43.943	-0.174	12.66
100	53.867	43.943	-0.174	12.66
500	53.867	43.943	-0.174	12.66
1000	53.867	43.943	-0.174	12.66

Berdasarkan hasil evaluasi model Random Forest dengan variasi nilai max depth, terlihat bahwa perubahan nilai max depth dari 5 hingga 1000 tidak memberikan perbedaan yang signifikan dalam metrik evaluasi utama seperti Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R²), dan Mean Absolute Percentage Error (MAPE). Pada max_depth = 5, model menunjukkan MAPE sebesar 12.70%. Ketika max_depth ditingkatkan ke 50, 100, 500, dan 1000, metrik evaluasi tetap stabil dengan RMSE 53.86, MAE 43.943, R² -0.174, dan MAPE 12.66%. Hasil ini menunjukkan bahwa setelah max_depth mencapai nilai tertentu (sekitar 50), peningkatan dalam kedalaman pohon tidak memberikan peningkatan kinerja model

4. Kesimpulan

Penelitian ini berhasil membangun model sales profiling menggunakan proses klustering dari hasil ekstraksi fitur gambar. Jumlah kombinasi grup cluster (Cluster ID) dan kluster dalam grup (Nilai K) dengan metode K-means yang sangat berpengaruh terhadap kinerja model prediksi. Ini dibuktikan dengan eksperimen skenario satu, dengan matriks evaluasi yang paling baik, MAPE yang bernilai 12.66% dengan Cluster ID = 8 dan Nilai K = 5. Eksperimen dengan parameter seperti jumlah leaf dan max_depth pada Random Forest terbukti tidak memberikan pengaruh signifikan pada nilai error dalam model prediksi.

Saran untuk penelitian selanjutnya adalah mengembangkan model lain seperti Extreme Gradient Boosting (XGBoost), Generative Adversarial Networks (GAN), atau model yang lebih canggih lainnya untuk meningkatkan akurasi prediksi penjualan. Selain itu, penelitian bisa menggunakan dataset yang lebih luas dan beragam untuk memperluas aplikasi model prediksi penjualan ke berbagai kategori produk pakaian lainnya. Alternatif lainnya adalah mengembangkan konsep seperti prediksi preferensi pelanggan berdasarkan gaya pakaian, dengan merekomendasikan produk yang mirip dengan yang pernah dibeli sebelumnya.

Daftar Pustaka

- [1] Reinartz, W., Wiegand, N., & Imschloss, M. (2019). The impact of digital transformation on the retailing value chain. *International Journal of Research in Marketing*, 36(3), 350–366. (Razaque et al., 2016). Razaque, A., Amsaad, F., Kumar, R., Abdulgader, M., Jagadabi, S. K., & Sheela, S. (2016). Pebble Watch security assessment. *2016 IEEE Long Island Systems, Applications and Technology Conference, LISAT 2016*. <https://doi.org/10.1109/LISAT.2016.7494138>
- [2] Lin, R. (2023). Digital fashion: Transforming design, technology, and industry. In *Proceedings of the 2023 5th International Conference on Literature, Art and Human Development (ICLAHD 2023)*, Atlantis Press, pp. 747–752. https://doi.org/10.2991/978-2-38476-170-8_84
- [3] Varanis, M., & Pederiva, R. (2015). Wavelet packet energy-entropy feature extraction and principal component analysis for signal classification. 3(Xxxv), 1–7. <https://doi.org/10.5540/03.2015.003.01.0471>
- [4] HubSpot. "What is a Competitive Analysis — and How Do You Conduct One?" 2024. URL: <https://blog.hubspot.com/marketing/competitive-analysis-kit>.
- [5] Kate Fletcher. "Sustainable fashion and textiles: design journeys". In: Earthscan (2008), p. 162.
- [6] Jehad Ali et al. "Random forests and decision trees". In: *International Journal of Computer Science Issues (IJCSI)* 9.5 (2012), p. 272
- [7] Learn AI. "Random Forest Algorithm — divakar1591." Medium, <https://medium.com/@divakar1591/random-forest-algorithm-772d3e5568af>. [Accessed 14-08-2024].
- [8] S. Thomassey, M. Happiette, and J. M. Castelain. "Mean-term Textile Sales Forecasting Using Families and Items Classification." *Studies in Informatics and Control*, vol. 12, no. 1, 2003, pp. 41–52.
- [9] Y. Arti and A. M. Arymurthy. "Face Spoofing Detection using Inception-v3 on RGB Modal and Depth Modal." *Jurnal Ilmu Komputer Dan Informasi*, vol. 16, no. 1, 2023, pp. 47–57. DOI: 10.21609/jiki.v16i1.1100. URL: <https://doi.org/10.21609/jiki.v16i1.1100>.
- [10] Sebastien Thomassey and Michel Happiette. "A Neural Clustering and Classification System for Sales Forecasting of New Apparel Items." *Applied Soft Computing*, vol. 7, no. 4, 2007, pp. 1177–1187.
- [11] K. Ferreira, B. Lee, and D. Simchi-Levi. "Analytics for an Online Retailer: Demand Forecasting and Price Optimization." *Manufacturing & Service Operations Management*, 2016.
- [12] Zhiming Liu et al. "Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, 2022, pp. 681–697.