Analisis Umpan Balik Pelanggan dan Efisiensi Produksi pada DataCo Smart Supply Chain

Lomba Analitik Big Data
FESMARO 2025

Winston Narada Kusumahadi - 12202023
Bryant Gabriel Effendi - 12302064
Farrany Sucitra Kusumahadi - 12402014

SWISS GERMAN UNIVERSITY
Prominence Tower Alam Sutera
Tangerang 15339
Indonesia

I. Latar Belakang

Di era industri 4.0, *smart supply chain* menjadi kunci dalam meningkatkan efisiensi produksi dan kepuasan pelanggan. DataCo, sebagai perusahaan yang menerapkan teknologi canggih seperti IoT dan *big data analytics*, menghadapi tantangan dalam menghubungkan umpan balik pelanggan dengan efisiensi produksinya. Umpan balik pelanggan berperan penting dalam menilai kualitas produk dan layanan, sementara efisiensi produksi mencerminkan sejauh mana perusahaan dapat memenuhi permintaan dengan optimal. Ketidakseimbangan antara keduanya dapat berdampak pada profitabilitas dan daya saing perusahaan.

Penelitian ini bertujuan untuk menganalisis hubungan antara umpan balik pelanggan dan efisiensi produksi dalam sistem *smart supply chain* DataCo. Dengan memahami pola dan wawasan yang diperoleh dari data pelanggan, perusahaan dapat mengoptimalkan operasionalnya dan meningkatkan kepuasan pelanggan. Hasil penelitian ini diharapkan dapat memberikan strategi berbasis data yang lebih efektif dalam menghadapi tantangan rantai pasok yang terus berkembang.

II. Metodologi

Metodologi yang digunakan dalam proyek ini bertujuan untuk menganalisis kepuasan pelanggan dan efisiensi produksi dengan memanfaatkan data dari **DataCo Smart Supply Chain** dan **ulasan Amazon**. Proses analisis dilakukan melalui beberapa tahapan sebagai berikut:

1. **Pembersihan Data (Data Cleaning)**
    - Menghapus **outlier** yang dapat mengganggu analisis.
    - Mengatasi **nilai kosong (missing values)** dengan teknik seperti imputasi atau penghapusan data yang tidak relevan.
2. **Pemfilteran Data (Data Filtering)**

- Menggunakan **Python** (dengan pustaka seperti Pandas dan NumPy) untuk menyaring data berdasarkan **kolom yang relevan** agar analisis lebih fokus dan akurat.

3. **Perhitungan dan Analisis**
   - Menghitung metrik penting seperti **Mean Absolute Percentage Error (MAPE)** untuk mengevaluasi akurasi prediksi.
   - Menggunakan **Log Loss** untuk menilai kualitas model klasifikasi jika diperlukan.
   - Menganalisis pola dan tren yang dapat memberikan wawasan mendalam terkait kepuasan pelanggan serta efisiensi rantai pasokan.
4. **Visualisasi Data**
   - Menggunakan pustaka untuk membuat grafik dan diagram agar hasil analisis lebih mudah dipahami.
   - Menampilkan hubungan antara variabel yang mempengaruhi kepuasan pelanggan dan efisiensi produksi melalui visualisasi data yang informatif.

Dengan metodologi ini, proyek ini dapat memberikan wawasan berbasis data yang dapat digunakan untuk meningkatkan kepuasan pelanggan dan mengoptimalkan efisiensi produksi.

III. Rincian Data dan Alat yang Digunakan

1. Deskripsi Dataset

Data yang digunakan dalam proyek ini terdiri dari dua sumber utama:

1. **Data Customer Amazon**: Menyediakan informasi tentang ratings yang diberikan oleh customer dengan 1 atau 2, dimana 1 digolongkan kategori buruk, sedangkan 2 termasuk di kategori baik.
2. **Data Smart Supply Chain**: Dataset *Smart Supply Chain* mencakup informasi mendetail tentang pesanan, pengiriman, pelanggan, produk, dan wawasan pasar. Data pesanan mencakup *Order Id, Tanggal Pemesanan, Status Pesanan, Keuntungan Per Pesanan*, dan *Total Penjualan*, yang digunakan untuk melacak pendapatan serta profitabilitas. Informasi pengiriman seperti *Hari untuk*

Pengiriman (Real), Hari untuk Pengiriman (Terjadwal), Mode Pengiriman, Status Pengiriman, dan *Risiko Keterlambatan Pengiriman* membantu dalam mengevaluasi efisiensi logistik. Data pelanggan meliputi *Customer Id, Nama Pelanggan, Email, Alamat, Segmen Pelanggan, serta Lokasi (Kota, Provinsi, Negara, dan Kode Pos)* yang memberikan wawasan tentang pola pembelian.

Informasi produk seperti *Product Id, Nama Produk, Deskripsi, Kategori, Departemen*, dan *Harga Produk* mendukung analisis inventaris dan penjualan. Setiap *Item Pesanan* juga memiliki atribut seperti *Product Id, Jumlah, Harga, Diskon, dan Rasio Keuntungan*, yang memungkinkan analisis lebih rinci terhadap kinerja penjualan. Selain itu, wawasan pasar dan geografis, termasuk *Wilayah Pasar, Lokasi Pesanan (Kota, Provinsi, Negara), Latitude, dan Longitude*, berguna untuk pengambilan keputusan strategis. Dataset ini sangat berharga untuk menganalisis efisiensi rantai pasok, pola pembelian pelanggan, serta kinerja bisnis secara keseluruhan.

2. **Proses Data Cleaning**

- **Menghapus kolom yang tidak relevan dan kolom kosong penuh**
  - drop_columns = [‘Customer Email’, ‘Customer Frame’, ‘Customer Lname’, ‘Customer Password’, ‘Product Description’, ‘Order Zipcode’]
    data.drop(columns=drop_columns, inplace=True)
- **Mengisi nilai hilang dengan nilai rata-rata**
  - for col in data.select_dtypes(include=[‘float64’, ‘int64’]).columns:
    data[col].fillna(data[col].mean(), inplace=True)
- **Encoding Kolom Kategorikal**
  - label_encoders={}
    for col in data.select_dtypes(include=[‘object’]).columns:
    le=LabelEncoder()
    data[col]=le.fit_transform(data[col].astype(str))
    label_encoders[col]=le

- **Standarisasi Fitur Numerik**
    - scaler=StandardScaler()
      numeric_cols=data.select_dtypes(includes=[‘float64’,’int64’]).columns
      scaled_data=scaler.fit_transform(data[numeric_cols])

3. **Alat Visualisasi**
    - Pandas: Digunakan untuk manipulasi dan analisis data dalam bentuk tabel (*DataFrame*). Berguna untuk membaca, membersihkan, dan mengelola data sebelum divisualisasikan.
    - Numpy: Digunakan untuk komputasi numerik, seperti operasi matriks dan vektor yang sering digunakan dalam analisis data dan machine learning.
    - Seaborn: Untuk membuat grafik statistik dengan tampilan yang lebih menarik dan informatif. Seaborn memudahkan analisis data dengan menyediakan berbagai fungsi untuk membuat grafik seperti heatmap, scatter plot, box plot, violin plot, histogram, dan pair plot.
    - Matplotlib.pyplot: Digunakan untuk membuat visualisasi statistik yang menarik dan informatif, termasuk ***heatmap***.
    - Joblib: Digunakan untuk menyimpan dan memuat model machine learning atau objek besar agar bisa digunakan kembali tanpa harus melatih ulang model.
    - Sklearn.preprocessing import StandardScaler, LabelEncoder:
        - **StandardScaler**: Digunakan untuk menstandarisasi fitur agar memiliki distribusi dengan *mean = 0* dan *standard deviation = 1*. Ini penting dalam algoritma machine learning yang peka terhadap skala data.
        - **LabelEncoder**: Digunakan untuk mengubah data kategorik menjadi angka.
    - Sklearn.cluster: Digunakan untuk melakukan *clustering* menggunakan algoritma ***K-Means***.
    - Sklearn.metrics:
        - Package sklearn.metrics dalam **scikit-learn** menyediakan berbagai matrik evaluasi untuk machine learning, termasuk **Davies Bouldin Score** dan **Calinski Harabasz score** yang digunakan untuk menilai kualitas hasil clustering tanpa label ground truth.

    - **Davies-Bouldin Score** mengukur seberapa baik cluster terbentuk dengan melihat rasio antara jarak intra-cluster dan inter-cluster (semakin rendah, semakin baik).
    - **Calinski-Harabasz Score** mengevaluasi kepadatan dan pemisahan antar cluster berdasarkan varians (semakin tinggi, semakin baik).
- Sklearn.decomposition import PCA:
    - ***Mengurangi Dimensi Data***: Berguna saat bekerja dengan dataset yang memiliki banyak fitur untuk menghindari *overfitting* dan meningkatkan efisiensi komputasi.
    - ***Menghilangkan Redundansi***: Membantu menghapus korelasi antar fitur dan mengekstrak informasi yang paling penting.
    - ***Meningkatkan Visualisasi Data***: Dengan mengurangi dimensi menjadi 2D atau 3D, kita bisa membuat grafik scatter plot dari data kompleks.

IV. Exploratory Data Analysis (EDA)

1. Pemilihan Model

Tahap pertama dalam proses analisis data adalah memuat dataset yang diperlukan. Dataset dimuat menggunakan library pandas dengan metode read_csv(), dan encoding yang digunakan adalah ISO-8859-1 untuk menangani karakter khusus.

Kode:

```
file_path = 'Dataset_Kedua.csv'
try:
   data = pd.read_csv(file_path, encoding='ISO-8859-1')
   print('Dataset berhasil dimuat!')
   print(data.info())
   print(data.describe())
except Exception as e:
   print(f'Error dalam memuat dataset: {e}')
```

Hasil:

```
Dataset berhasil dimuat!
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180519 entries, 0 to 180518
Data columns (total 53 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   Type                           180519 non-null  object
 1   Days for shipping (real)       180519 non-null  int64
 2   Days for shipment (scheduled)  180519 non-null  int64
 3   Benefit per order              180519 non-null  float64
 4   Sales per customer             180519 non-null  float64
 5   Delivery Status                180519 non-null  object
 6   Late_delivery_risk             180519 non-null  int64
 7   Category Id                    180519 non-null  int64
 8   Category Name                  180519 non-null  object
 9   Customer City                  180519 non-null  object
 10  Customer Country               180519 non-null  object
 11  Customer Email                 180519 non-null  object
 12  Customer Fname                 180519 non-null  object
 13  Customer Id                    180519 non-null  int64
 14  Customer Lname                 180511 non-null  object
 15  Customer Password              180519 non-null  object
 16  Customer Segment               180519 non-null  object
 17  Customer State                 180519 non-null  object
 18  Customer Street                180519 non-null  object
...
75%       199.990005            0.0
max      1999.989990            0.0
```

Dataset berhasil dimuat dengan baik, dan dilakukan inspeksi awal untuk melihat informasi dasar serta deskripsi statistik dari dataset.

2. Pemeriksaan Missing Values

Pada tahap ini, kami melakukan pemeriksaan data yang hilang pada setiap kolom dengan cara menghitung jumlah nilai yang hilang (isnull().sum()).

Untuk mempermudah analisis, kami juga membuat visualisasi bar plot menggunakan Seaborn untuk menunjukkan jumlah nilai yang hilang pada 20 kolom teratas.

Kode:

```
missing_data = data.isnull().sum().sort_values(ascending=False)
print('Missing Data per Kolom:')
print(missing_data)

plt.figure(figsize=(12, 8))
sns.barplot(x=missing_data.index[:20], y=missing_data.values[:20], palette='viridis')
plt.title('Jumlah Missing Values pada 20 Kolom Teratas')
plt.xticks(rotation=90)
```

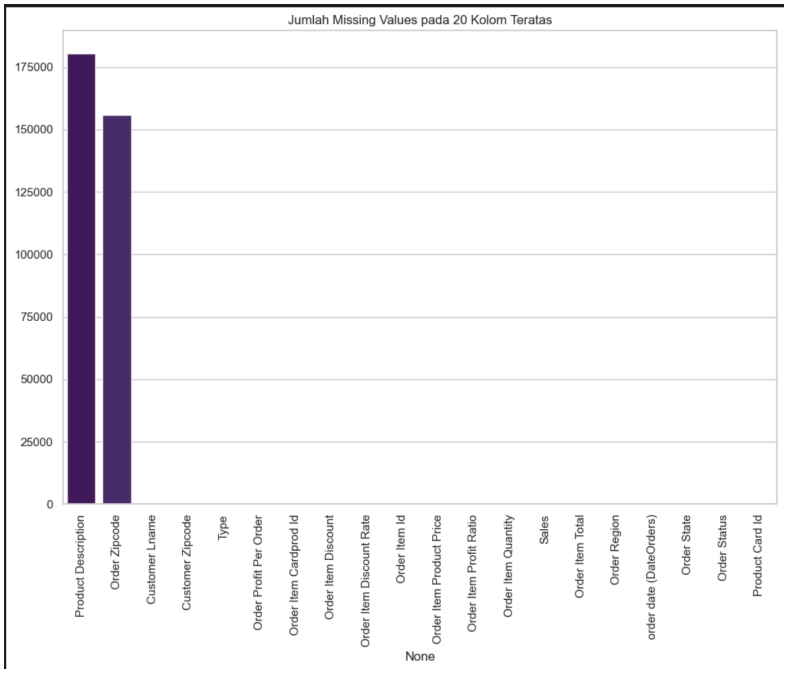
plt.show()

Hasil:

```
Missing Data per Kolom:
Product Description             180519
Order Zipcode                   155679
Customer Lname                       8
Customer Zipcode                     3
Type                                 0
Order Profit Per Order               0
Order Item Cardprod Id               0
Order Item Discount                  0
Order Item Discount Rate             0
Order Item Id                        0
Order Item Product Price             0
Order Item Profit Ratio              0
Order Item Quantity                  0
Sales                                0
Order Item Total                     0
Order Region                         0
order date (DateOrders)              0
Order State                          0
Order Status                         0
Product Card Id                      0
Product Category Id                  0
Product Image                        0
Product Name                         0
Product Price                        0
...
Market                               0
Order City                           0
Shipping Mode                        0
dtype: int64
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
C:\Users\winst\AppData\Local\Temp\ipykernel_4176\1973165464.py:8: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x

  sns.barplot(x=missing_data.index[:20], y=missing_data.values[:20], palette='viridis')
```



Dari hasil visualisasi, dapat dilihat bahwa beberapa kolom memiliki banyak nilai yang hilang, sehingga diputuskan untuk menghapus kolom dengan persentase nilai hilang yang sangat besar (lebih dari 50%).

Untuk kolom numerik dengan nilai hilang yang lebih sedikit, dilakukan imputasi dengan nilai rata-rata.

3. Statistik Deskriptif

Analisis statistik deskriptif dilakukan pada kolom numerik untuk memahami distribusi data. Statistik yang dihitung meliputi:

- Mean (Rata-rata)
- Standard Deviation (Standar Deviasi)
- Minimum dan Maksimum
- Kuartil (25%, 50%, 75%)

Kode:

```
print('Statistik Deskriptif:')
print(data.describe().T)
```

Hasil:

```
Statistik Deskriptif:
                              count          mean           std  \
Days for shipping (real)     180519.0      3.497654      1.623722
Days for shipment (scheduled)180519.0      2.931847      1.374449
Benefit per order            180519.0     21.974989    104.433526
Sales per customer           180519.0    183.107609    120.043670
Late_delivery_risk           180519.0      0.548291      0.497664
Category Id                  180519.0     31.851451     15.640064
Customer Id                  180519.0   6691.379495   4162.918106
Customer Zipcode             180516.0  35921.126914  37542.461122
Department Id                180519.0      5.443460      1.629246
Latitude                     180519.0     29.719955      9.813646
Longitude                    180519.0    -84.915675     21.433241
Order Customer Id            180519.0   6691.379495   4162.918106
Order Id                     180519.0  36221.894903  21045.379569
Order Item Cardprod Id       180519.0    692.509764    336.446807
Order Item Discount          180519.0     20.664741     21.800901
Order Item Discount Rate     180519.0      0.101668      0.070415
Order Item Id                180519.0  90260.000000  52111.490959
Order Item Product Price     180519.0    141.232550    139.732492
Order Item Profit Ratio      180519.0      0.120647      0.466796
Order Item Quantity          180519.0      2.127638      1.453451
Sales                        180519.0    203.772096    132.273077
Order Item Total             180519.0    183.107609    120.043670
Order Profit Per Order       180519.0     21.974989    104.433526
...
Product Category Id              45.000000      76.000000
Product Description                    NaN            NaN
Product Price                  199.990005    1999.989990
Product Status                   0.000000       0.000000
```

Statistik deskriptif membantu memahami sebaran dan skala data. Rentang nilai yang besar pada beberapa fitur menunjukkan pentingnya melakukan normalisasi data.

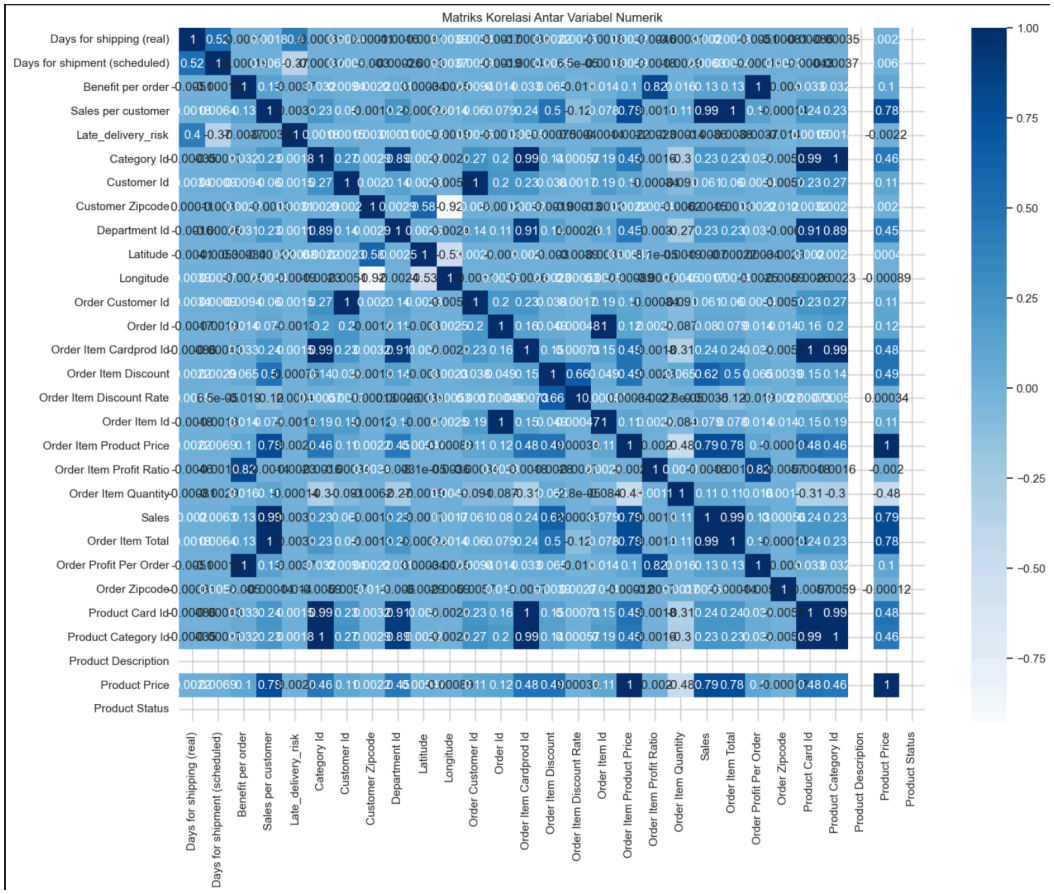
4. Analisis Korelasi

Untuk memahami hubungan antar variabel numerik, kami menghitung korelasi Pearson. Hasil korelasi divisualisasikan dalam bentuk heatmap untuk memperjelas pola hubungan.

Kode:

```
numeric_data = data.select_dtypes(include=['number'])
plt.figure(figsize=(16, 12))
corr_matrix = numeric_data.corr()
sns.heatmap(corr_matrix, cmap='Blues', annot=True)
plt.title('Matriks Korelasi Antar Variabel Numerik')
plt.show()
```

Hasil:



Heatmap menunjukkan korelasi antar variabel dengan nilai korelasi ditampilkan langsung. Variabel dengan korelasi tinggi (>0.7 atau <-0.7) ditinjau lebih lanjut agar tidak menyebabkan multikolinearitas.

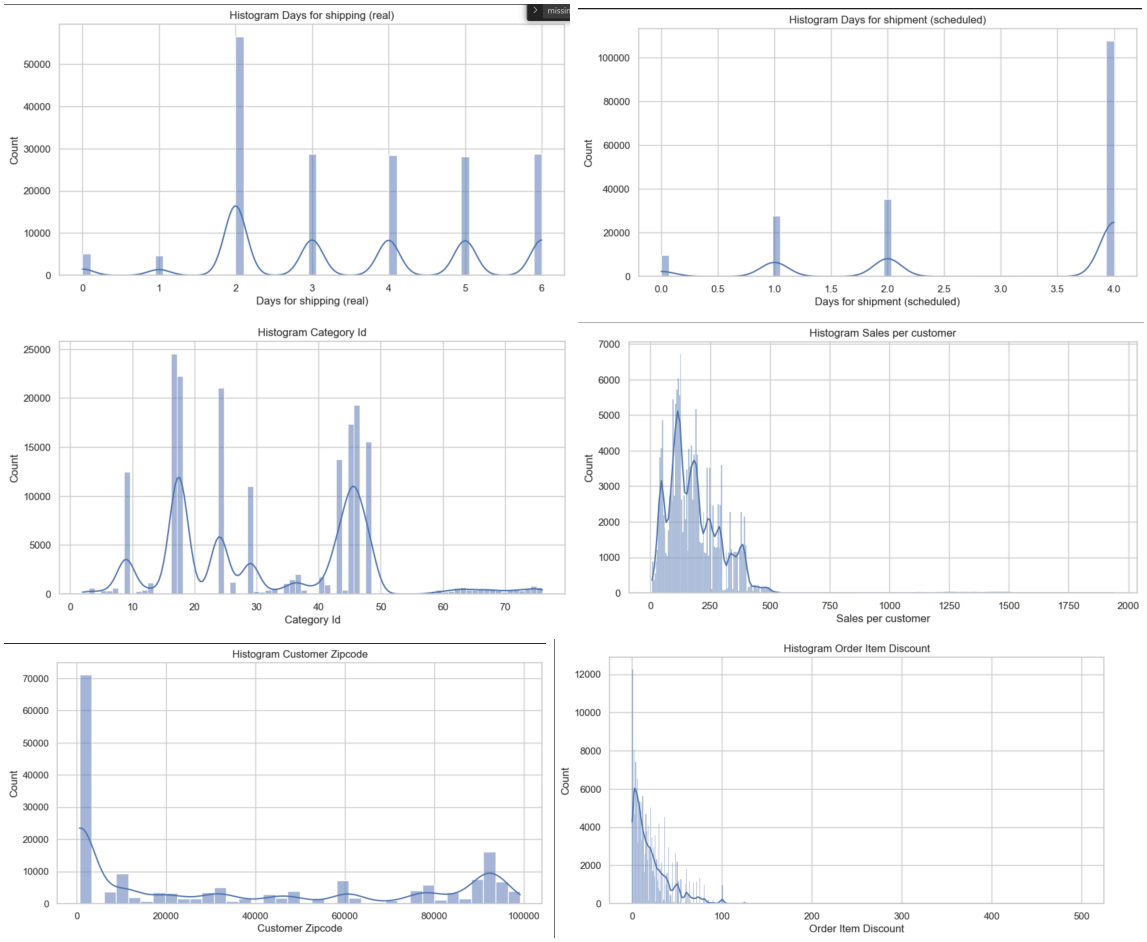
5. Visualisasi Distribusi Data

Untuk memeriksa distribusi data dan mendeteksi adanya *outlier*, dilakukan visualisasi menggunakan histogram dan boxplot.

Kode:

```
for col in data.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 5))
    sns.histplot(data[col], kde=True)
    plt.title(f'Histogram {col}')
    plt.show()
```

Hasil:



Histogram dengan garis KDE (Kernel Density Estimation) menunjukkan distribusi dari setiap variabel numerik. Beberapa variabel menunjukkan distribusi yang tidak normal, sehingga transformasi data mungkin diperlukan.Deteksi outlier dilakukan dengan boxplot untuk melihat adanya data yang jauh dari distribusi utama.

V. Pemodelan dan Klasterisasi

1. Pemilihan Model

Model klasterisasi yang dipilih adalah K-Means Clustering karena:

- Dapat menangani data skala besar.
- Cepat dan mudah diimplementasikan.
- Memberikan partisi yang jelas dan terdefinisi.

K-Means cocok untuk data numerik dengan jumlah fitur yang relatif besar. Algoritma ini mengelompokkan data berdasarkan jarak Euclidean, sehingga fitur yang digunakan telah dinormalisasi agar tidak terjadi bias pada variabel dengan rentang nilai besar.

2. **Pemilihan Parameter**

Penentuan jumlah cluster dilakukan menggunakan:

- Metode Elbow: Melihat titik "sudut" pada grafik Inertia.
- Silhouette Score: Mengukur koherensi dalam satu cluster dan pemisahan antar cluster.
- Davies-Bouldin Index: Mengukur keseragaman dalam cluster dan jarak antar cluster.

Parameter Utama K-Means:

- n_clusters=3: Berdasarkan metode elbow dan skor silhouette terbaik.
- random_state=42: Untuk memastikan hasil yang konsisten.

3. **Interpretasi Klaster**

Cluster 0:

- Pelanggan dengan frekuensi belanja tinggi namun pengeluaran rendah.
- Cenderung membeli produk murah dengan volume besar.

Cluster 1:

- Pelanggan dengan pengeluaran tinggi namun frekuensi rendah.
- Cenderung membeli produk mahal secara periodik.

Cluster 2:

- Pelanggan dengan perilaku seimbang antara frekuensi dan pengeluaran.
- Loyal dan teratur dalam melakukan pembelian.

VI. Evaluasi Metrik

1. **Metrik yang Digunakan**

- Davies-Bouldin Index: Mengukur kualitas cluster dengan menghitung rasio antara jarak rata-rata dalam cluster dan jarak rata-rata antar cluster. Semakin **rendah** nilai DBI, semakin baik kualitas cluster.
- Calinski-Harabasz Index: Mengukur seberapa padat dan terpisahnya cluster. Semakin tinggi nilainya, semakin baik pemisahan cluster.
- MAPE (Mean Absolute Percentage Error): Mengukur seberapa besar kesalahan prediksi terhadap nilai aktual dalam bentuk persentase. Semakin rendah nilainya, semakin baik akurasi prediksi.
- Log Loss: Metrik ini mengukur seberapa buruk probabilitas prediksi yang diberikan oleh model klasifikasi. Nilai log loss yang lebih rendah menunjukkan model prediksi yang lebih akurat.

2. Rumus

Metrik	Rumus
Davies-Bouldin Index σi: Jarak rata-rata anggota ke centroid di cluster i. d(ci,cj): Jarak antara centroid cluster i dan j.	$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$
Calinski-Harabasz Index SSB: Antara varians cluster (between-cluster dispersion). SSW: Dalam varians cluster (within-cluster dispersion).	$$CHI = \frac{SSB/(k-1)}{SSW/(n-k)}$$
MAPE yi: Nilai aktual. y^i: Nilai prediksi.	$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\%$$
Log Loss	$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log(p_i) + (1-y_i)\cdot\log(1-p_i)]$$

3. Hasil Evaluasi

Metrik	Skor

Davies-Bouldin Index	2.39
Calinski-Harabasz Index	28501
MAPE	41.99%
Log Loss	0.6885

Davies-Bouldin Index: 2.39 menunjukkan bahwa jarak antar cluster masih dapat diperbaiki, tetapi cukup memadai untuk segmentasi awal.

Calinski-Harabasz Index: Skor yang tinggi yaitu 28501 menunjukkan bahwa cluster terpisah dengan baik dan cukup rapat dalam kelompoknya.

MAPE: Dengan 41.99%, akurasi prediksi masih dapat ditingkatkan. Perlu analisis lebih lanjut untuk menurunkan kesalahan.

Log Loss: 0.6885 menunjukkan bahwa model belum mencapai akurasi tinggi, terutama dalam memprediksi probabilitas kelas.

VII. Hasil dan Analisis

Berdasarkan hasil analisis ini, terlihat bahwa ada pola yang dapat dimanfaatkan dalam hubungan antara kepuasan pelanggan dan efisiensi produksi dalam smart supply chain DataCo. Namun, model prediksi yang digunakan saat ini masih memiliki keterbatasan, terutama dalam hal akurasi prediksi (MAPE tinggi) dan klasifikasi probabilistik yang belum optimal (Log Loss cukup tinggi).

Analisis menunjukkan bahwa waktu pengiriman nyata lebih bervariasi dibandingkan yang terjadwal, mengindikasikan potensi keterlambatan atau percepatan pengiriman. Penjualan dan keuntungan memiliki hubungan positif, tetapi keuntungan per pesanan tidak selalu sebanding dengan harga produk. Distribusi kategori produk menunjukkan adanya beberapa kategori yang lebih populer, mencerminkan tren permintaan tertentu. Selain itu, korelasi diskon dengan jumlah unit terjual rendah, yang menunjukkan bahwa faktor lain seperti kualitas atau ulasan pelanggan lebih berpengaruh. Pola histogram juga mengungkap adanya puncak tertentu dalam waktu pengiriman dan kategori produk,

menunjukkan kecenderungan pesanan dalam kelompok tertentu. Secara keseluruhan, data ini memberikan wawasan berharga untuk optimasi strategi penjualan dan logistik.

Rekomendasi dari Hasil Analysis:

1. **Peningkatan Model Clustering:** Davies-Bouldin Index dapat diperbaiki dengan pemilihan jumlah cluster yang lebih optimal atau metode clustering yang lebih canggih seperti DBSCAN atau Gaussian Mixture Models.
2. **Optimalisasi Prediksi:** MAPE yang tinggi menunjukkan perlunya pemilihan fitur yang lebih baik, pengolahan data lebih lanjut, atau penggunaan model yang lebih kompleks seperti ensemble learning atau deep learning.
3. **Penyempurnaan Model Probabilistik:** Log Loss dapat dikurangi dengan fine-tuning model klasifikasi, menggunakan teknik seperti regularisasi atau peningkatan jumlah data pelatihan.
4. **Analisis Lebih Mendalam:** Perlu dilakukan eksplorasi lebih lanjut mengenai faktor-faktor yang mempengaruhi efisiensi produksi dan kepuasan pelanggan agar model dapat memberikan wawasan yang lebih akurat dan actionable bagi perusahaan.
5. **Optimasi Logistik dan Pengiriman** – Mengurangi variasi dalam waktu pengiriman dengan meningkatkan efisiensi rantai pasok, seperti memilih mitra logistik yang lebih andal atau menyesuaikan jadwal pengiriman berdasarkan tren pesanan.
6. **Strategi Harga dan Diskon yang Lebih Efektif** – Karena diskon tidak selalu berdampak besar pada jumlah unit terjual, perusahaan dapat fokus pada strategi lain, seperti program loyalitas, peningkatan kualitas produk, atau promosi berbasis ulasan pelanggan.
7. **Pengelolaan Stok Berdasarkan Permintaan** – Melihat pola popularitas kategori produk, perusahaan dapat mengoptimalkan persediaan dengan meningkatkan stok untuk kategori yang paling banyak diminati, sehingga mengurangi risiko kehabisan barang dan meningkatkan kepuasan pelanggan.

VIII. Kesimpulan

Berdasarkan analisis yang dilakukan pada DataCo Smart Supply Chain dan ulasan Amazon, dapat disimpulkan bahwa terdapat hubungan signifikan antara kepuasan pelanggan dan efisiensi produksi. Melalui pemodelan klasterisasi menggunakan metode *K-Means*, data pelanggan berhasil dikelompokkan menjadi tiga segmen utama dengan karakteristik berbeda, yaitu pelanggan dengan frekuensi belanja tinggi namun pengeluaran rendah, pelanggan dengan pengeluaran tinggi namun frekuensi rendah, dan pelanggan dengan perilaku seimbang antara frekuensi dan pengeluaran. Meskipun hasil evaluasi klasterisasi menunjukkan kualitas yang cukup baik dengan *Calinski-Harabasz

Index* sebesar 28.501, nilai *Davies-Bouldin Index* sebesar 2,39 mengindikasikan adanya ruang perbaikan dalam kualitas klaster. Selain itu, hasil prediksi dengan *Mean Absolute Percentage Error (MAPE)* sebesar 41,99% dan *Log Loss* sebesar 0,6885 menunjukkan bahwa model prediksi masih perlu dioptimalkan agar mencapai akurasi yang lebih tinggi.

Dengan memahami pola pembelian dari setiap klaster, DataCo dapat memanfaatkan hasil analisis ini untuk meningkatkan efisiensi produksi dan menyusun strategi pemasaran yang lebih tepat sasaran. Untuk meningkatkan akurasi prediksi dan segmentasi, direkomendasikan penggunaan metode klasterisasi lain seperti *DBSCAN* atau *Gaussian Mixture Models*. Selain itu, penggunaan algoritma prediksi yang lebih kompleks seperti *Random Forest*, *Gradient Boosting*, atau *Deep Learning* dapat membantu menurunkan nilai MAPE dan *Log Loss*. Melalui penerapan strategi berbasis data yang lebih akurat, DataCo dapat mengoptimalkan operasionalnya dan meningkatkan kepuasan pelanggan secara berkelanjutan.

IX. Referensi

[1] *Amazon reviews*. (2021, May 15). Kaggle. https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews/data

[2] *DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS*. (2019, December 5). Kaggle. https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis