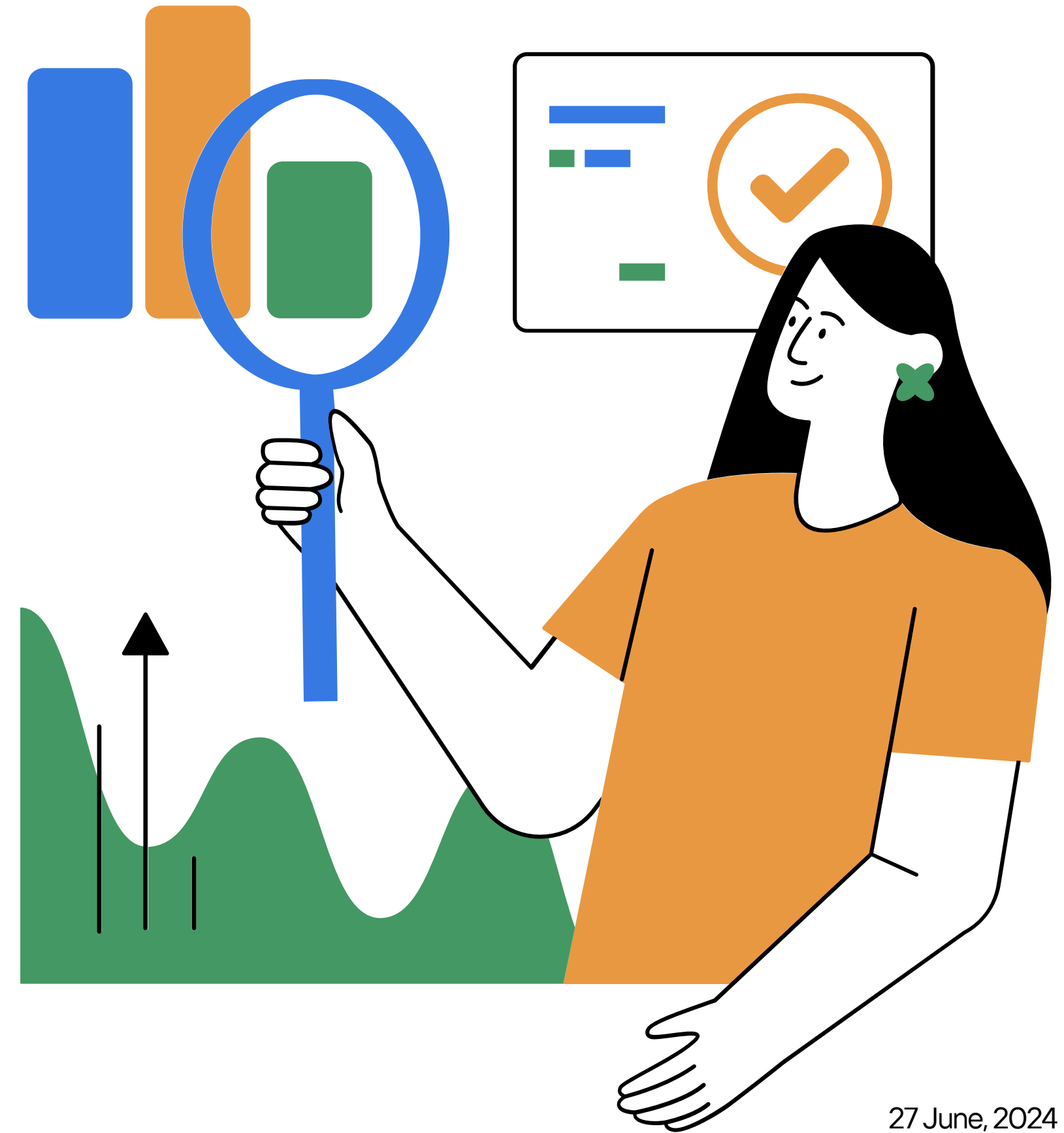


Data Analytics

GROUP 2





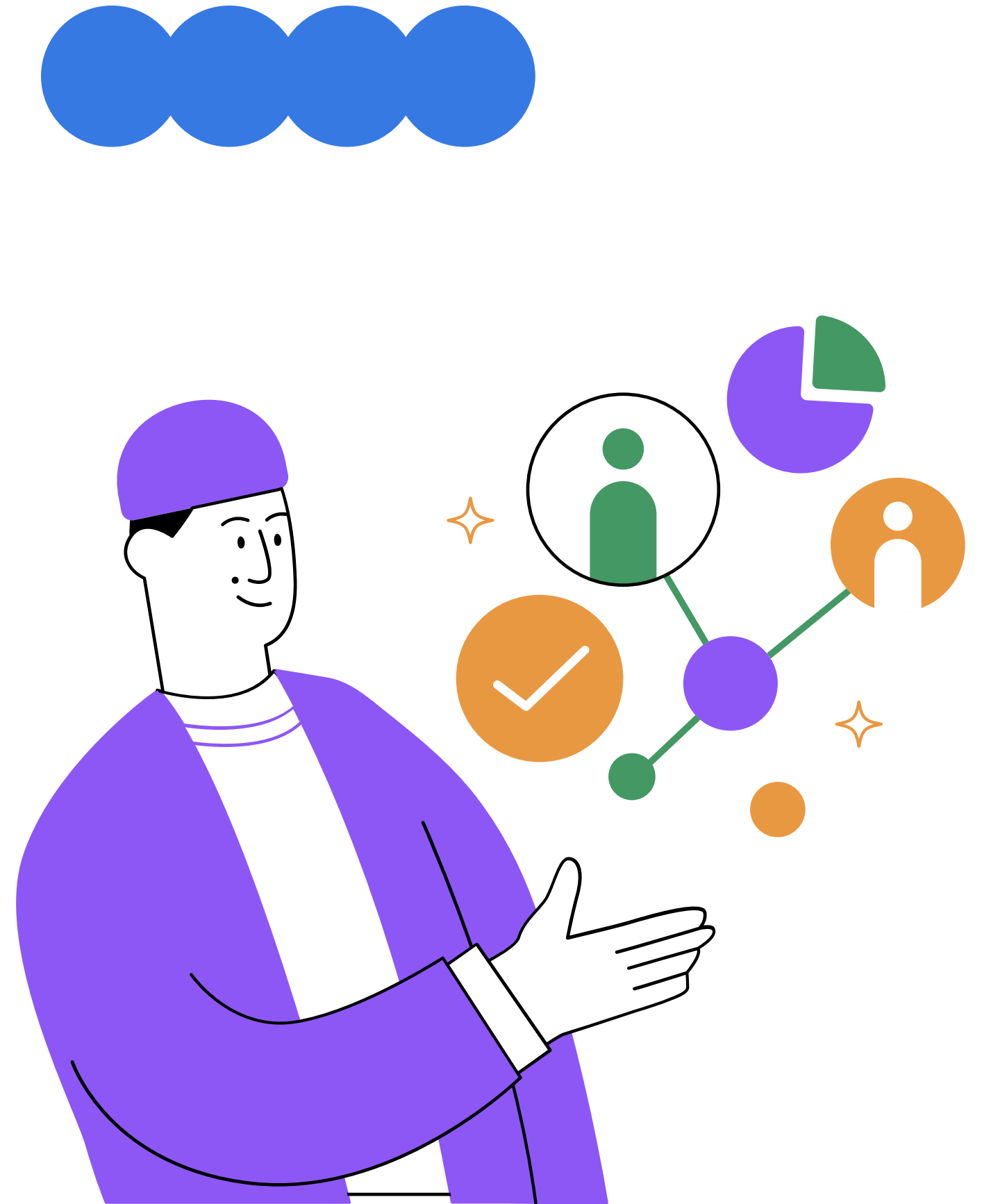
Latar Belakang

Industri Food & Beverage (F&B) memiliki persaingan yang ketat, di mana data transaksi menjadi faktor penting dalam memahami performa bisnis. "Pizza Place" mencatat lebih dari 48.000 transaksi pada tahun 2015 yang dapat digunakan untuk menganalisis pola penjualan seperti produk terlaris dan jam sibuk. Namun, saat ini pemanfaatan data masih terbatas pada analisis deskriptif, sehingga belum sepenuhnya mendukung pengambilan keputusan yang lebih mendalam dan berbasis prediksi.

Literature Review

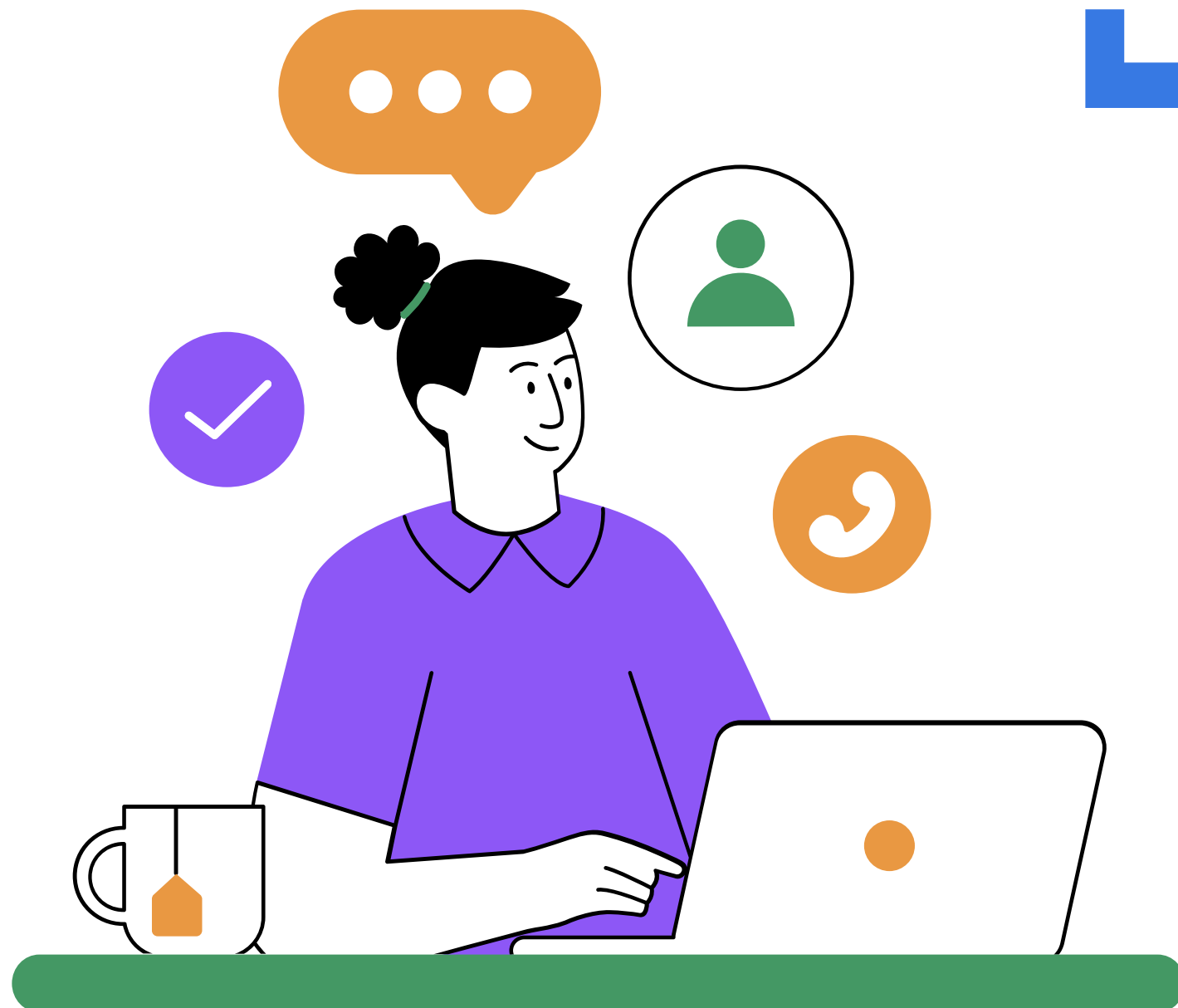
2.1 Kualitas Data terhadap Pemodelan

Kualitas data merupakan faktor utama dalam keberhasilan model Machine Learning. Dimensi seperti akurasi, kelengkapan, konsistensi, dan ketepatan waktu sangat memengaruhi hasil pemodelan. Data mentah sering mengandung anomali dan inkonsistensi sehingga perlu dilakukan preprocessing atau data wrangling. Data yang berkualitas baik akan meningkatkan akurasi, ketahanan model, dan kemampuan generalisasi.





Literature Review



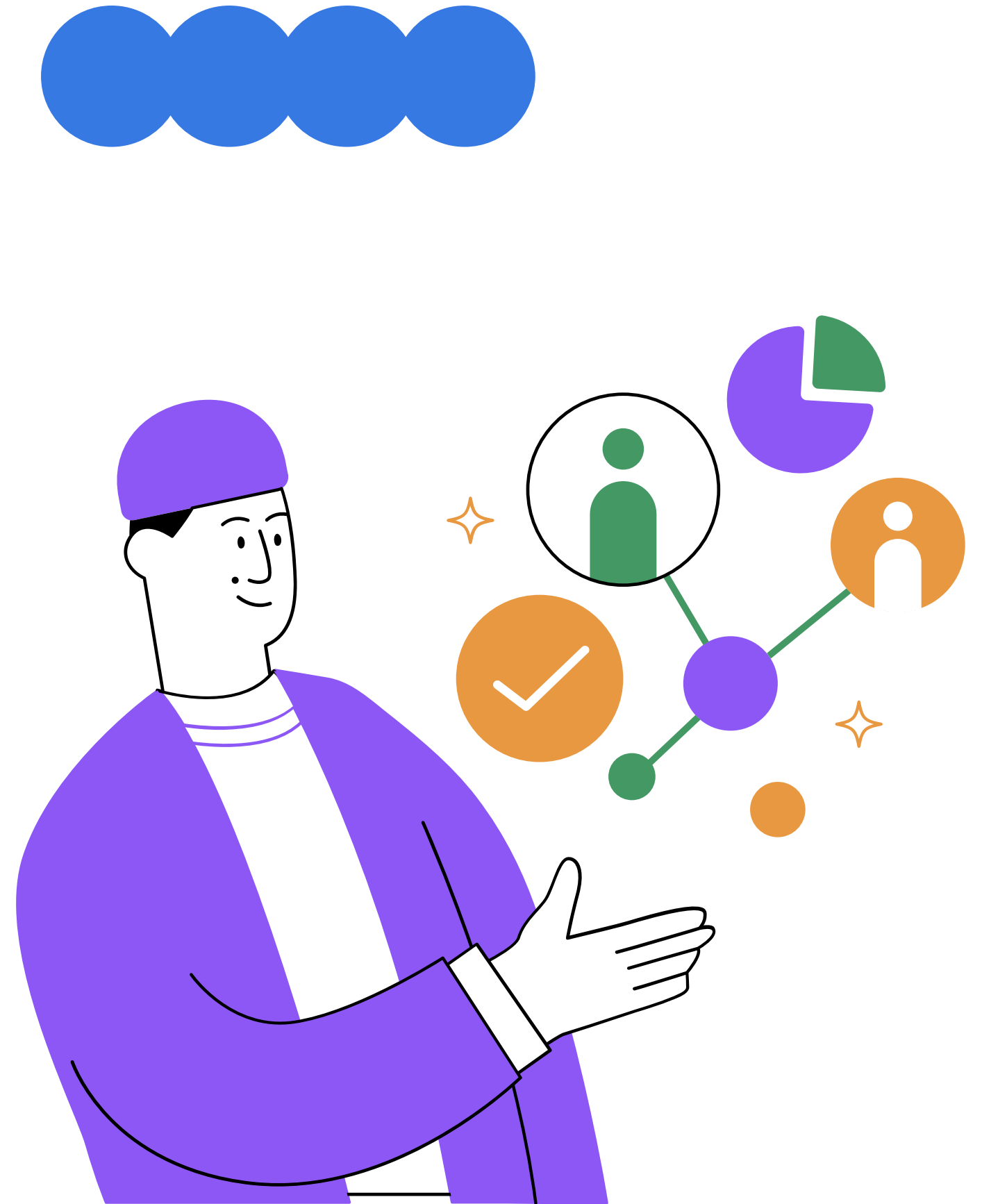
2.2 Pemodelan Prediktif

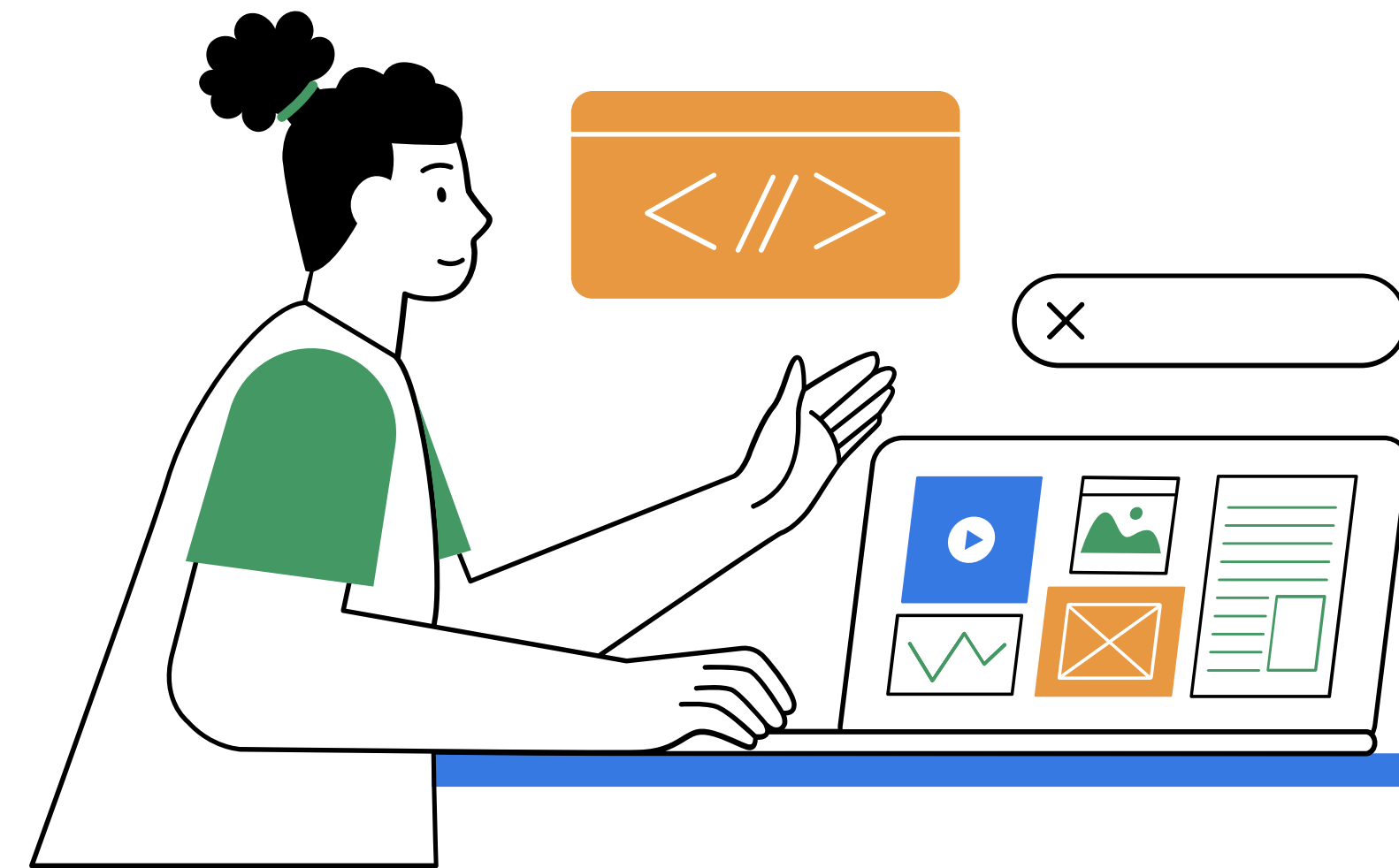
Dalam pemodelan prediktif, digunakan berbagai algoritma sesuai kompleksitas data. Linear Regression sering digunakan sebagai model dasar karena sederhana dan mudah diinterpretasikan. Untuk pendekatan yang lebih kompleks, Random Forest Regressor digunakan sebagai metode ensemble learning yang menggabungkan banyak decision tree untuk meningkatkan akurasi dan mengurangi overfitting.

Literature Review

2.3 Metrik Evaluasi Model dan Keterbatasan Interpretasi

Evaluasi model umumnya menggunakan metrik seperti MAE dan RMSE untuk mengukur selisih antara nilai prediksi dan aktual. Namun, akurasi statistik yang tinggi tidak selalu mencerminkan keberhasilan bisnis. Hal ini karena metrik tersebut tidak mempertimbangkan dampak asimetris dari kesalahan prediksi. Oleh karena itu, evaluasi model perlu dipahami bersama keterbatasannya dalam konteks pengambilan keputusan bisnis.

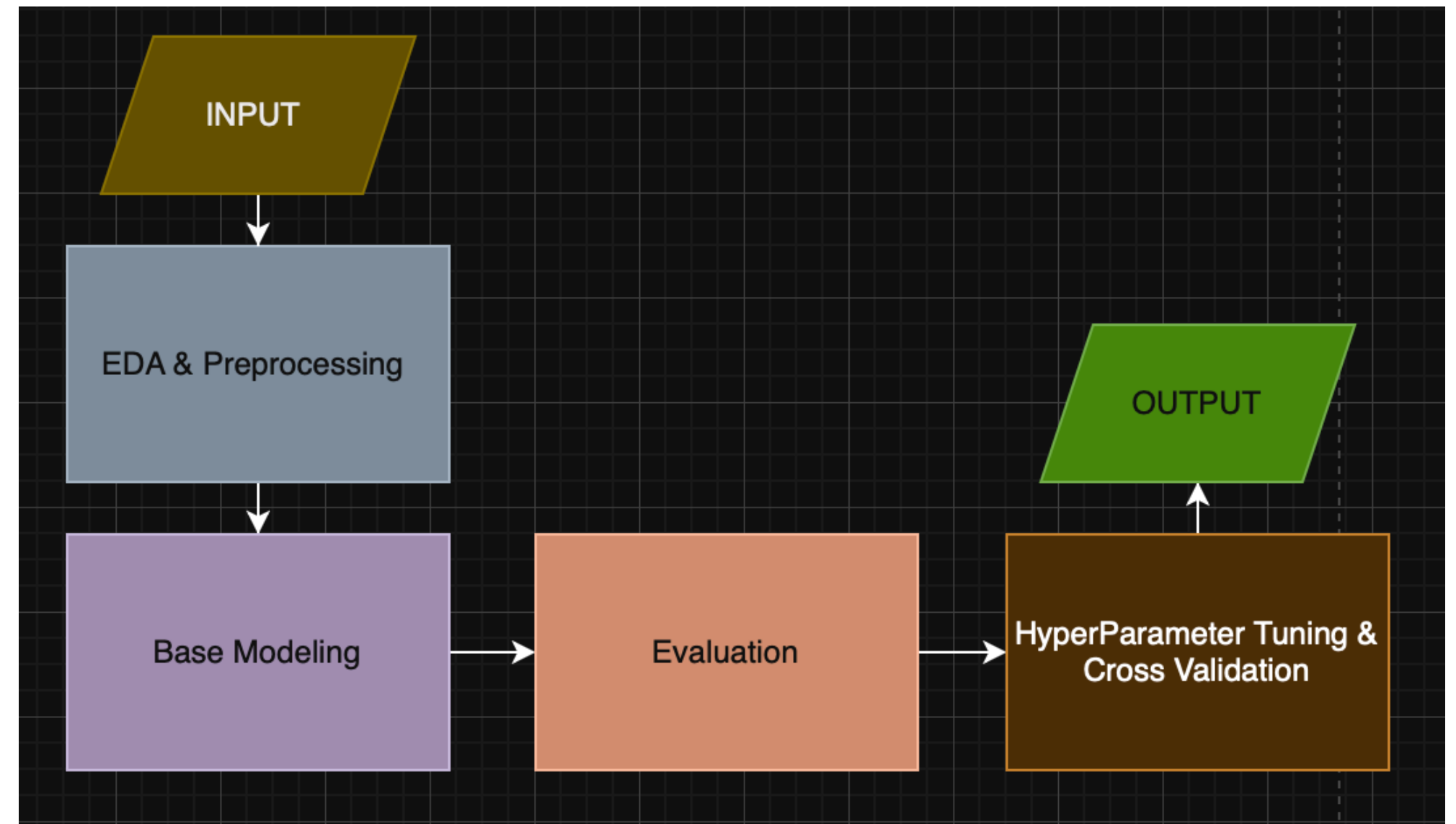




Methodologi

3.1. Analytical Flow

Metodologi dalam proyek ini dirancang menggunakan pendekatan sistematis yang terdiri dari enam tahapan utama.



3.2. Penjelasan Tahapan Metodologi

1. Input Data

Tahap awal mencakup pemuatan dataset mentah transaksi "Pizza Place" tahun 2015. Data ini diinspeksi untuk mengidentifikasi anomali struktural bawaan yang dapat menghambat proses agregasi dan pemodelan, seperti kesalahan format desimal pada harga dan tipe data temporal.

2. EDA & Preprocessing

Fase ini terbagi menjadi dua proses krusial:

- Data Wrangling & Cleaning: Memperbaiki format desimal harga dari koma menjadi titik untuk memungkinkan kalkulasi matematis (float), mengonversi atribut waktu ke dalam format datetime, dan menghilangkan data duplikat.
- Feature Engineering & Encoding: Mengekstraksi variabel turunan seperti Jam (Hour), Hari (Day_Name), dan Bulan (Month). Atribut kategorikal (Kategori dan Ukuran Pizza) ditransformasikan menjadi representasi numerik menggunakan metode Label Encoding agar dapat diproses oleh algoritma matriks korelasi dan Machine Learning.

3. Base Modeling

Untuk mengidentifikasi hubungan prediktif antara spesifikasi pesanan dan volume permintaan, kami mendefinisikan quantity (jumlah pizza) sebagai target prediksi (dependent variable). Variabel unit_price dan total_price secara ketat dikeluarkan dari fitur input untuk mencegah Data Leakage, mengingat keduanya adalah turunan matematis dari target. Dua teknik analitik yang digunakan adalah:

- Linear Regression: Digunakan sebagai model baseline (dasar) untuk menguji apakah terdapat hubungan linier sederhana antara waktu/kategori dengan jumlah pesanan.
- Random Forest Regressor: Digunakan sebagai model lanjutan berbasis decision trees ensemble untuk menangkap potensi pola non-linier yang kompleks pada perilaku pesanan pelanggan.

4. Evaluasi Model (Leandra Belva VPN / 2802499876)

Kinerja kedua model dasar diukur secara kuantitatif menggunakan metrik evaluasi regresi berikut:

- Mean Absolute Error (MAE) & Mean Squared Error (MSE): Digunakan untuk mengukur rata-rata deviasi absolut dan kuadrat antara prediksi volume pesanan dengan angka aktual. MSE memberikan penalti lebih besar pada kesalahan prediksi yang ekstrem.
- R-Squared (R2): Dipilih sebagai metrik rasio krusial untuk mengukur seberapa besar persentase variansi dari jumlah pesanan yang benar-benar dapat dijelaskan oleh variabel independen (waktu dan ukuran/kategori pizza).

5. HyperParameter Tuning & Cross Validation (Leandra Belva VPN / 2802499876)

Model Random Forest dieksekusi ulang menggunakan GridSearchCV yang dipadukan dengan 3-fold Cross Validation. Tahap ini tidak bertujuan untuk sekadar menaikkan akurasi secara buta, melainkan untuk membatasi kompleksitas model (seperti membatasi max_depth menjadi 3) guna mencegah overfitting. Cross-validation memastikan metrik yang dihasilkan stabil dan tidak bias terhadap satu partisi data pelatihan (training data) tertentu.

6. Output

Tahap akhir berupa ekstraksi insight dari perbandingan performa model sebelum dan sesudah tuning. Hasil evaluasi ini kemudian ditranslasikan menjadi identifikasi limitasi data historis dan rekomendasi bisnis strategis bagi manajemen.

Methodologi

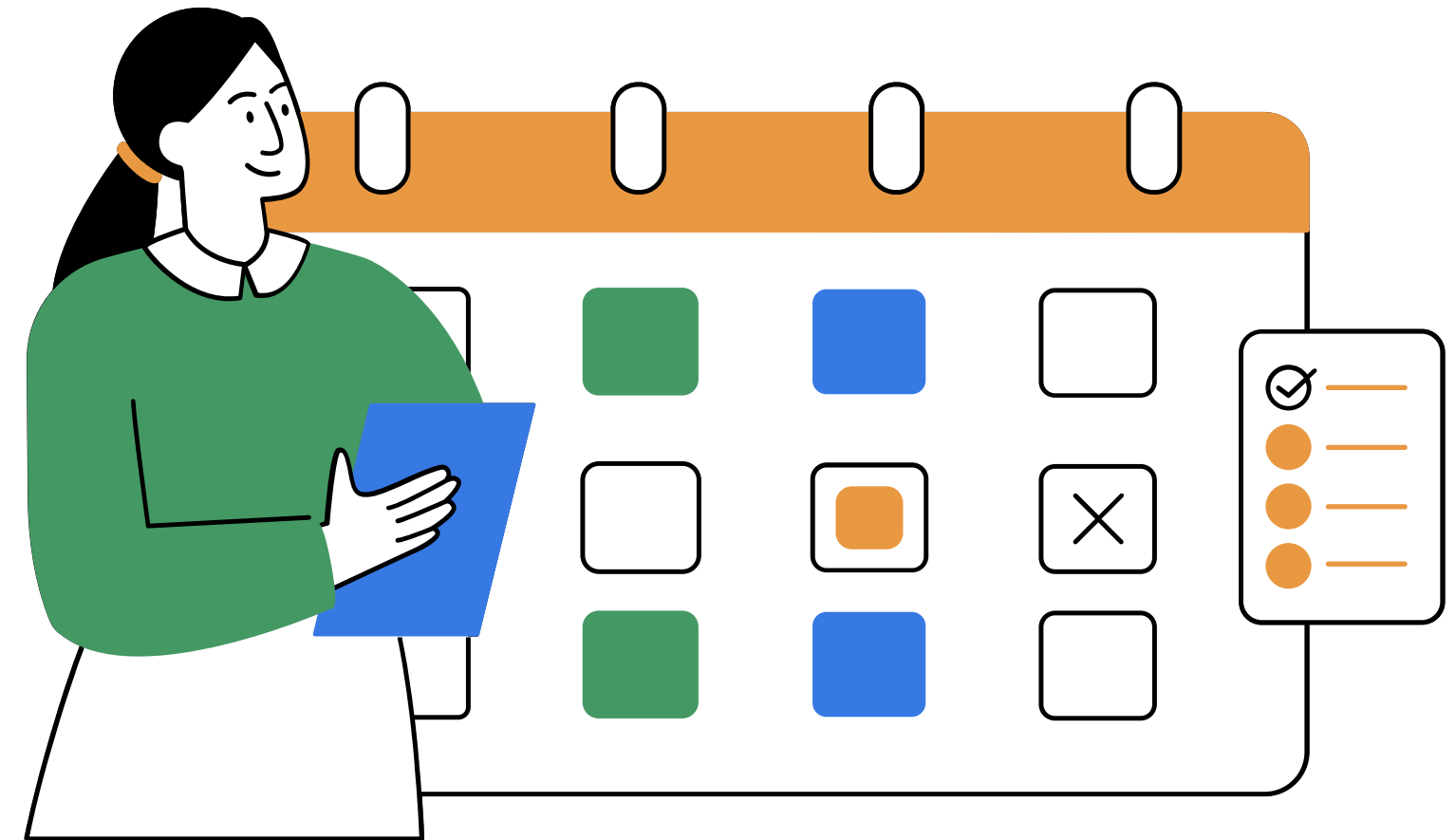


Nathasya Umboh / 2802533043



Leandra Belva Varissa PN / 2802499876

Implementasi & Evaluasi



4.1 Review dan Evaluasi Kualitas Data

a. Identifikasi Masalah Utama

Masalah kualitas data primer yang diidentifikasi pada fase awal (UTS) bukan terletak pada missing values atau duplikasi (keduanya terkonfirmasi nol mutlak), melainkan pada kesalahan format struktural dan tipe data numerik. Kolom harga (unit_price dan total_price) terekam dengan tanda koma sebagai pemisah desimal, menyebabkan sistem membacanya sebagai teks (string). Kolom waktu (order_date dan order_time) juga direkam dalam format teks murni.



b. Efektivitas Proses Data Wrangling

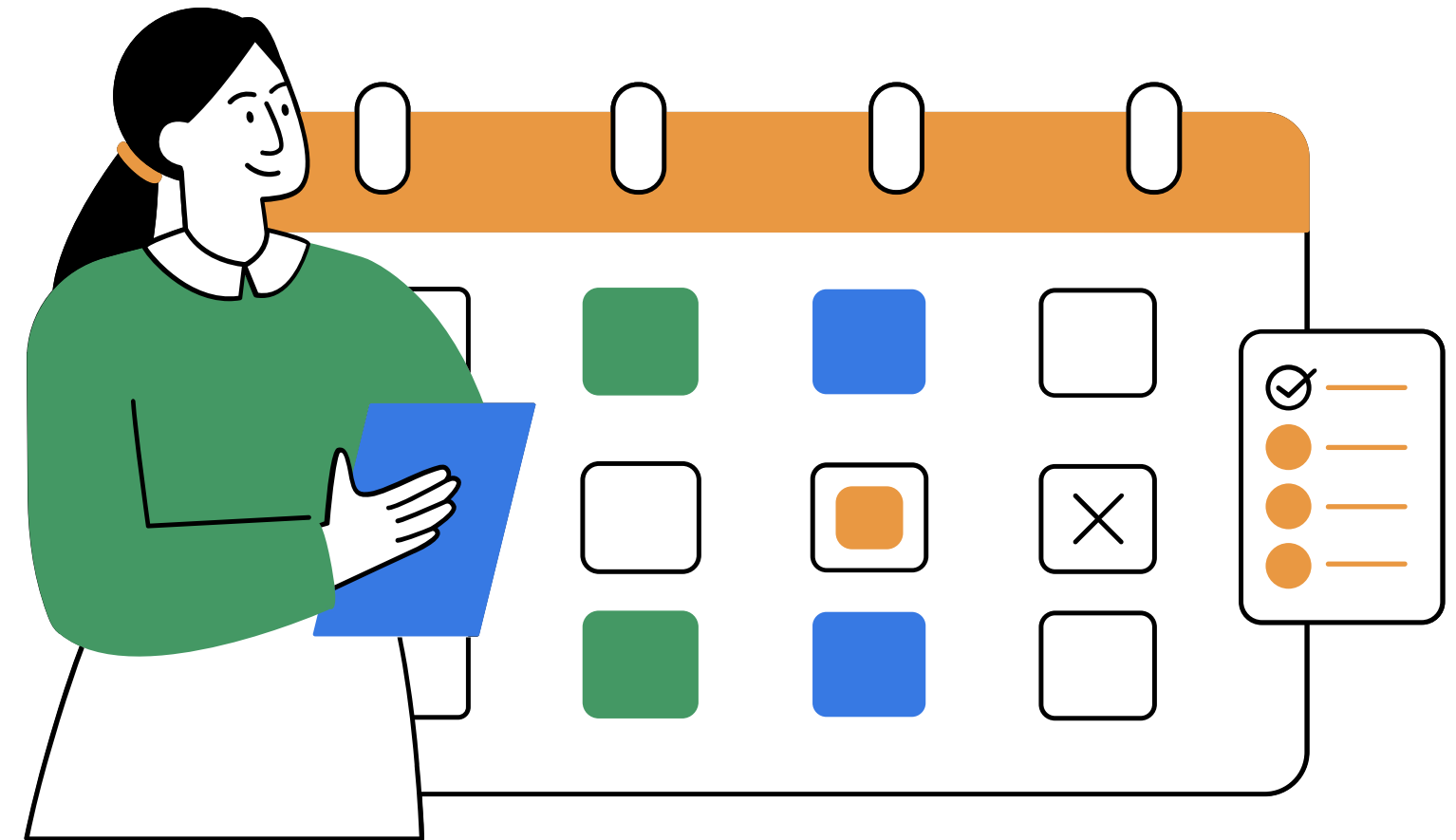
Proses pembersihan data difokuskan pada manipulasi format. Kami mengimplementasikan teknik coercion untuk menyubstitusi karakter koma menjadi titik, dilanjutkan dengan konversi paksa (type-casting) ke format desimal (float64) dan waktu standar (datetime64[ns]).



c. Indikator Keberhasilan

Indikator mutlak keberhasilan wrangling ini adalah transisi tipe data dari object/string menjadi tipe numerik dan temporal. Tanpa perbaikan fundamental ini, algoritma matriks korelasi dan Random Forest akan mengalami syntax error akibat kegagalan sistem komputasi dalam mengalikan variabel teks.

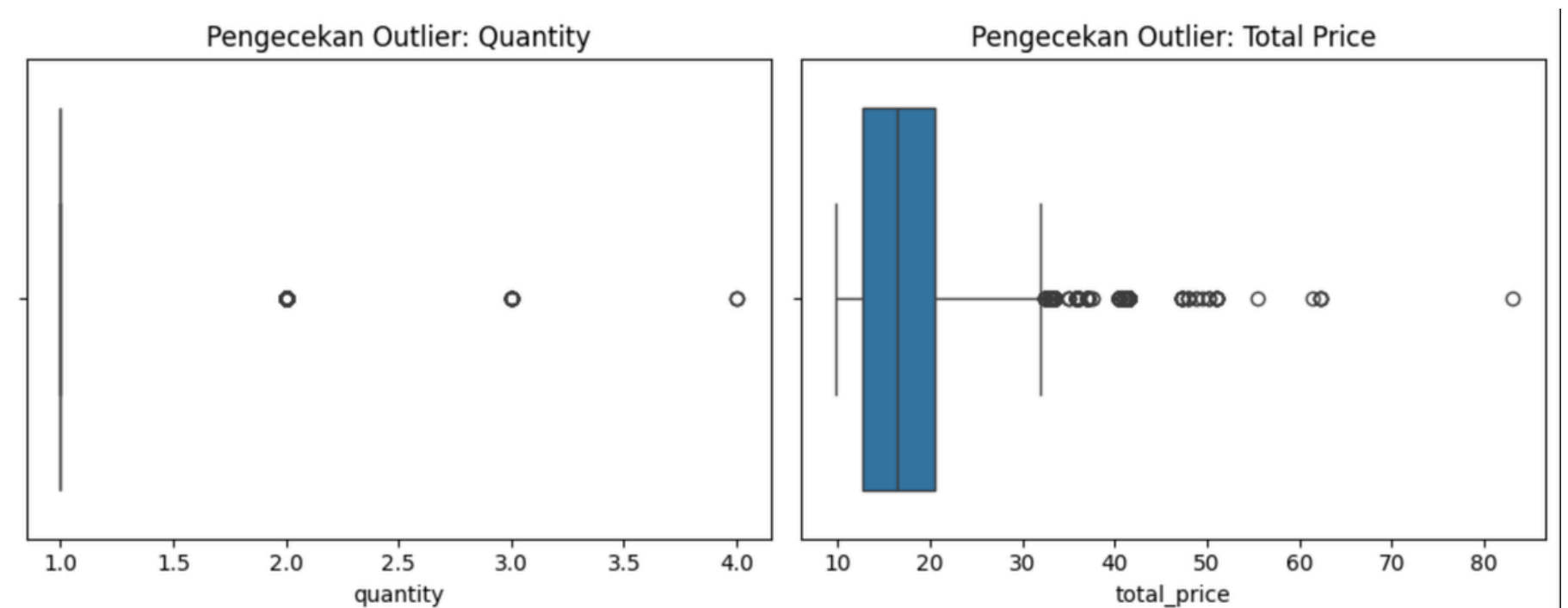
Implementasi & Evaluasi



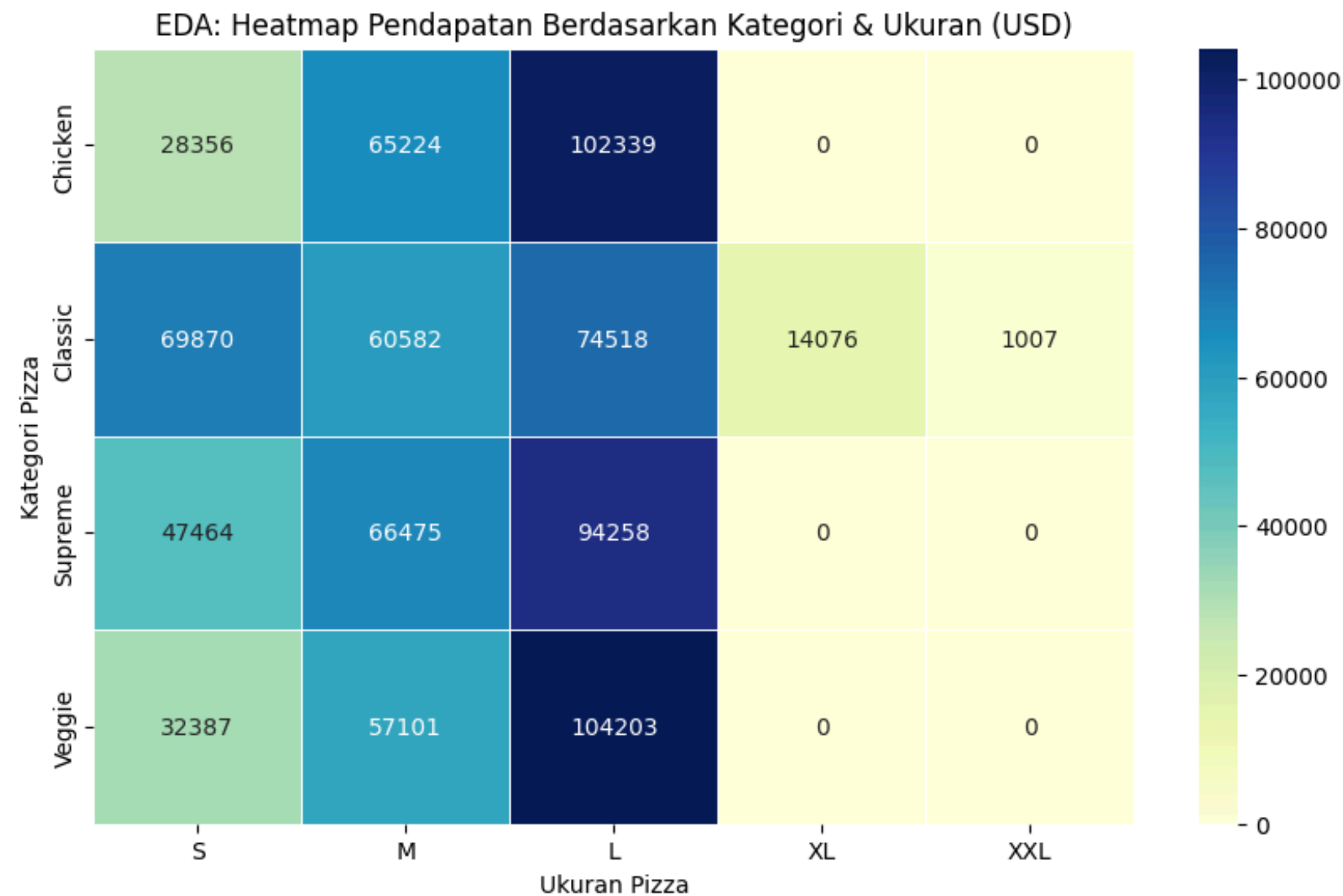
4.1 Review dan Evaluasi Kualitas Data

d. Analisis Outliers

Visualisasi boxplot menunjukkan adanya outliers pada variabel quantity dan total_price. Namun, outliers tersebut tidak dihapus karena dianggap merepresentasikan transaksi yang valid, seperti pesanan dalam jumlah besar untuk acara atau kebutuhan korporat. Menghapus data tersebut berisiko mengurangi nilai pendapatan yang sebenarnya dan menghasilkan analisis bisnis yang kurang akurat.

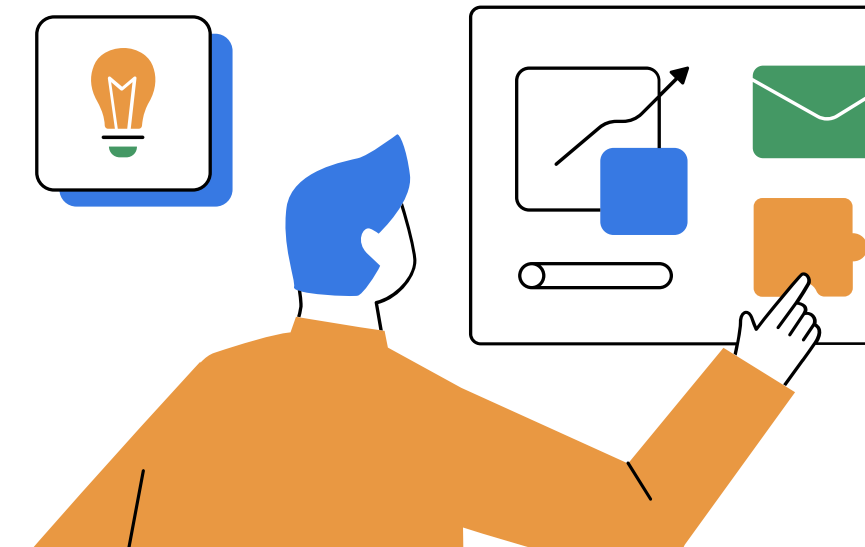


4.2 EDA & Variabel Paling Berpengaruh

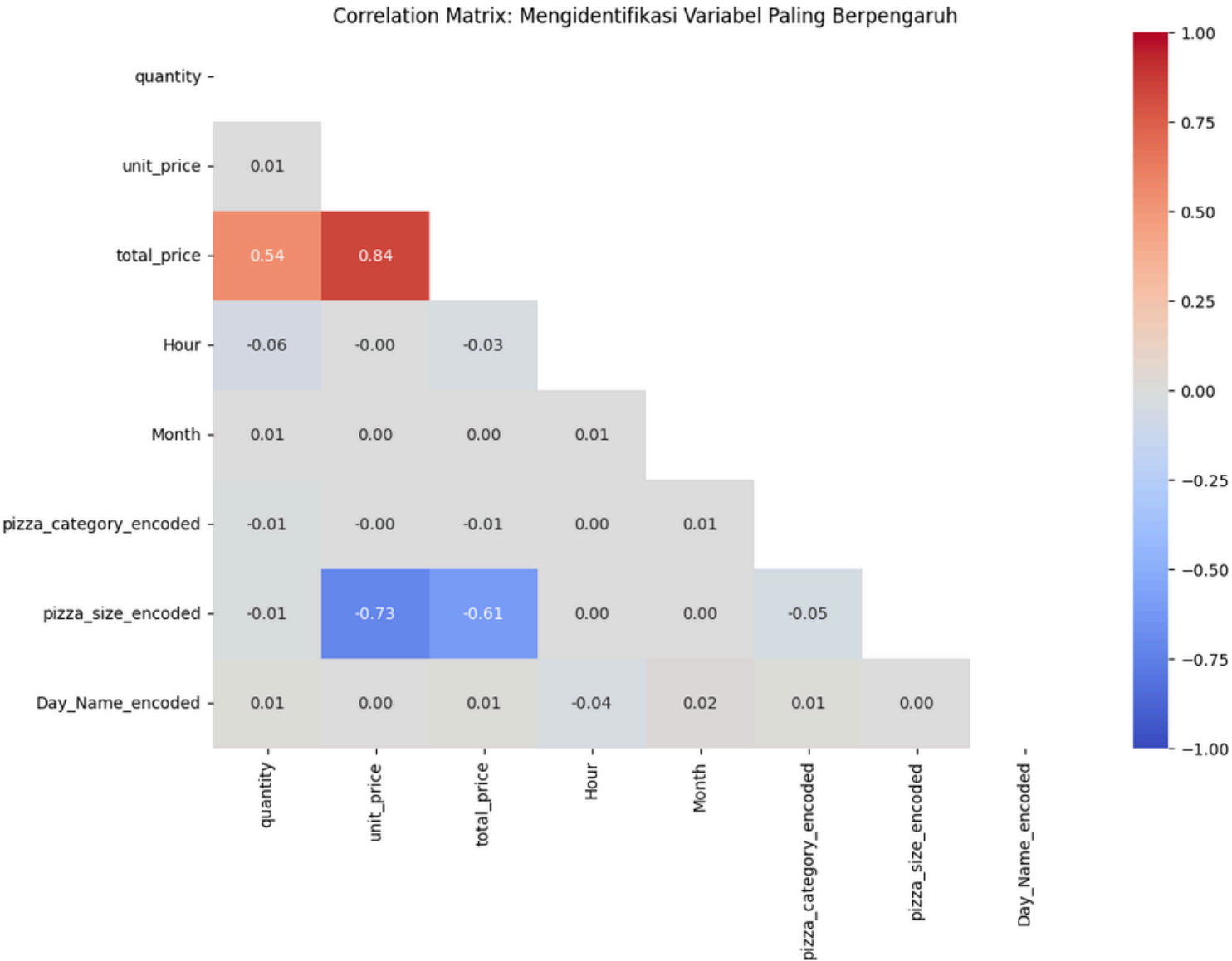


a. Analisis Tersegmentasi

Analisis heatmap menunjukkan bahwa ukuran Large (L) menjadi kontributor pendapatan terbesar di hampir seluruh kategori pizza. Sebaliknya, ukuran XL dan XXL memberikan kontribusi pendapatan yang sangat rendah meskipun tetap tersedia pada kategori tertentu. Temuan ini mengindikasikan bahwa pelanggan cenderung memilih ukuran L sebagai ukuran yang paling optimal, sehingga ukuran ekstrem memiliki nilai bisnis yang relatif rendah.

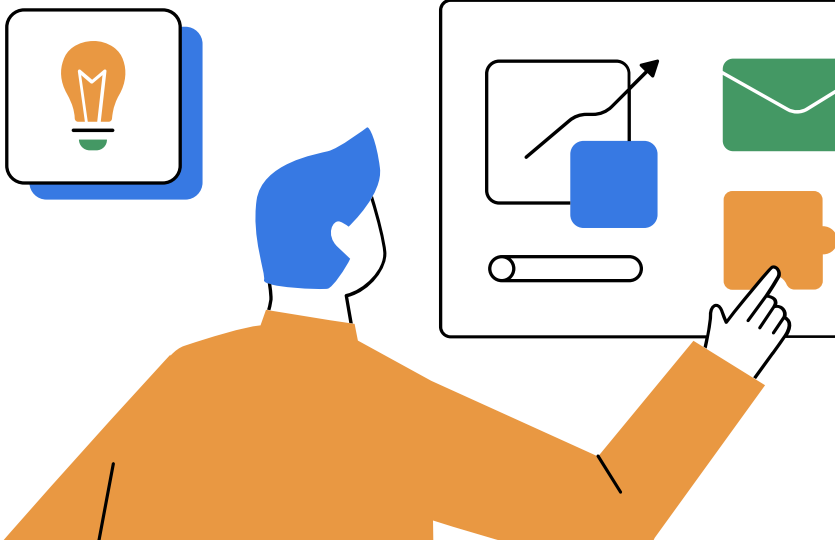


4.2 EDA & Variabel Paling Berpengaruh



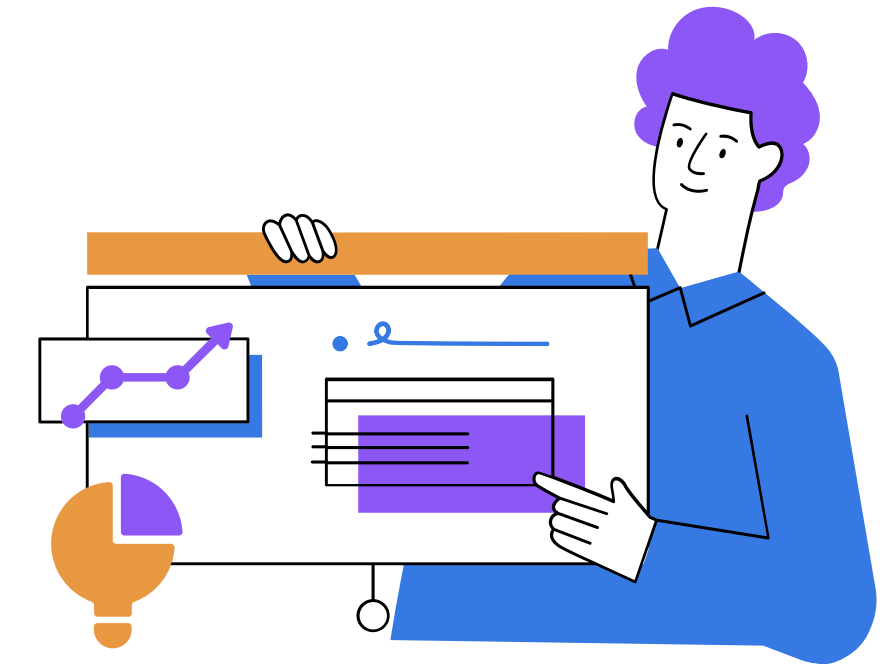
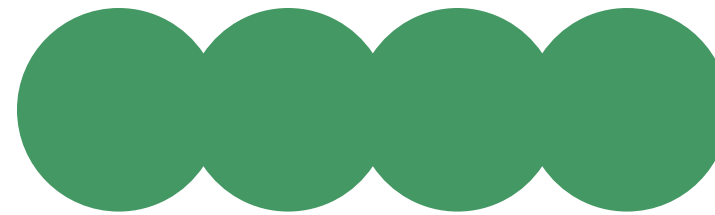
b. Variabel Paling Berpengaruh

Hasil analisis korelasi menunjukkan bahwa variabel operasional seperti jam, bulan, hari, kategori, dan ukuran pizza memiliki hubungan yang sangat lemah terhadap jumlah pesanan (quantity). Hal ini mengindikasikan bahwa fluktuasi permintaan pelanggan cenderung sulit diprediksi menggunakan variabel-variabel tersebut. Temuan ini juga menjelaskan mengapa model regresi diperkirakan akan mengalami kesulitan dalam membentuk pola prediksi yang akurat.



✿ Leandra Belva Varissa PN / 2802449876

✿ Nathasya Umboh / 2802533043



4.3. Implementasi Pemodelan dan Evaluasi

Evaluasi performa berbasis metrik menunjukkan dinamika berikut:

- Base Model: Linear Regression gagal menangkap pola ($R^2 = 0.0033$). Random Forest dasar yang dibiarkan tanpa batasan parameter terjebak dalam overfitting ekstrem akibat menghafal kebisingan data, menyebabkan akurasi hancur dengan R^2 negatif (-0.2314).
- Tuned Model: Intervensi Hyperparameter Tuning dengan Cross-Validation sukses memangkas kompleksitas pohon ($\text{max_depth} = 3$, $\text{min_samples_split} = 5$). Intervensi ini menghentikan overfitting, memulihkan nilai R^2 menjadi positif (0.0048), dan menekan angka error melampaui kemampuan model baseline.

--- EVALUASI BASE MODEL ---

	Model	MAE	MSE	R-Squared
1	Linear Regression	0.038707	0.021536	0.003373
2	Random Forest	0.041332	0.026611	-0.231471

--- MEMULAI HYPERPARAMETER TUNING ---

Fitting 3 folds for each of 12 candidates, totalling 36 fits

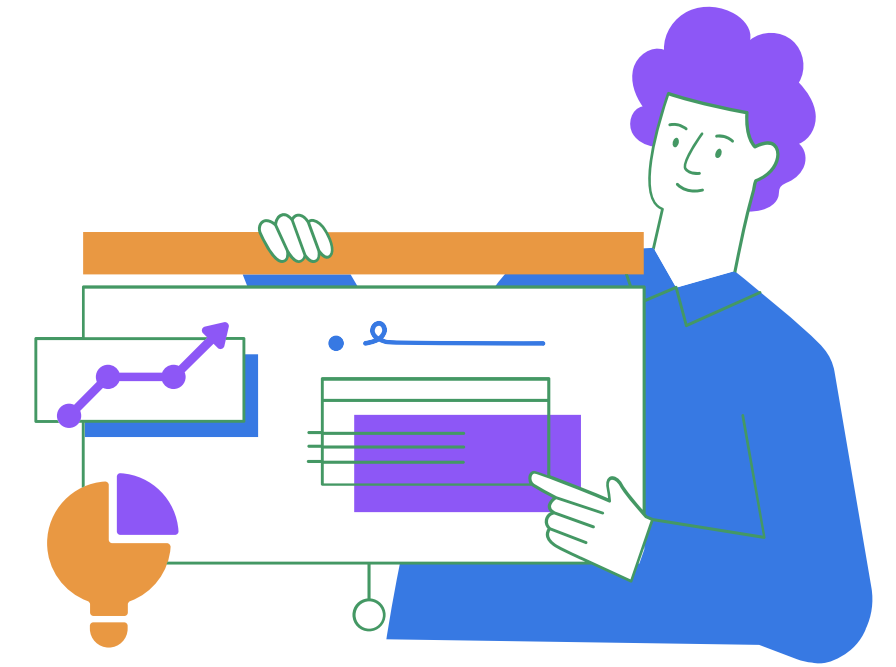
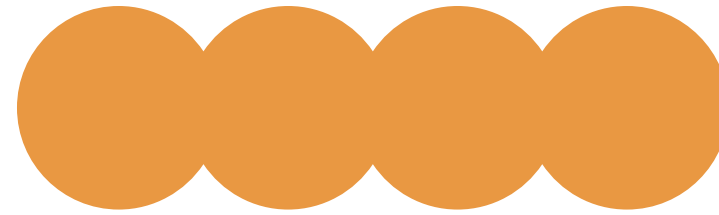
HASIL EVALUASI SETELAH TUNING (SESUDAH)

Best Parameters: {'max_depth': 3, 'min_samples_split': 5, 'n_estimators': 100}

MAE Tuned: 0.038677

MSE Tuned: 0.021504

R-Squared Tuned: 0.004856



4.4. Limitasi model

Meskipun proses tuning secara teknis telah berhasil menekan error dan mencegah overfitting, nilai R^2 akhir sebesar 0,0048 merupakan bukti matematis dari limitasi fundamental sistem prediksi ini. Artinya, model prediktif hanya sanggup menjelaskan tidak lebih dari 0,48% dari total variansi jumlah pesanan harian.

Akar masalahnya ada pada keterbatasan variabel dalam dataset itu sendiri. Volume transaksi restoran bergantung pada pemicu eksternal yang tidak terekam di sistem Point of Sales "Pizza Place": program promosi, ukuran rombongan tamu, cuaca harian, hari libur lokal. Tanpa variabel-variabel itu, algoritma Machine Learning tidak punya bahan untuk mendeteksi lonjakan transaksi dan bukan karena modelnya lemah, tapi karena informasinya memang tidak ada. Model ini tidak layak dipakai sebagai alat otomasi inventaris atau prakiraan rantai pasok.

```
--- MEMULAI HYPERPARAMETER TUNING ---  
Fitting 3 folds for each of 12 candidates, totalling 36 fits  
  
HASIL EVALUASI SETELAH TUNING (SESUDAH)  
Best Parameters: {'max_depth': 3, 'min_samples_split': 5, 'n_estimators': 100}  
MAE Tuned: 0.038677  
MSE Tuned: 0.021504  
R-Squared Tuned: 0.004856
```

Hasil & Diskusi

5.1 Kesimpulan

1. Model prediktif tidak mampu menghasilkan prediksi yang akurat, ditunjukkan oleh nilai R^2 sebesar 0,0048. Hal ini menunjukkan bahwa variabel operasional seperti waktu, hari, dan jenis menu memiliki pengaruh yang sangat rendah terhadap volume transaksi.
2. Keterbatasan prediksi disebabkan oleh tidak tersedianya faktor eksternal penting, seperti cuaca, event lokal, dan promosi, yang memengaruhi keputusan pembelian pelanggan tetapi tidak tercatat dalam dataset.

5.2 Rekomendasi Bisnis

1. Upselling pada Peak Hours

Meningkatkan nilai transaksi saat jam sibuk (12.00–13.00 dan 17.00–19.00) dengan menawarkan produk berpendapatan tinggi, terutama pizza kategori Veggie dan Chicken ukuran L.

2. Standardisasi Inventaris

Menggunakan distribusi historis sebagai dasar pengelolaan stok, dengan fokus pada ukuran L yang paling diminati serta mengurangi persediaan ukuran XL dan XXL yang memiliki permintaan rendah.



Hasil & Diskusi

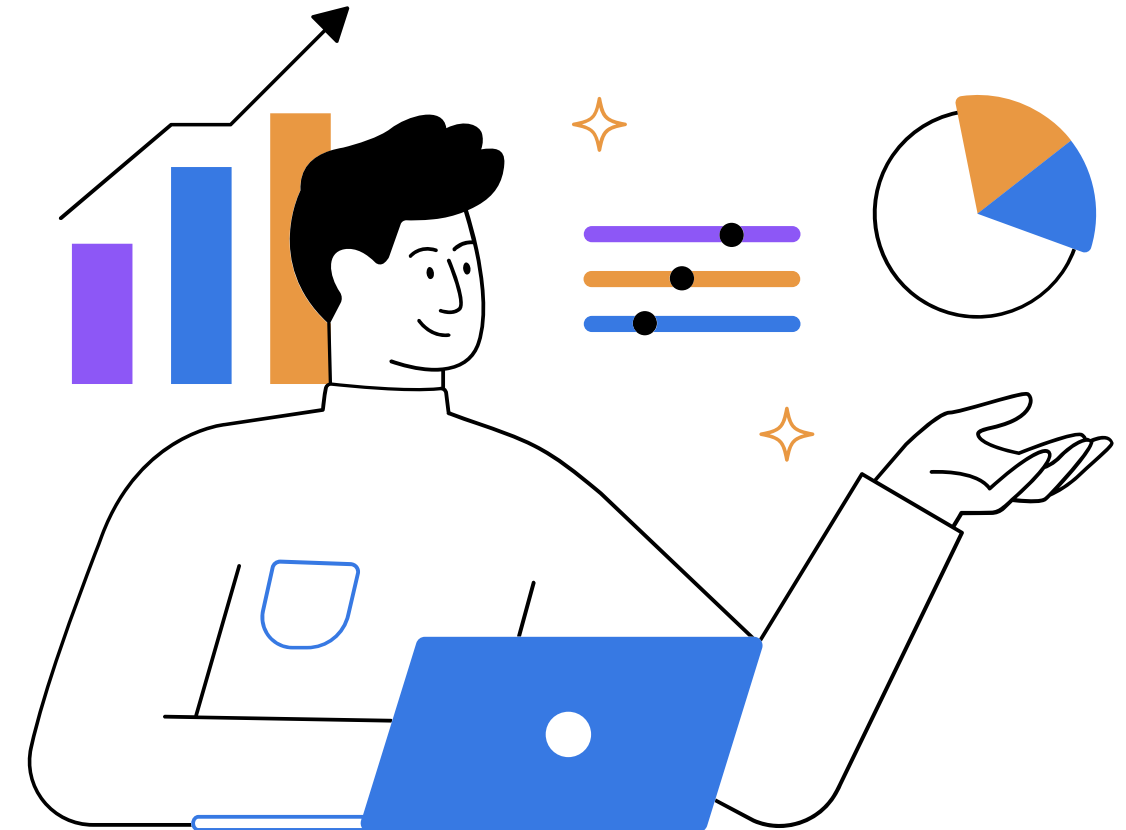
5.3 Evaluasi Risiko

1. Strategi upselling dapat menambah waktu pelayanan dan berpotensi meningkatkan antrean pada jam sibuk.
2. Sistem inventaris berbasis data historis tidak dapat merespons perubahan permintaan yang mendadak, sehingga berisiko menyebabkan kelebihan stok dan pemborosan bahan baku.



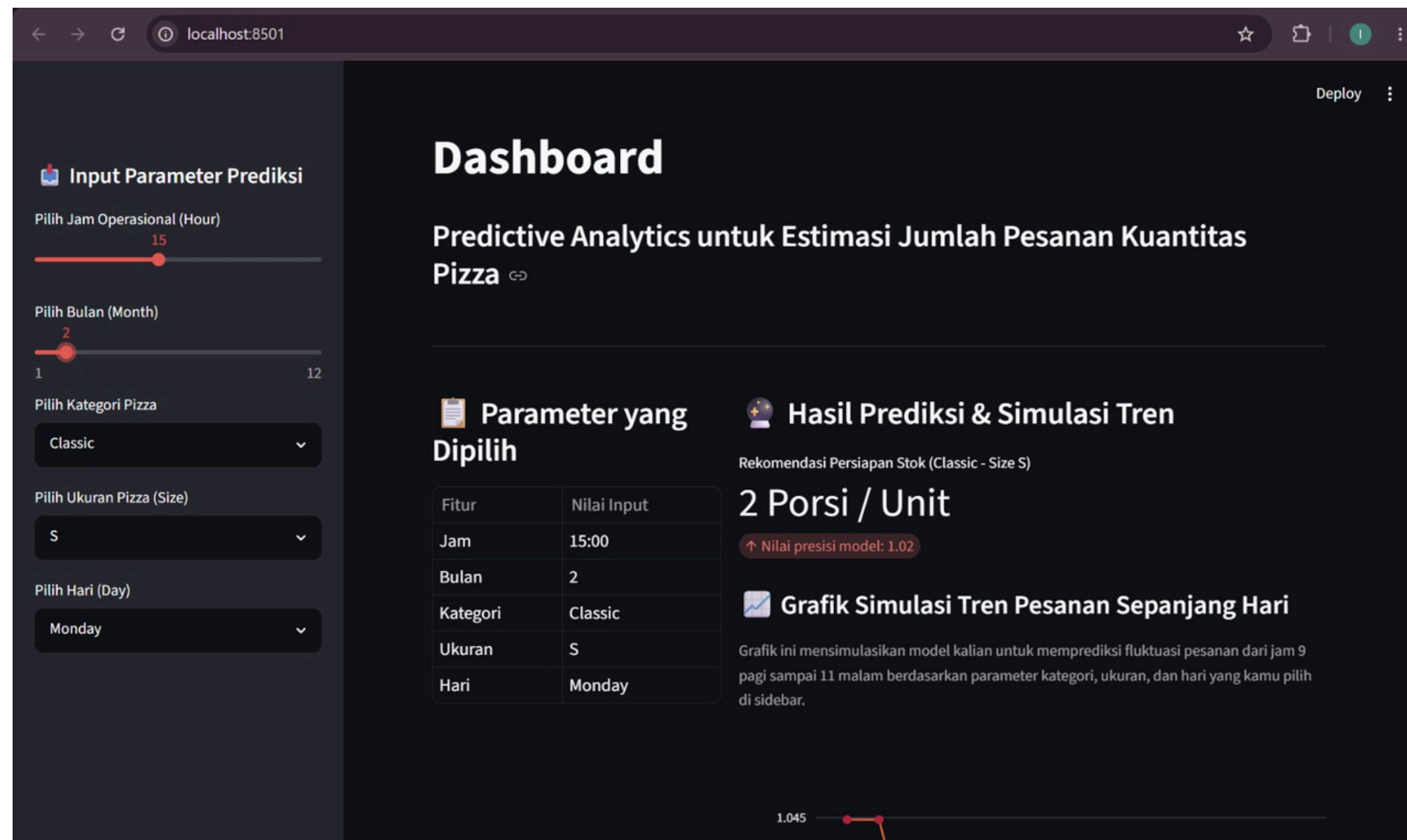


Kontribusi Anggota



Nama	Kontribusi
Leonardo / 2802465912	EDA, Segmented Analysis & Influential Variables
Nathasya Umboh / 2802533043	Base Modeling & Evaluation
Leandra Belva Varissa Perwira Negara / 2802499876	Hyperparameter tuning & Cross Validation, Dashboard

Dashboard



Thank You

