

# e-Commerce Shipping Prediction

External Data Scientist Team X BUSDATA



---

Date Presented:  
March 11th, 2023

# Meet Our Team!

Edwin Juan Sugianto

Nabila Rahmadhani  
Kusuma Putri

Nurul Azizah

Shally Indhani

Vemby Somadias

Ardianto

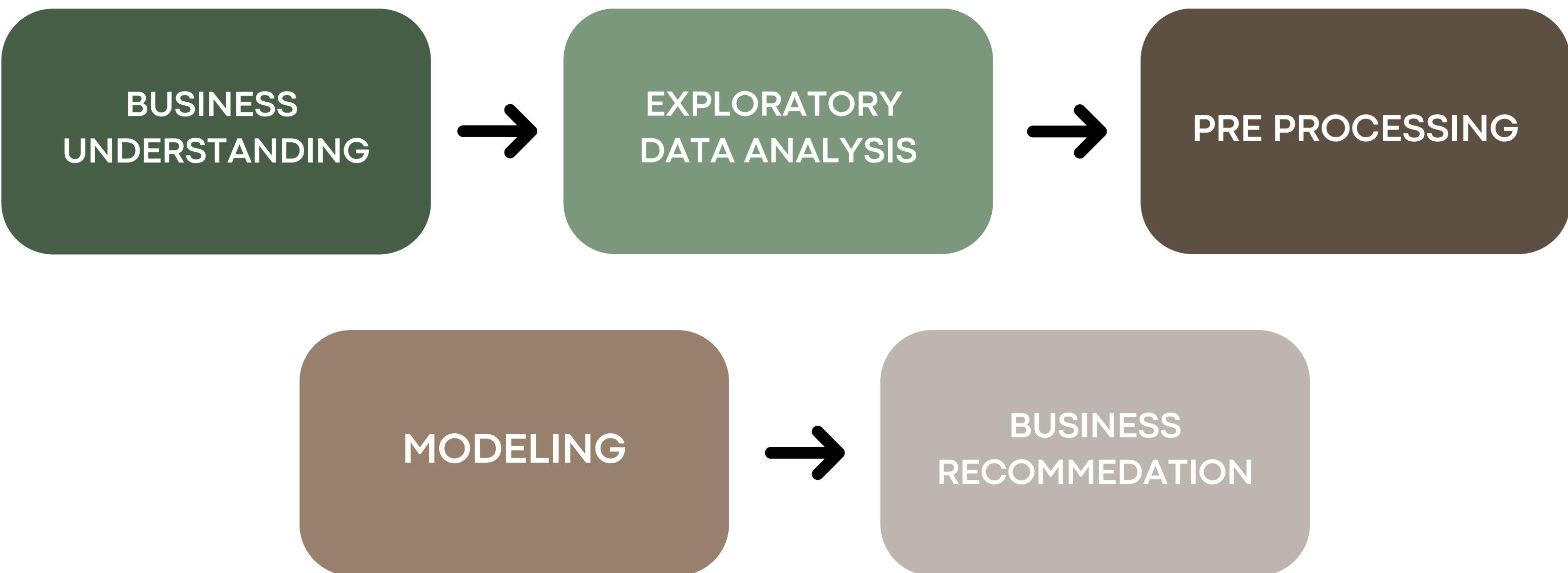
Nurul Fadilah Syahrul

Syella Dwi Safitri

Special Thanks To :

Fadilah Nur Imani

# Work Flow





# Business Understanding

External Data Scientist Team X BUSDATA

[Back to Agenda](#)

# BUSINESS BACKGROUND.

BUSDATA merupakan perusahaan e-commerce yang menjual berbagai produk elektronik.

Sebagai perusahaan data science consultant, kami dipekerjakan secara kontrak oleh BUSDATA untuk mengolah suatu dataset shipment untuk mencari solusi dari permasalahan yang mereka punya melalui dataset yang tersedia.



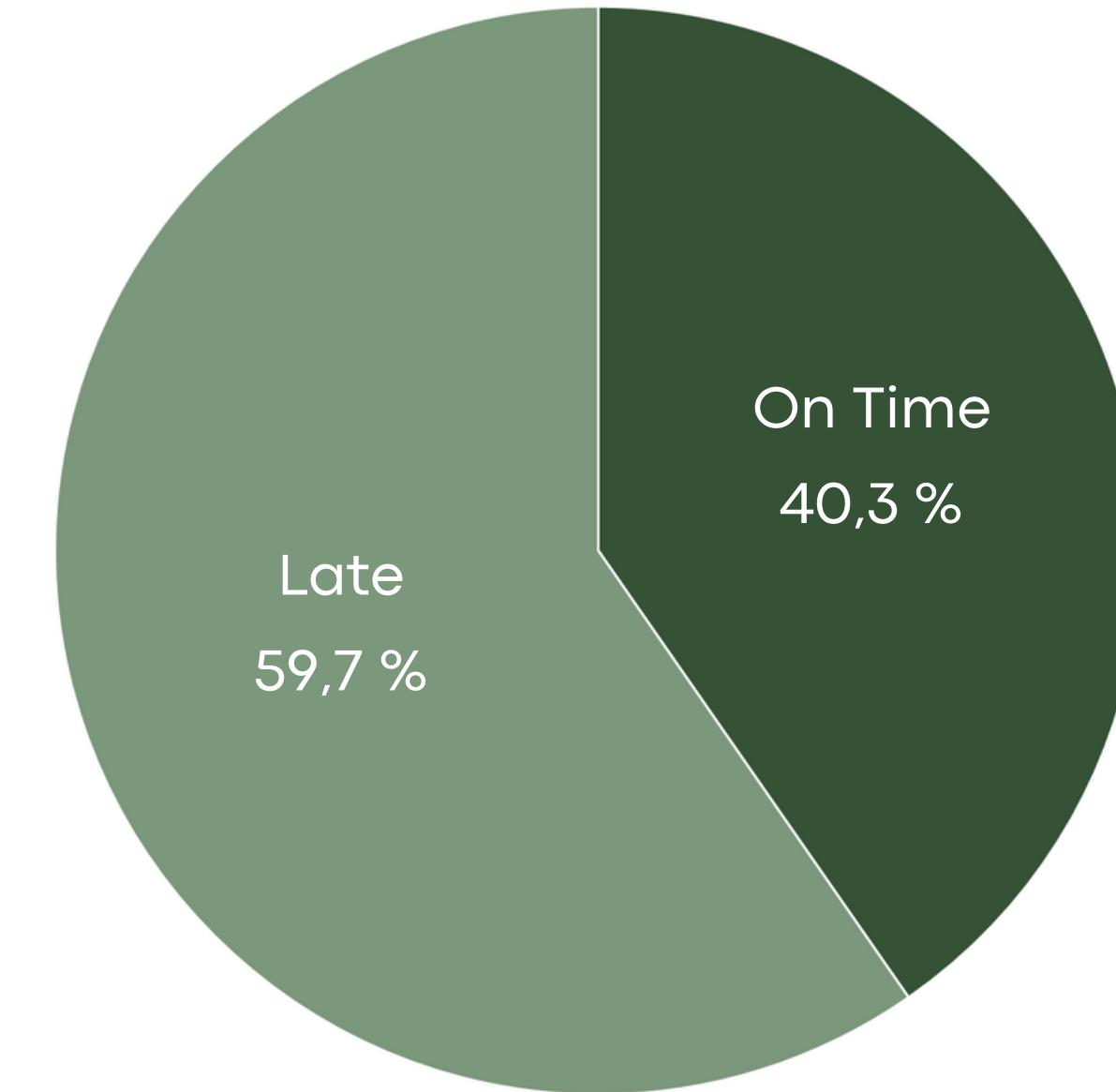
Link Dataset



Data Science Consultant

# Problem Statement

Data set BUSDATA E-Commerce Shipping menunjukan bahwa dari **10.999** sampel transaksi pembelian online, **59.7%** transaksi diantaranya mengalami **keterlambatan** dalam pengiriman, ini mengakibatkan banyaknya complaint yang di terima. Dikhawatirkan apabila dibiarkan akan menurunkan angka customer rating. Oleh karena itu, E-Commerce BUSDATA ingin meningkatkan performa dalam ketepatan pengiriman barang.



Late Shipment Percentage

# Objective, Business Metrics, Goal.

## Objective.

Membangun model untuk dapat memprediksi ketepatan waktu pengiriman barang dan Memberikan rekomendasi dari insight bisnis untuk mengatasi keterlambatan waktu pengiriman.

## Business Metrics.

- On Time Delivery
- Average Customer Rating

## Goals.

Menjaga agar pelayanan berkualitas lewat angka customer rating

# Exploratory Data Analysis



[Back to Agenda](#)

# HASIL EKSPLORASI

ID  
Warehouse\_block  
Mode of Shipment  
Customer\_care\_calls  
Customer\_rating  
Cost\_of\_the\_Product  
Prior\_purchases  
Product\_importance  
Gender  
Discount\_offered  
Weight\_in\_gms

Reached.on.Time\_Y.N



10.999 data pelanggan



tidak ada missing value  
maupun data duplikat



data feature konsisten



2 feature memiliki  
distribusi tidak normal  
dan memiliki outliers

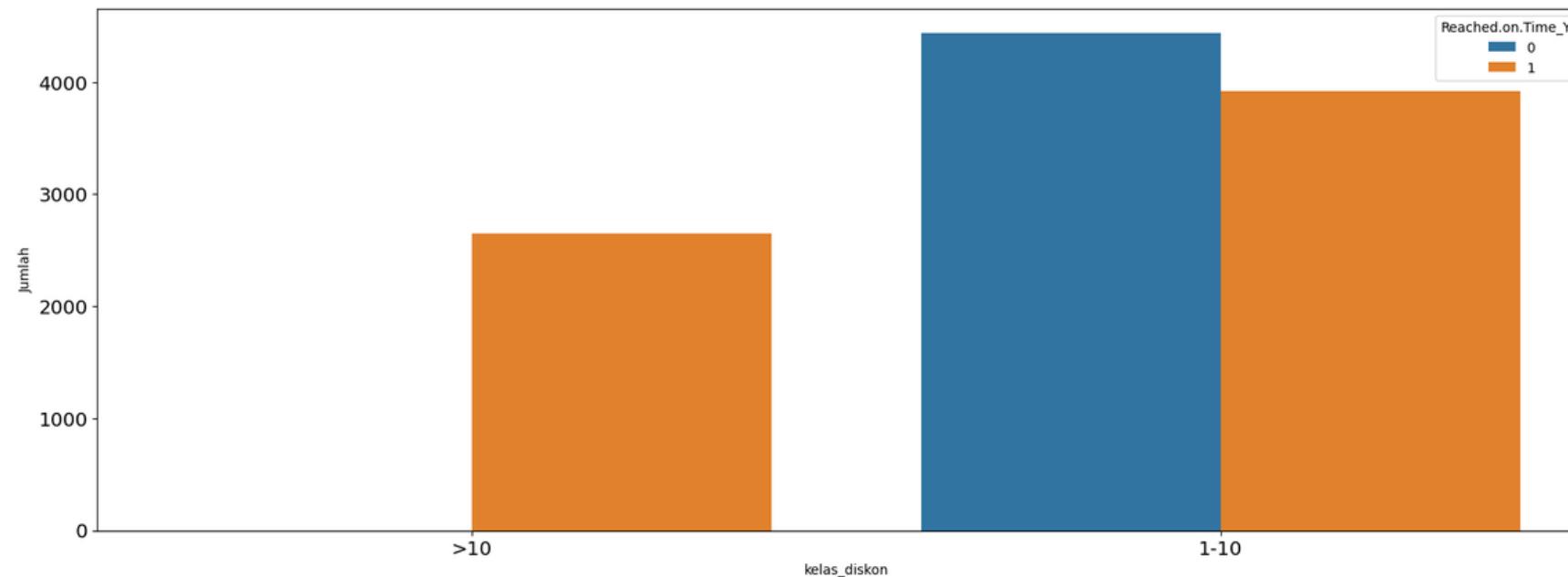


Reached\_on\_time memiliki korelasi positif dengan Discount\_offered sebesar 0.40

Reached\_on\_time memiliki korelasi negatif dengan Weight\_in\_gms sebesar -0.27

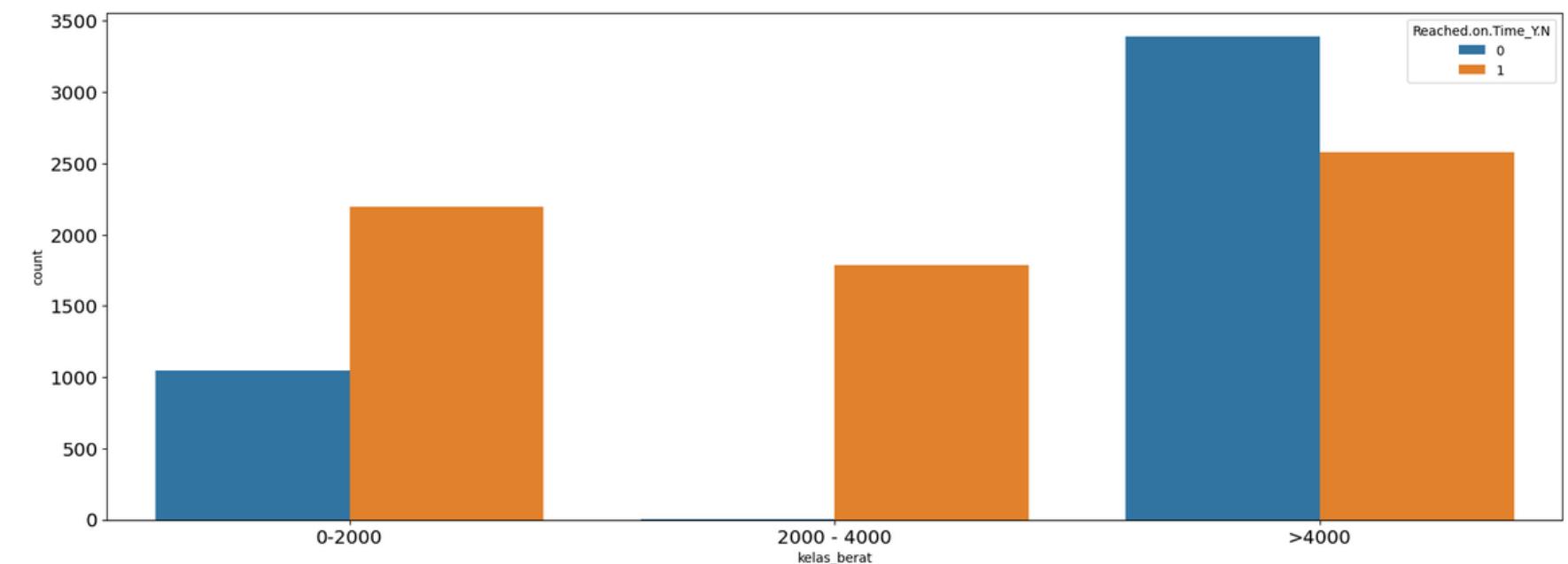
Weight\_in\_gms memiliki korelasi negatif dengan Discount\_offered sebesar -0.38

█ Late    █ On time



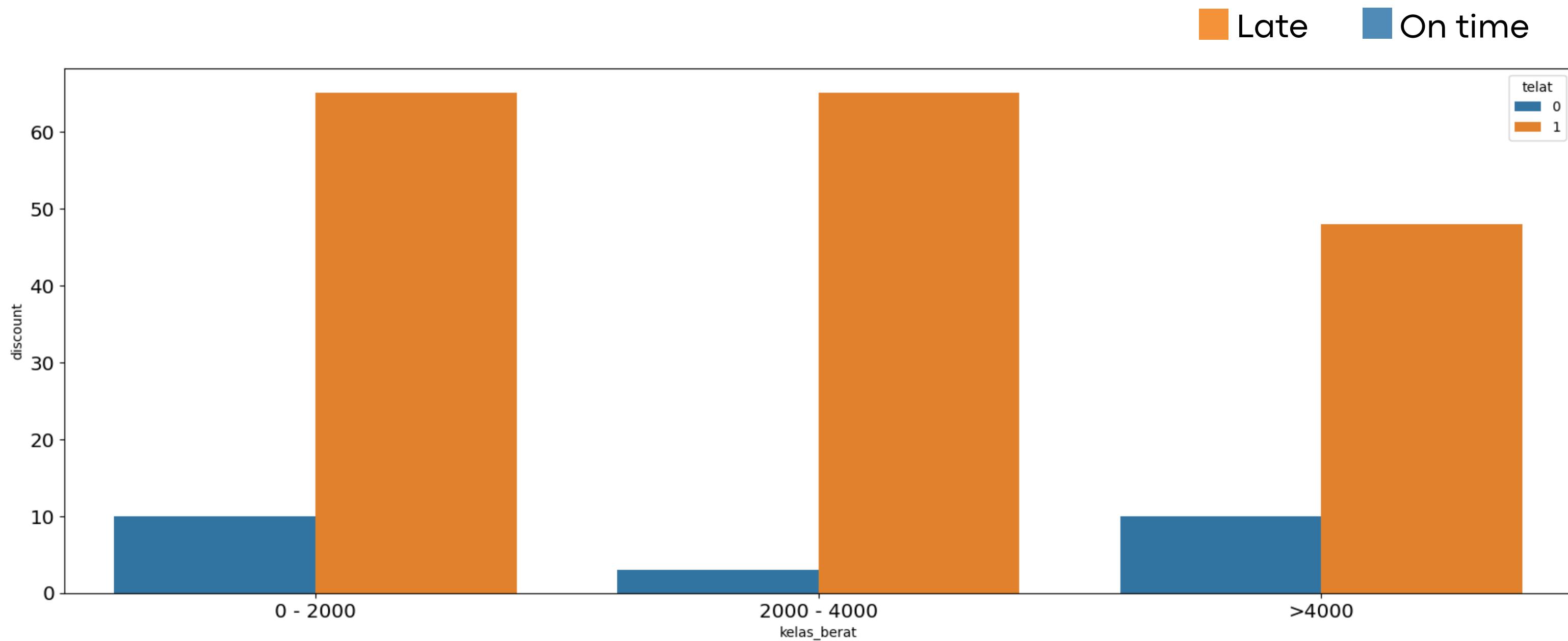
### Target vs Diskon

Transaksi dengan diskon di atas 10% mengalami 100% keterlambatan

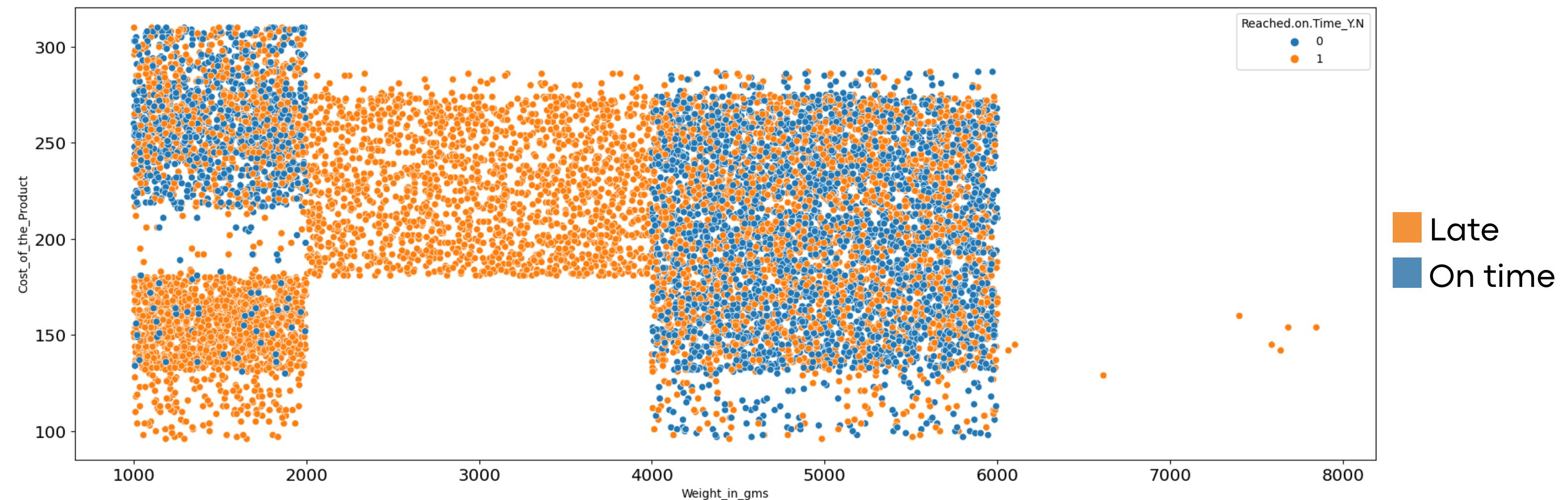


### Target vs Berat

- Transaksi dengan berat di bawah 2000 g didominasi terlambat.
- Transaksi dengan berat produk 2001 - 4000 g mengalami 100% keterlambatan

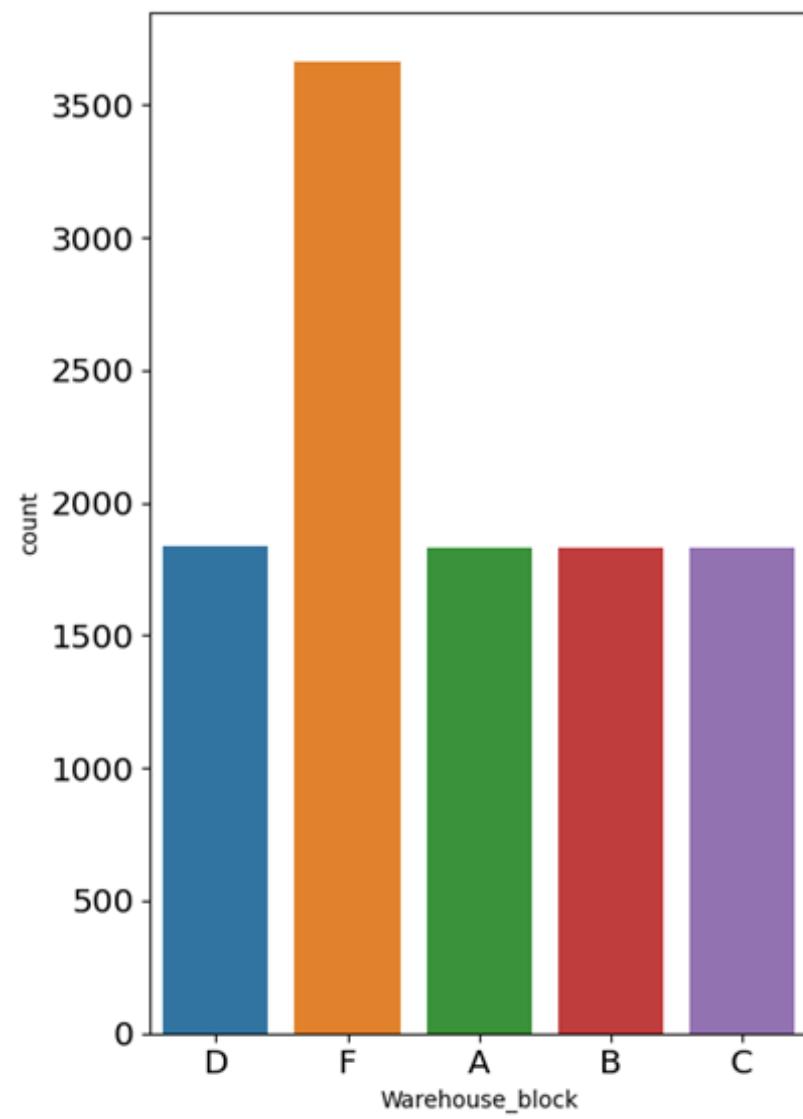


Terlihat produk dengan distribusi discount hingga 65% mayoritas berada di bawah beban 4000 gr sedangkan beban pada 4001-6000 gr mayoritas hanya mendapat discount maksimal 10%

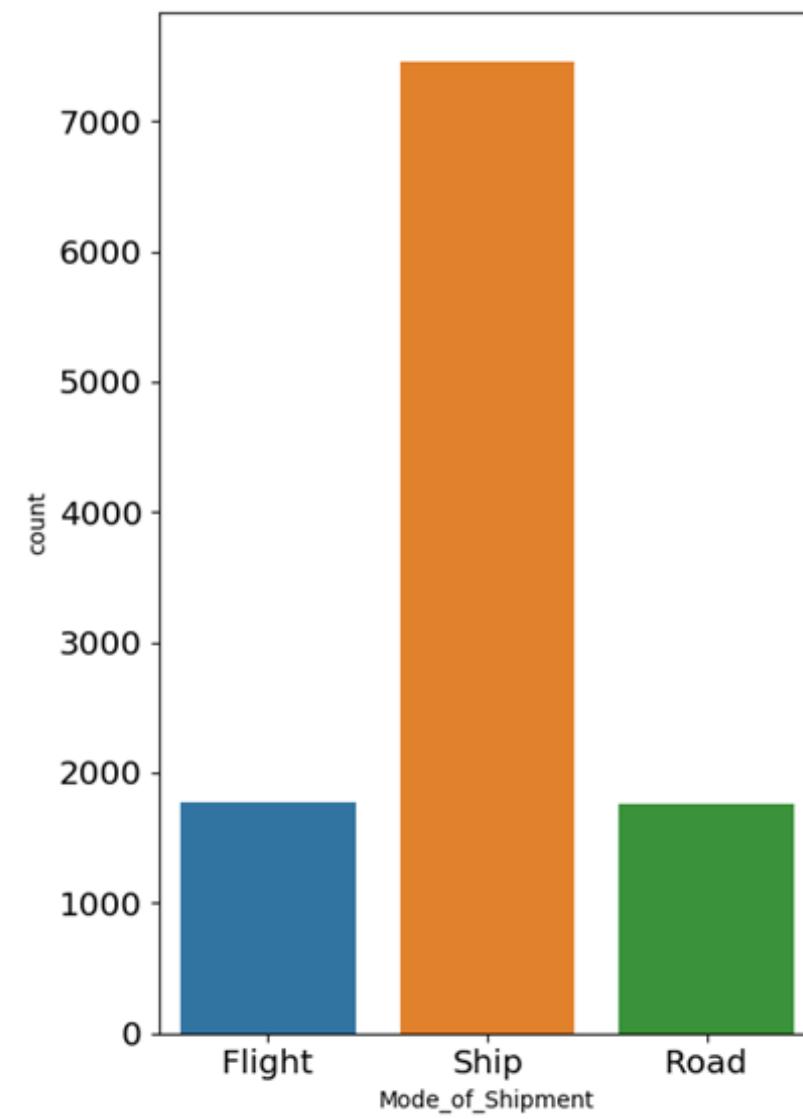


Pada area sekitar median  
Weight\_in\_gms memiliki frekuensi  
produk terendah namun dari data  
reach on time tercatat produk  
100% terlambat dan data  
Cost\_of\_the\_Product relatif tinggi

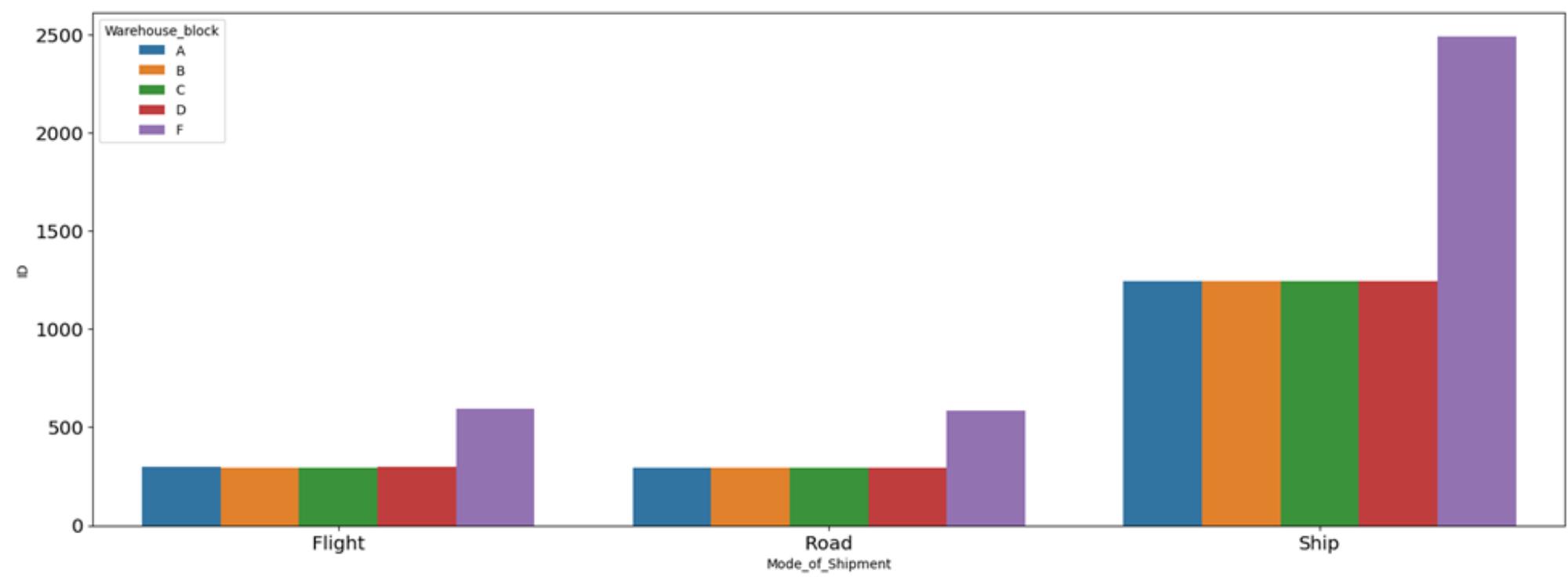
Barang mahal  
tidak sampai  
tepuk waktu



**Warehouse blok F**  
mendominasi  
sekitar 33%.



**Jalur laut (Ship)**  
mendominasi  
sekitar 67,8%.



**Pengiriman dari warehouse F dan jalur laut**  
**overcrowded**

adanya 'overcrowded' pada warehouse F dan pengiriman menggunakan jalur laut.

# Data Preprocessing



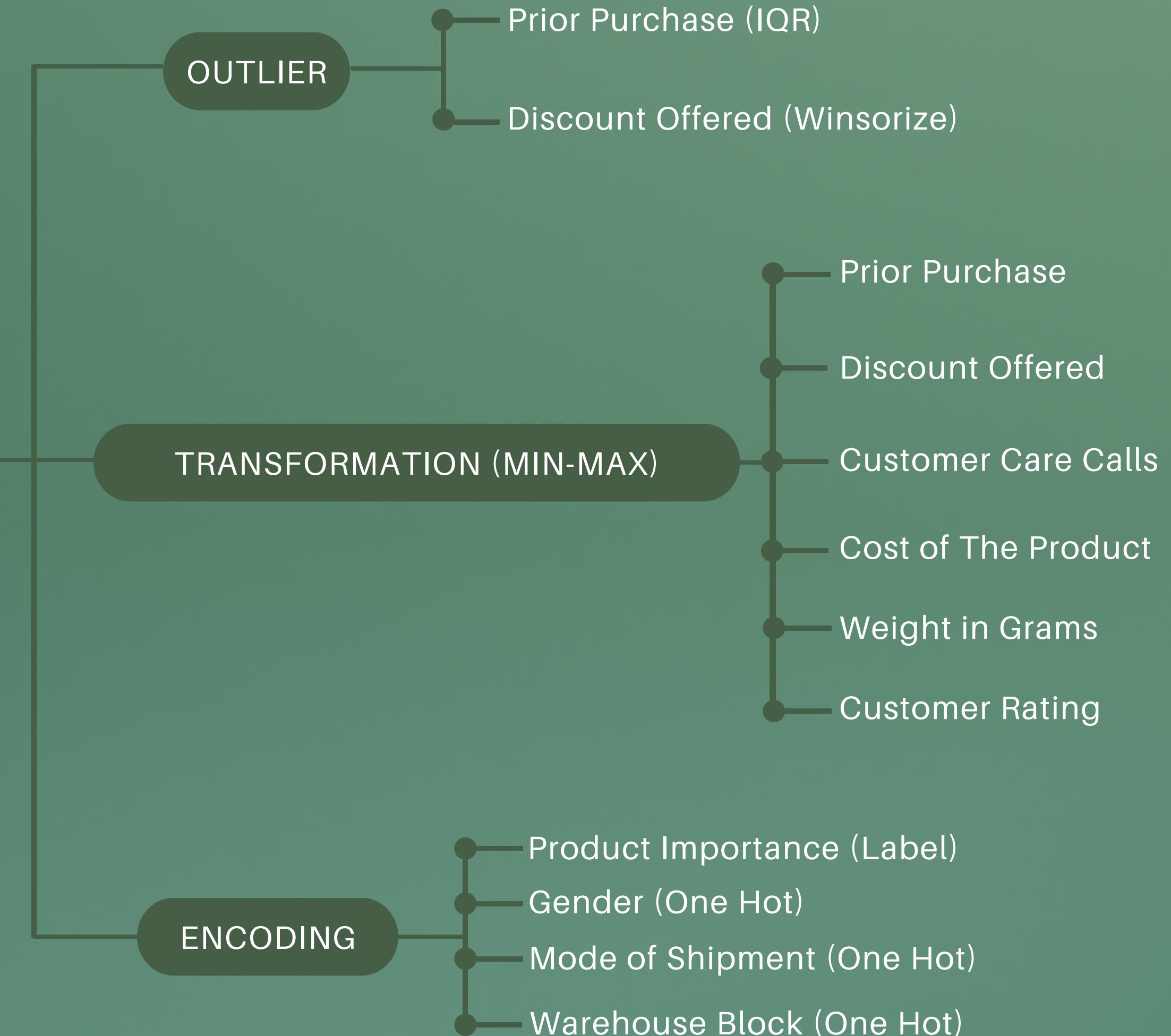
[Back to Agenda](#)

# SPLIT DATASET

TRAIN	TEST
8799	2200



# PRE- PROCESSING





# Modeling & Business Recommendation

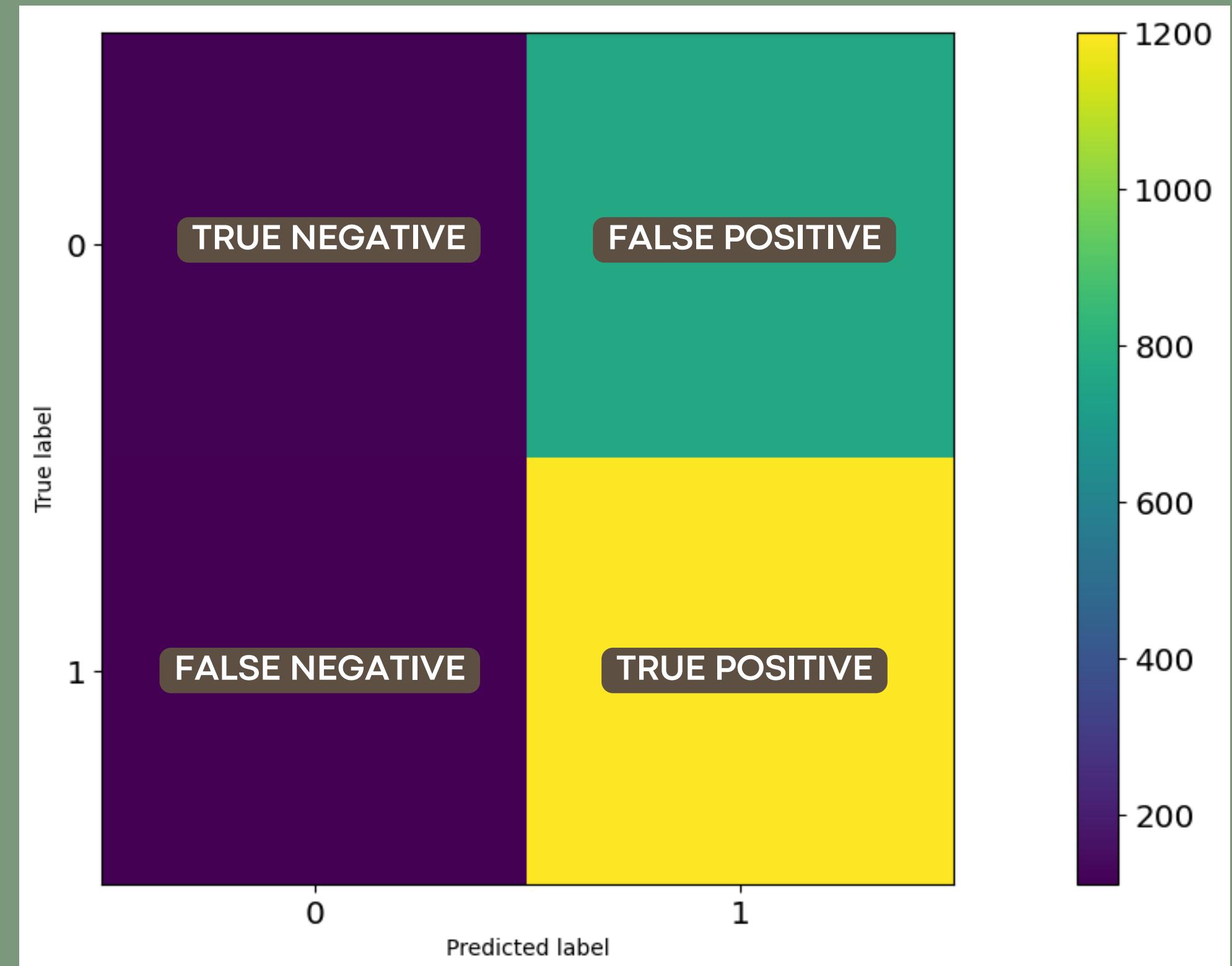
[Back to Agenda](#)

# Fokus Metric Value

**RECALL**

$$\frac{TP}{TP+FN}$$

- Prediksi tepat waktu (0) padahal telat (1) -> **FALSE NEGATIVE**
- Menyebabkan ekspektasi pelanggan terhadap suatu produk sampai tepat waktu meningkat namun setelah barang sampai dengan telat, satisfaction pelanggan menurun -> **rating rendah.**



	Test		Train	
	RECALL	ACCURACY	RECALL	ACCURACY
LOGISTIC REGRESSION	0.91	0.60	0.95	0.64
KNN	0.70	0.63	0.73	0.66
DECISION TREE	0.68	0.65	1.00	0.75
RANDOM FOREST	0.81	0.65	0.81	0.65

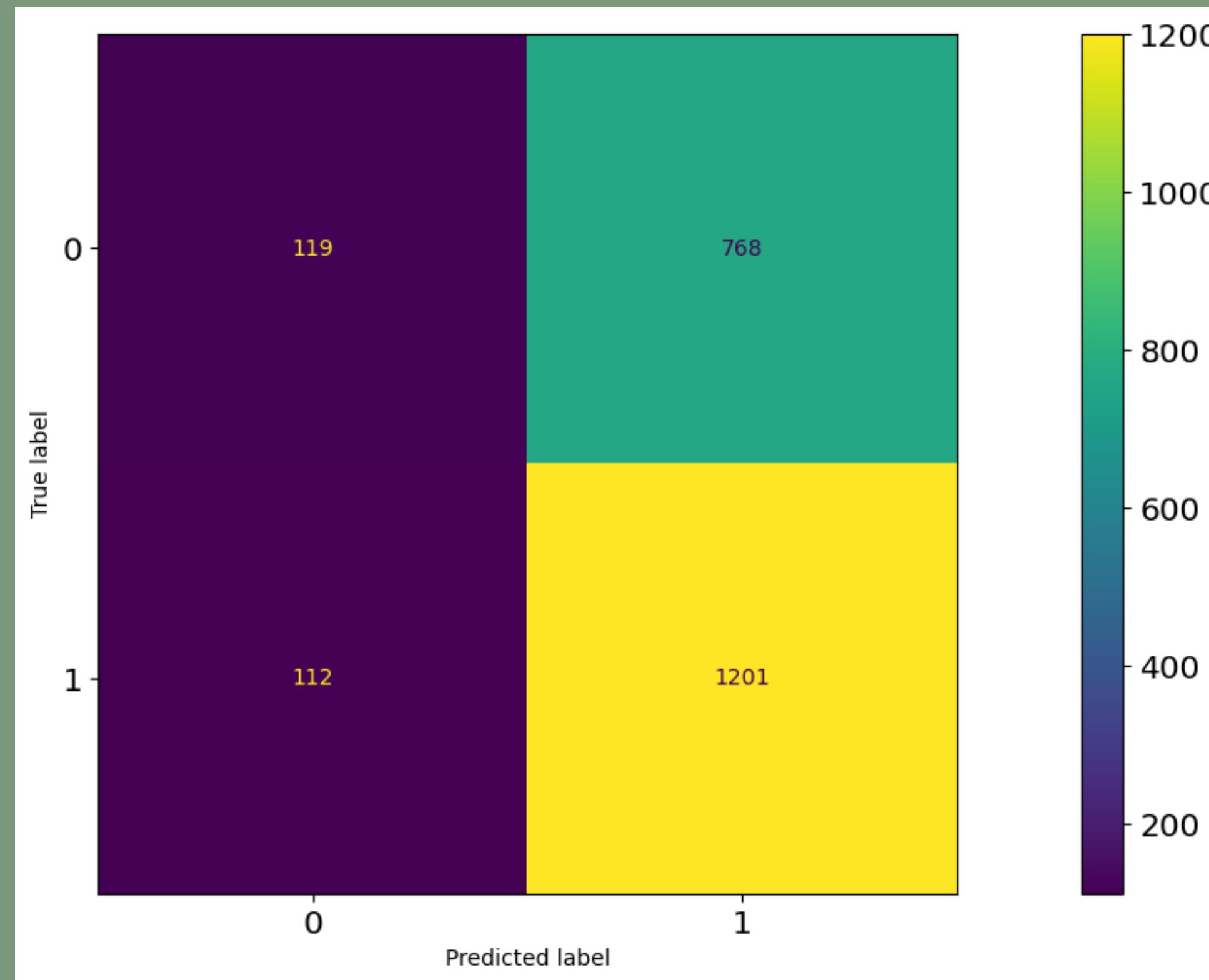
Pengujian dilakukan dengan 4 model machine learning

Logistic Regression menjadi model yang dapat menghasilkan performa terbaik

- Hubungan linear antar fitur.
- Fitur dan jumlah data yang digunakan tidak terlalu besar.
- Model kompleks overfit

# dengan Hyperparameter Tuning

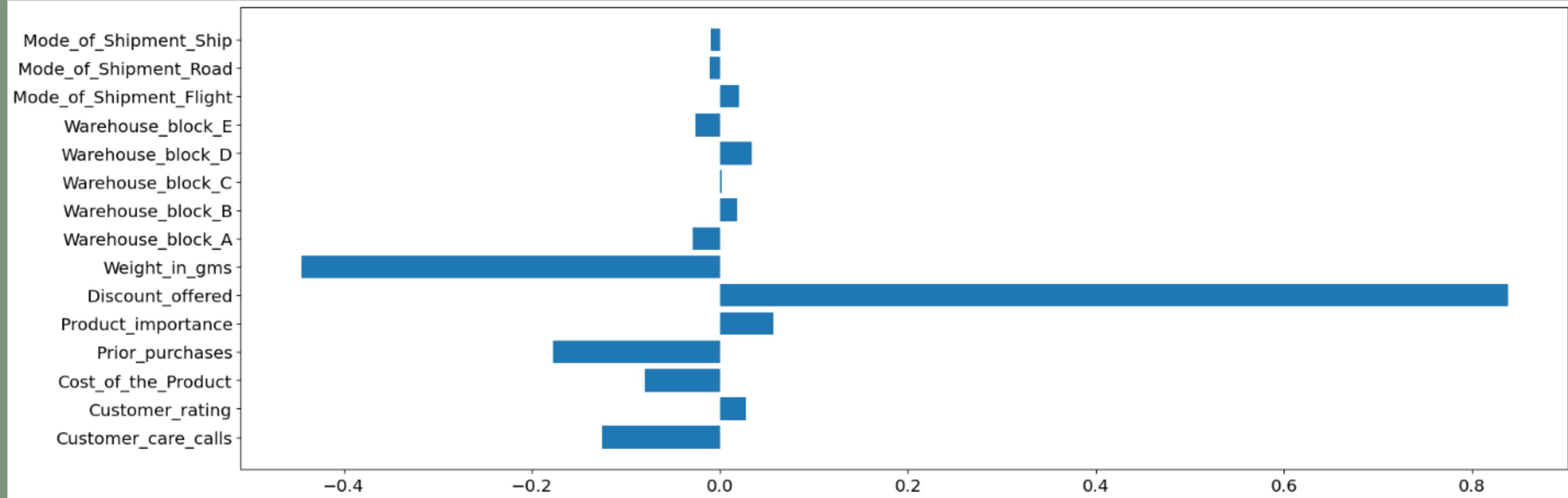
# Confusion Matrix



Jika dilihat dari **confusion matrix**:

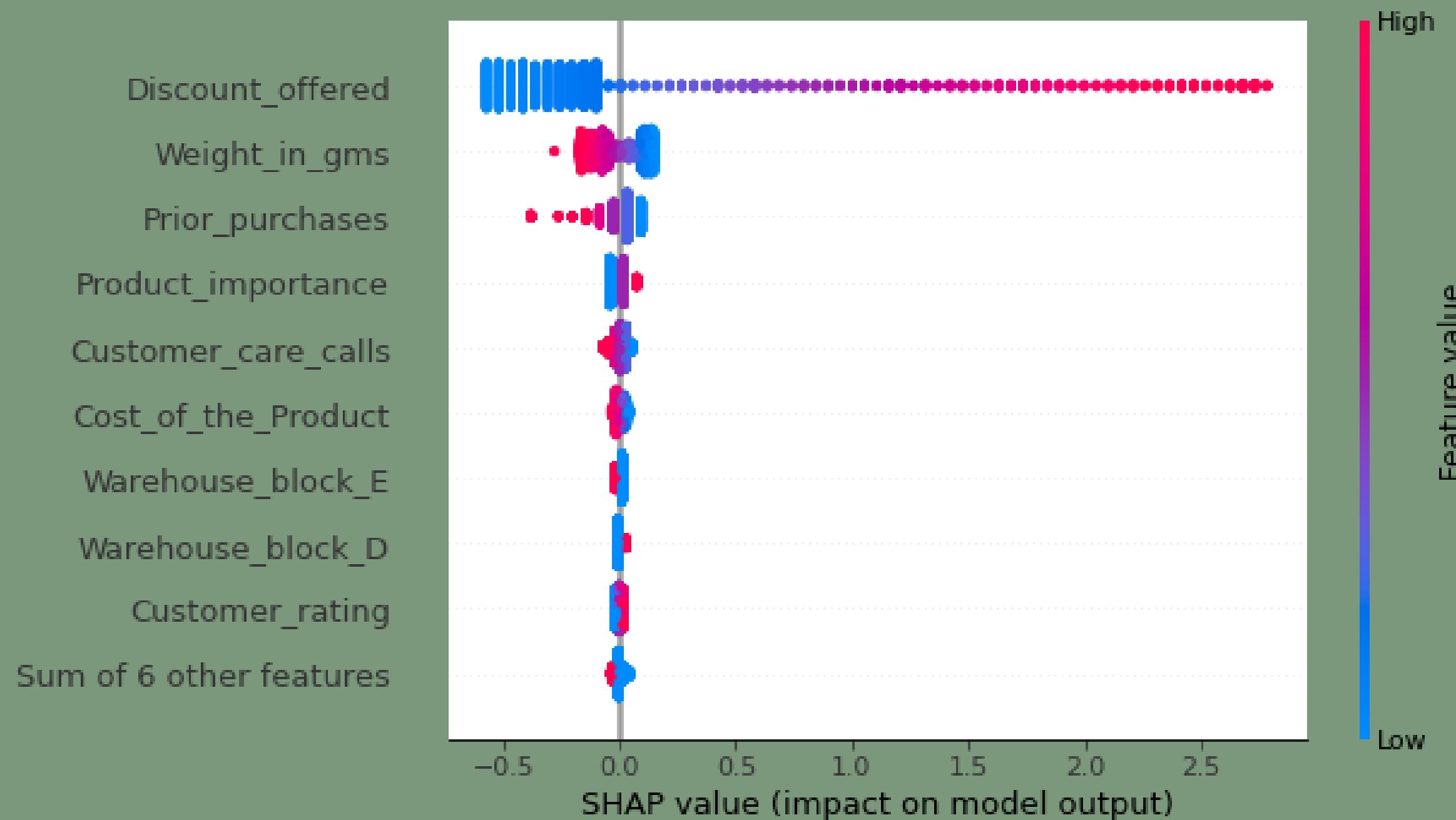
- Dari **1313 produk (60% test)** tidak datang tepat waktu, model memprediksi ada **112 produk** yang diprediksi tepat waktu (**FN**).
- Dari **887 produk (40% test)** datang tepat waktu, model memprediksi ada **768 produk** yang diprediksi tidak tepat waktu (**FP**).

# Feature Importance



Weight\_in\_gms dan Discount\_offered menjadi feature yang paling berpengaruh pada model

# Feature Importance



**Weight\_in\_gms** dan **Discount\_offered** menjadi feature yang paling berpengaruh pada model

# Recommendations

## Manajemen Shipping

### INTERNAL (by BUSDATA)

- Penambahan dan pelatihan SDM dalam hal pick up, packing, dan manajemen gudang.
- Pemerataan penggunaan gudang untuk staging, transit, dan tracking.

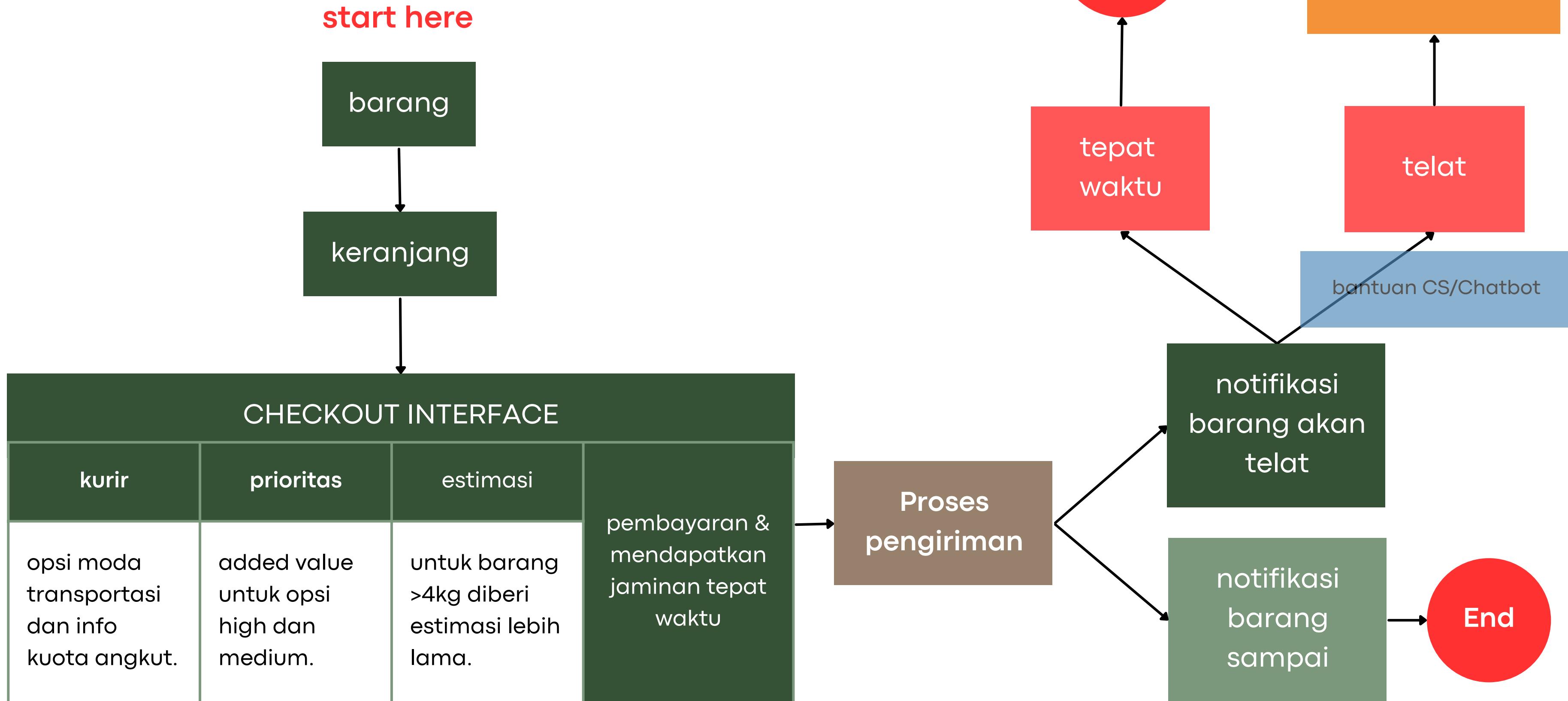
sumber: [bringg.com](http://bringg.com)

### INTERACTION WITH CUSTOMERS (by system)

- Pemberian opsi kurir dengan informasi estimasi datang.
- Opsi prioritas khusus barang yang mahal (mis. kurir same day/instan) dengan biaya tambahan.
- Khusus untuk barang yang berat (>4kg) diberikan estimasi lebih lama.
- Jaminan tepat waktu berupa refund ongkos kirim jika barang masih telat datang.
- Notifikasi barang yang rawan telat beserta alasan yang akan muncul di akun pembeli (mis. akibat high season/kejadian tidak terduga).



## HOW IT WORKS - interaction by system



# Business Simulation

Business Metrics	before	after
On Time Delivery	LATE = 59.7% ON TIME = 40.3%	LATE = 5.4% ON TIME = 94.6% 134.7% dari sebelum modeling.
Customer Rating (mean)	2.99	3.42 (14.3% dari sebelum modeling)

Karena model dapat mendeteksi ketelatan dengan baik (91%) dan dengan asumsi transaksi yang diprediksikan telat tersebut diberi perlakuan khusus hingga menjadi 100% on time.



**TERIMA KASIH**

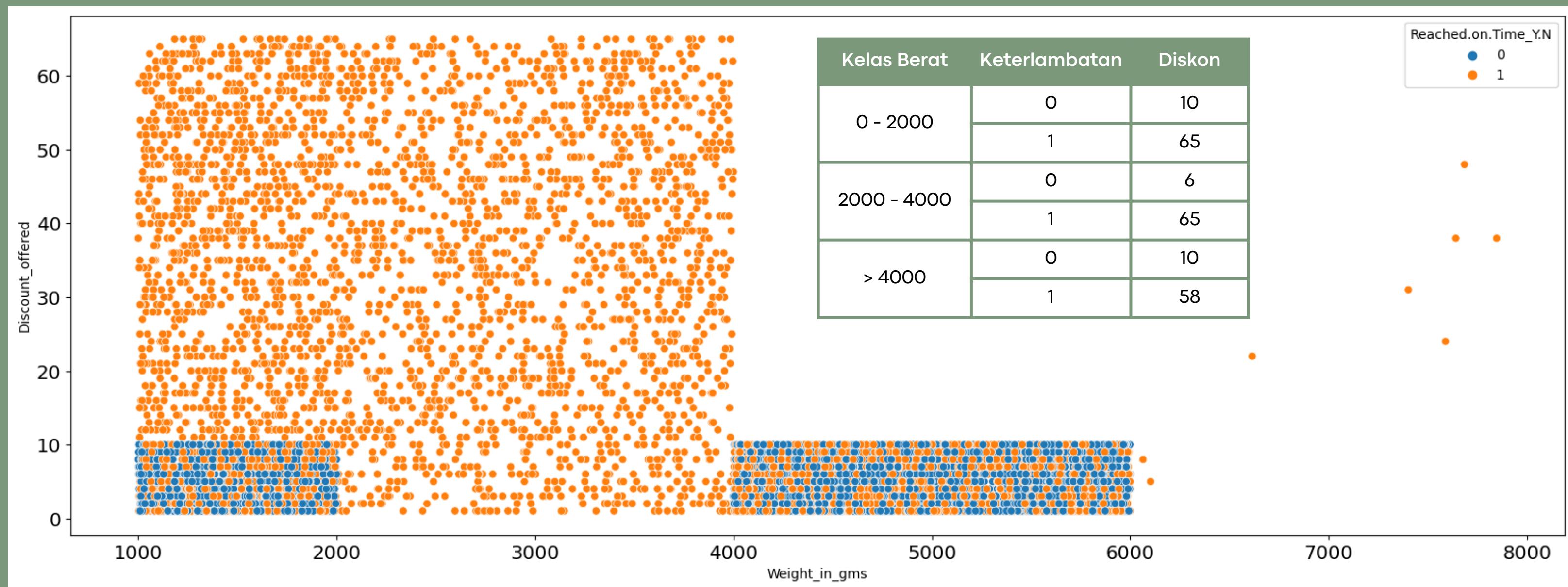


Add Company Name

# APPENDIX

[Back to Agenda](#)

# Scatterplot Berat vs Diskon

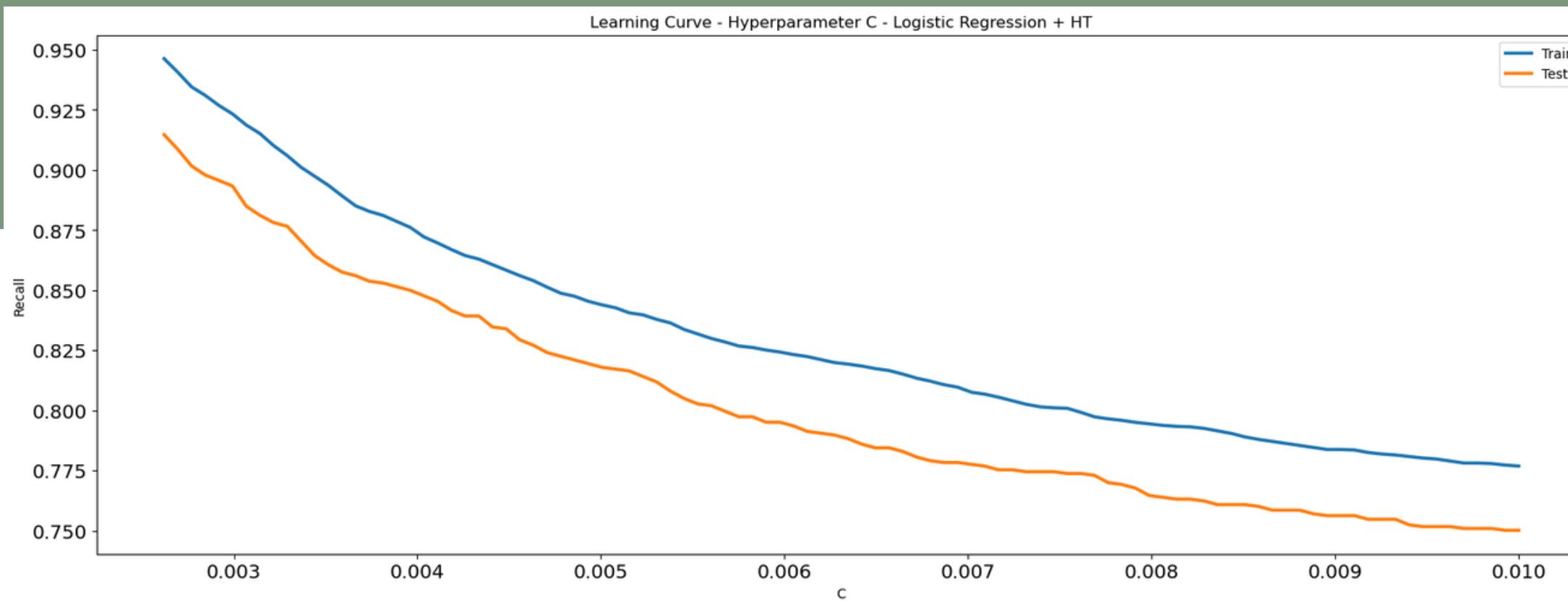


# All Metrics Value (with Hyperparameter Tuning)

	Test					Train				
	RECALL	ACCURACY	Precission	ROC_AUC	F1 SCORE	RECALL	ACCURACY	Precission	ROC_AU C	F1 SCORE
LOGISTIC REGRESSION	0.91	0.60	0.61	0.71	0.73	0.95	0.64	0.61	0.72	0.74
KNN	0.70	0.63	0.68	0.71	0.69	0.73	0.66	0.71	0.74	0.72
DECISION TREE	0.68	0.65	0.72	0.73	0.70	1.00	0.75	0.81	0.84	0.79
RANDOM FOREST	0.81	0.65	0.67	0.75	0.73	0.81	0.65	0.67	0.75	0.73



# HT Settings (1)



## LOGISTIC REGRESSION

- `random_state = 42`
- `penalty = l2`
- `C = 0.00262020202020205`



# HT Settings (2)

## KNN

- `n_neighbors` = 81 - 87
- `weights` = uniform
- `p` = 2
- `algorithm` = kd\_tree

## Decision Tree

- `random_state` = 42
- `max_depth` = 34
- `min_samples_split` = 10
- `min_samples_leaf` = 10
- `max_features` = auto
- `criterion` = entropy
- `splitter` = best

## Random Forest

- `random_state` = 42
- `n_estimators` = 9
- `max_depth` = 2
- `min_samples_split` = 4
- `min_samples_leaf` = 13



# All Recall CV Value (cv = 5) on Tuned Models

	Test	Train
LOGISTIC REGRESSION	0.68	0.80
KNN	0.53	0.66
DECISION TREE	0.59	0.74
RANDOM FOREST	0.61	0.81