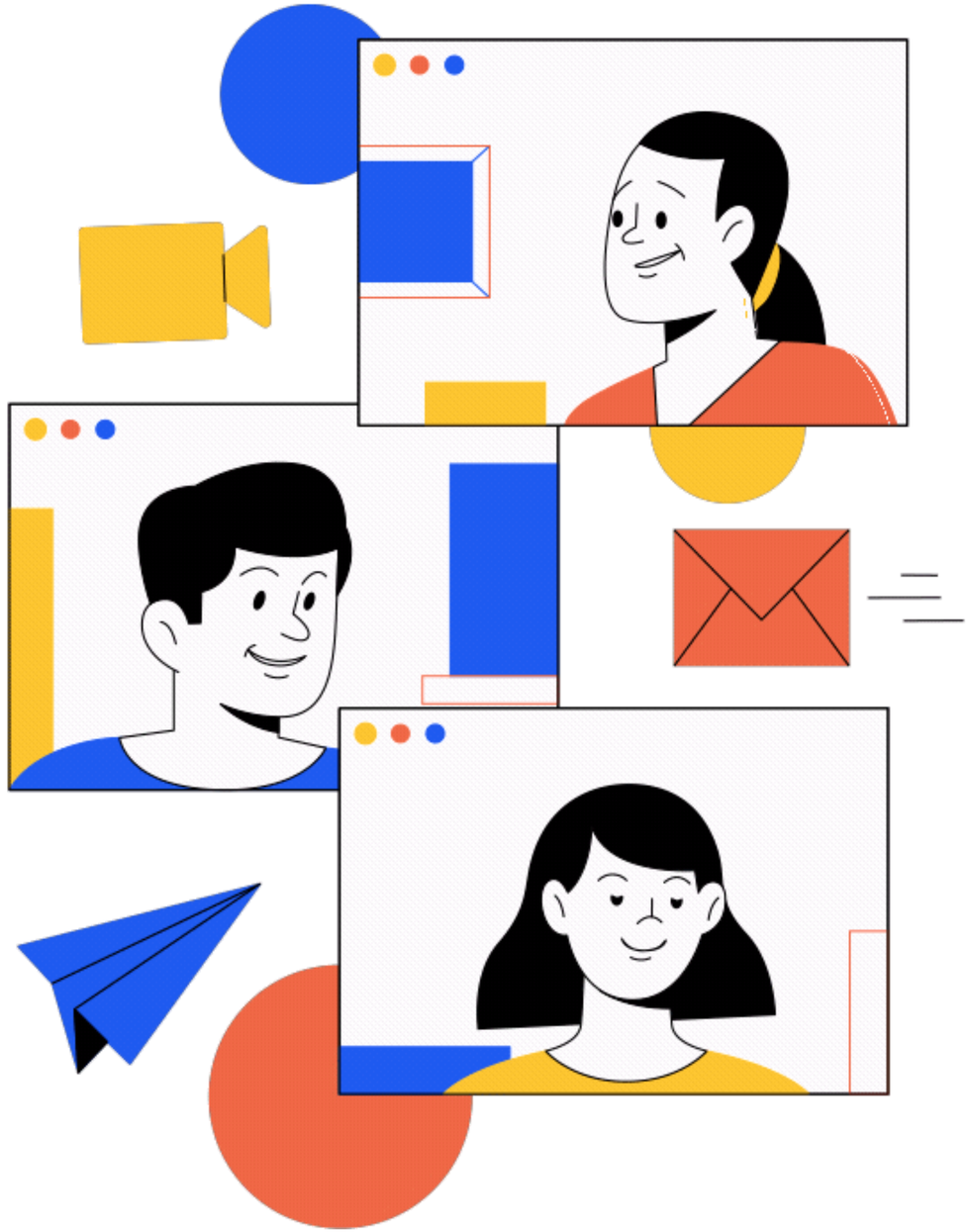
# FEATURE IMPORTANCE ANALYSIS

using ANOVA and Machine Learning



# ABOUT DATA

Dataset Superstore berisi data transaksi penjualan dari sebuah perusahaan yang mencakup informasi pelanggan, produk, lokasi, dan detail pemesanan.

**51.290**
**Row Data**

## SuperStore Dataset

- Order_Id
- Order_Date
- Ship_Date
- Customer_Name
- Product_Id
- Product_Name
- Segment
- State
- Country
- Market
- Region
- Category
- Sub_Category
- Sales
- Quantity
- Discount
- Profit
- Shipping_Cost
- Order_Priority

# TUJUAN ANALISIS

Mengetahui Feature Importance dari variabel target Sales.

Membandingkan hasil Feature Importance dari ANOVA dengan hasil Machine Learning.

Membandingkan performa berbagai machine learning, termasuk Gradient Boosting, XGBoost, Random Forest, Linear Regression, dan Decision Tree.

# FEATURE IMPORTANCE BY ANOVA

- Menguji pengaruh dari dua atau lebih faktor
- Mendeteksi perbedaan varians antar kelompok

P-Value < a, **Tolak HO**, Terdapat Perbedaan Signifikan
P-Value > a, **Terima HO**, Tidak Terdapat Perbedaan Signifikan

| No | Variabel | p-value (PR(>F)) | Keterangan Singkat |
|---|---|---|---|
| 1 | **ship_mode** | 9.282117e-01 | Tidak signifikan |
| 2 | **segment** | 8.704638e-01 | Tidak signifikan |
| 3 | **order_priority** | 5.762584e-01 | Tidak signifikan |
| 4 | **state** | 8.717133e-71 | Sangat signifikan, meskipun banyak level |
| 5 | **region** | 7.153350e-134 | Sangat signifikan |
| 6 | **market** | 1.200073e-155 | Sangat signifikan |
| 7 | **country** | 1.512257e-172 | Sangat signifikan |
| 8 | **category** | 0.000000e+00 | Signifikan (paling kecil) |
| 9 | **sub_category** | 0.000000e+00 | Signifikan (paling kecil) |

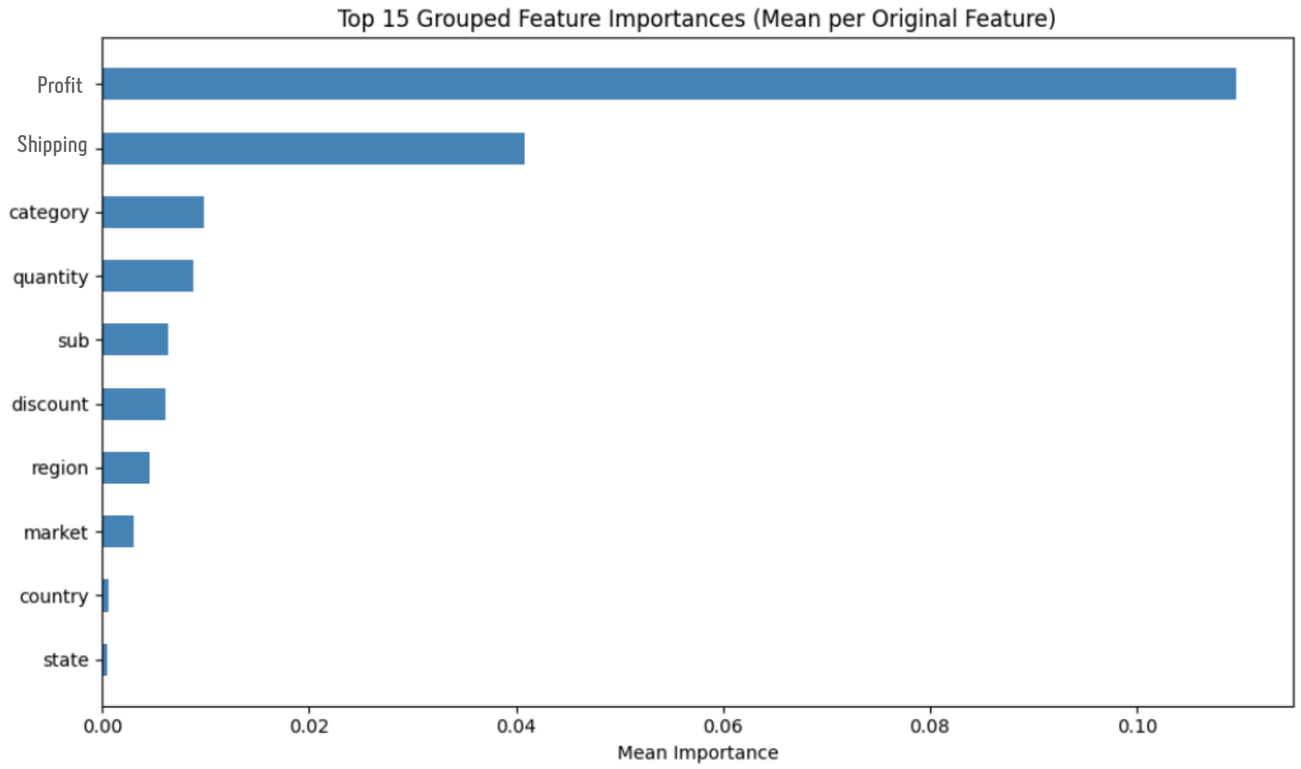
## Variabel hasil pengujian ANOVA

1. **Sub_Category**
2. **Category**
3. **Country**
4. **Market**
5. **Region**
6. **State**

# XGBOOST

Membagi dataset menjadi data train dan data testing,
Dengan variabel tambahan **Shipping Cost**, **Profit**, **Quanitity**, dan **Diskon**.

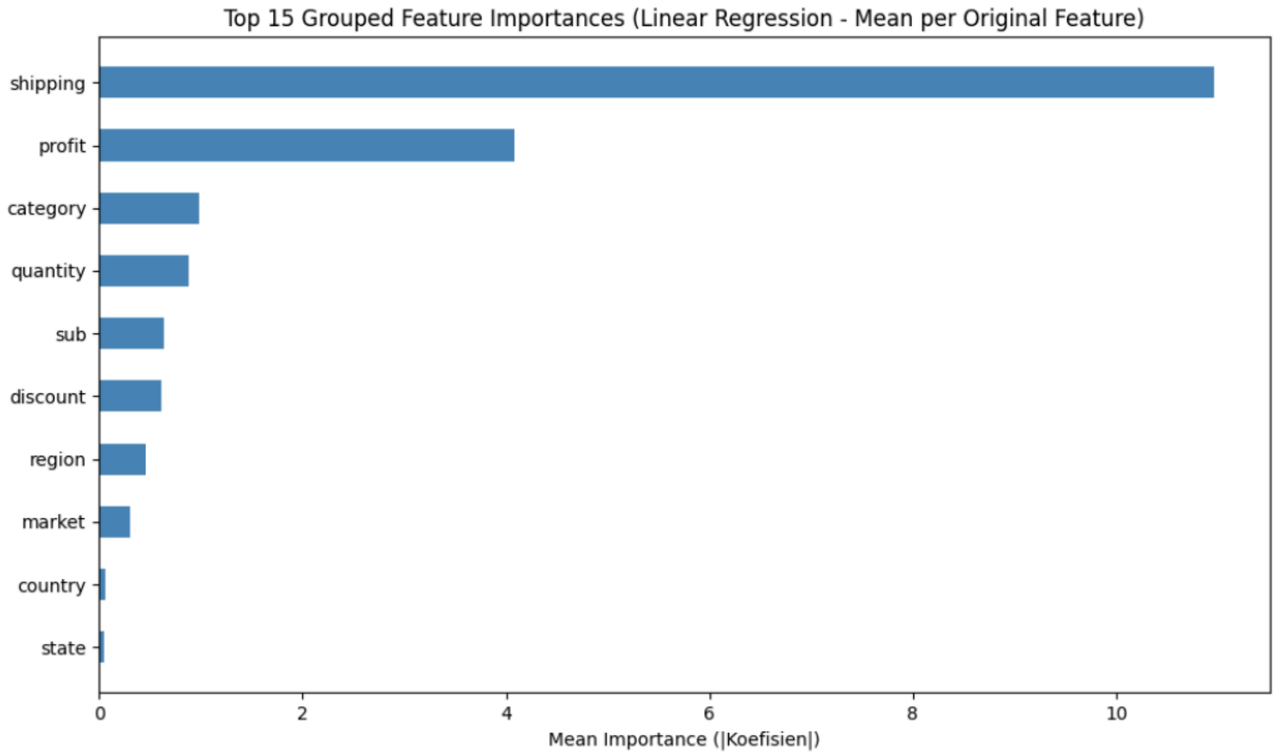
Dilanjutkan pemodelan Machine Learning menggunakan target '**Sales**'.




```
R² Score: 0.7831
RMSE: 230.34
Persentase Error: 92.67% dari rata-rata penjualan
```

- Model XGBoost cukup baik secara $R^2$, namun RMSE dan % error masih tinggi
- Feature Importance tertinggi adalah variabel **shipping_cost** dan **profit**

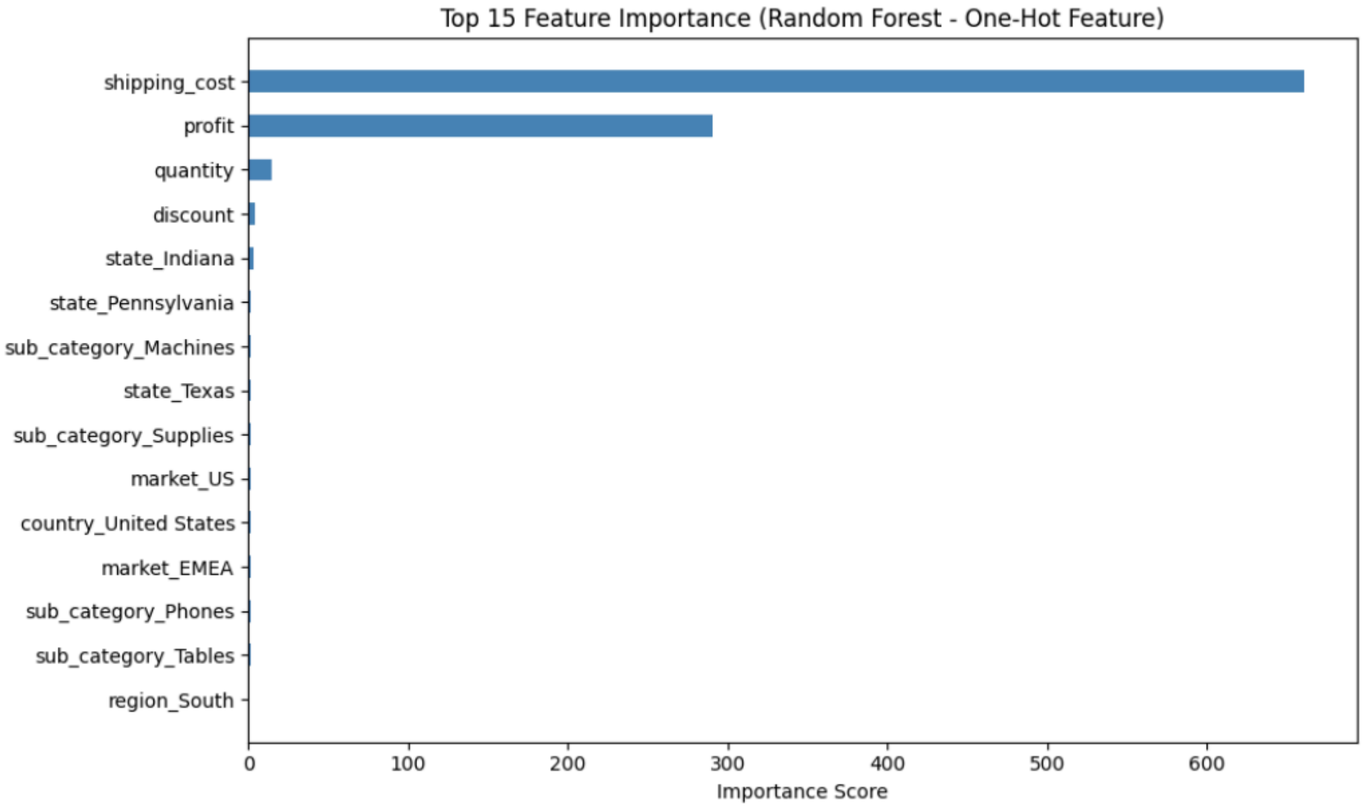
# LINEAR REGRESSION




```
R² Score: 0.7183
RMSE: 262.52
Persentase Error: 105.62% dari rata-rata penjualan
```

- Model Linear Regression cukup baik secara $R^2$, namun RMSE dan % error masih tinggi
- Feature Importance tertinggi adalah variabel **shipping_cost** dan **profit**

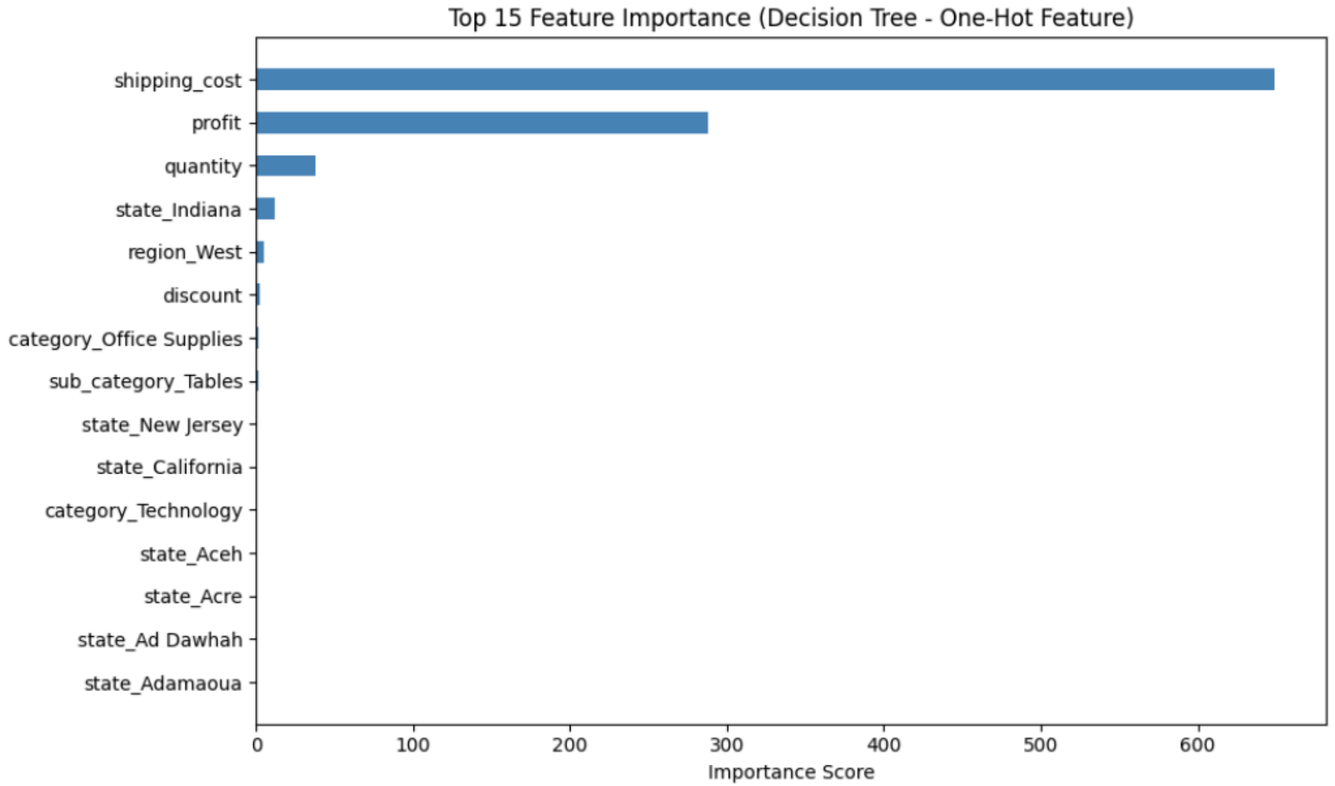
# RANDOM FOREST




```
R² Score: 0.7403
RMSE: 252.07
Persentase Error: 101.41% dari rata-rata penjualan
```

- Model Random Forest cukup baik secara $R^2$, namun RMSE dan % error masih tinggi
- Feature Importance tertinggi adalah variabel **shipping_cost** dan **profit**

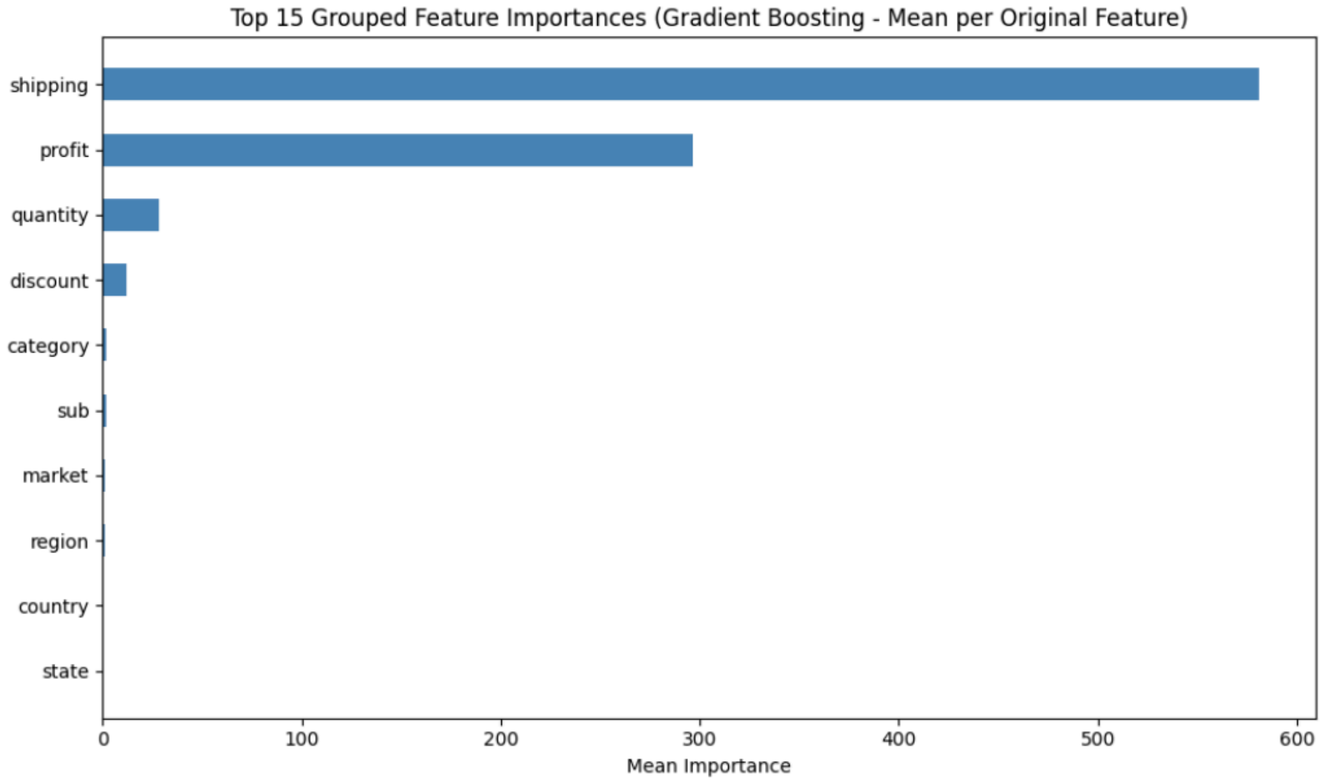
# DECISION TREE




```
R² Score: 0.5872
RMSE: 317.80
Persentase Error: 127.86% dari rata-rata penjualan
```

- Model Decision Tree cukup baik secara $R^2$, namun RMSE dan % error masih tinggi
- Feature Importance tertinggi adalah variabel **shipping_cost** dan **profit**

# GRADIENT BOOSTING REGRESSOR




```
R² Score: 0.8083
RMSE: 216.55
Persentase Error: 87.12% dari rata-rata penjualan
```

- Model Gradient Boosting Regressor cukup baik secara $R^2$, dengan RMSE dan % error lebih kecil dibandingkan model yang lain.
- Feature Importance tertinggi adalah variabel **shipping_cost** dan **profit**

# PERBANDINGAN HASIL PEMODELAN

| Model | $R^2$ Score | RMSE | Persentase Error terhadap Rata-rata Penjualan |
|---|---|---|---|
| **Gradient Boosting** | **0.8083** | **216.55** | **87.12%** |
| **XGBoost** | 0.7831 | 230.34 | 92.67% |
| Random Forest | 0.7403 | 252.07 | 101.41% |
| Linear Regression | 0.7183 | 262.52 | 105.62% |
| Decision Tree | 0.5872 | 317.80 | 127.86% |

Gradient Boosting Regressor menunjukkan performa terbaik dengan **nilai $R^2$ tertinggi**, **RMSE terendah**, dan **persentase error terkecil dari rata-rata penjualan**.
Oleh karena itu, Gradient Boosting Regressor direkomendasikan sebagai model utama dalam prediksi data SuperStore.

# PERBANDINGAN DENGAN ANOVA

**ANOVA**

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

**XGBoost**

1. Profit
2. Shipping_Cost
3. Category
4. Quantity
5. Sub_Category
6. Discount
7. Region
8. Market
9. Country
10. State

1. Category
2. Sub_Category
3. Market
4. Region
5. Country
6. State

**ANOVA**

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

**Linear Regression**

1. Shipping_Cost
2. Profit
3. Category
4. Quantity
5. Sub_Category
6. Discount
7. Region
8. Market
9. Country
10. State

1. Category
2. Sub_Category
3. Region
4. Market
5. Country
6. State

# PERBANDINGAN DENGAN ANOVA

**ANOVA**

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

**Random Forest**

1. Shipping Cost
2. Profit
3. Quantity
4. Discount
5. State
6. Sub_Category
7. Market
8. Country
9. Region

1. State
2. Sub_Category
3. Market
4. Country
5. Region

**ANOVA**

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

**Decision Tree**

1. Shipping_Cost
2. Profit
3. Quantity
4. State
5. Region
6. Discount
7. Category
8. State
9. Sub_Category

1. State
2. Region
3. Discount
4. Category
5. State
6. Sub_Category

# PERBANDINGAN DENGAN ANOVA

- Hasil Feature Importance terbaik menggunakan Gradient Boosting Regressor
- Lalu dilakukan perbandingan untuk melihat hasil Feature Importance ANOVA dengan hasil Feature Importance Machine Learning.

## ANOVA

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

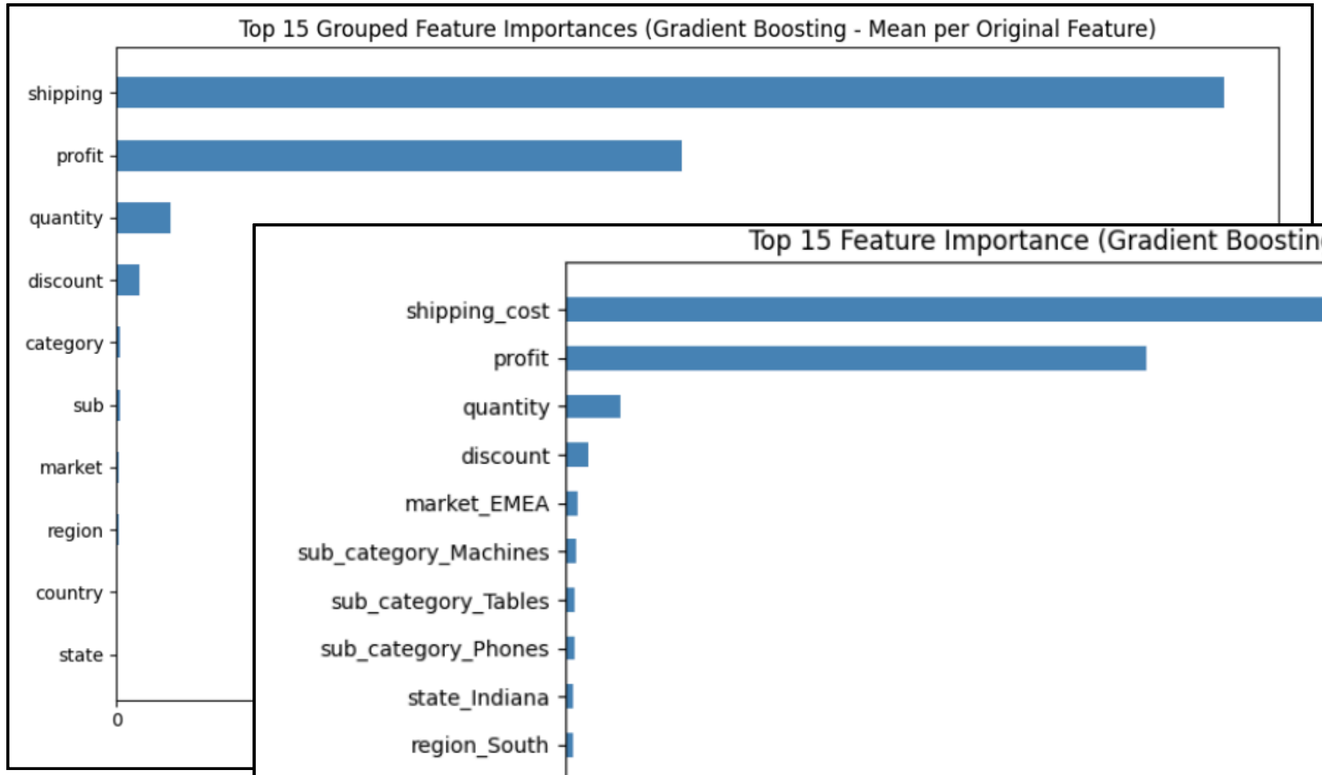
## Gradient Boosting Regressor

1. Shipping_Cost
2. Profit
3. Quantity
4. Discount
5. Category
6. Sub_Category
7. Market
8. Region
9. Country
10. State

1. Category
2. Sub_Category
3. Market
4. Region
5. Country
6. State

# TOP 15 FEATURE IMPORTANCE

## Dari Gradient Boosting







Isi dari Feature Importance yang dimodelkan oleh

**Gradient Boosting Regressor**

# KESIMPULAN

**ANOVA**

1. Sub_Category
2. Category
3. Country
4. Market
5. Region
6. State

**Gradient Boosting Regressor**

1. Shipping_Cost
2. Profit
3. Quantity
4. Discount
5. Category
6. Sub_Category
7. Market
8. Region
9. Country
10. State

1. Category
2. Sub_Category
3. Market
4. Region
5. Country
6. State

- Gradient Boosting Regressor adalah model terbaik untuk memprediksi penjualan karena menghasilkan akurasi tertinggi dan error terkecil dibandingkan model lainnya.
- Variabel numerik seperti Shipping Cost, Profit, Quantity, dan Discount terbukti memiliki pengaruh paling besar terhadap Sales, berdasarkan hasil feature importance dari model.
- Sementara itu, variabel kategorikal seperti Category, Sub_Category, Market, Region, Country, dan State dinilai signifikan secara statistik menurut hasil One-Way ANOVA.

# SARAN

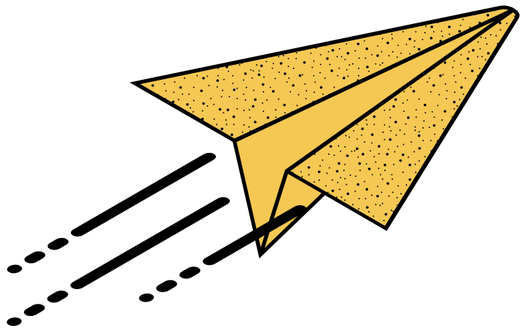
- Mengoptimalkan Shipping Cost, karena Shipping Cost dan Profit memiliki dampak signifikan terhadap Sales.

→ Cth : Mengurangi beban **Shipping Cost** misalnya dengan menetapkan minimum order agar pengiriman lebih efisien.

- Fokus pada pengelolaan stok dan strategi Discount, karena Quantity dan Discount juga berpengaruh signifikan terhadap hasil Sales.

→ Cth : Mengoptimalkan pengadaan **Product** dilihat dari kuantitas penjualan tertinggi per **Sub-Kategori** nya. Seperti **Sub_Category_Machine**, **Sub_Category_Tables**, dan **Sub_Category_Phones**

Nabila Karin | 71478



# THANK YOU!