

# Prompt Log

## 1. Perbaikan Pipeline Preprocessing

Pipeline preprocessing direstrukturisasi agar mengikuti praktik machine learning yang benar, terutama untuk mencegah terjadinya data leakage. Perubahan utama yang dilakukan meliputi tiga aspek berikut.

- Split data dilakukan sebelum proses target encoding.
- Scaling hanya diterapkan pada data training, bukan keseluruhan dataset.
- Penanganan missing value dibuat secara eksplisit dan terstruktur.

### Prompt yang Digunakan

*"Perbaiki preprocessing agar tidak terjadi data leakage pada encoding dan scaling."*

### Hasil

Pipeline yang dihasilkan mengikuti urutan berikut:

```
Data --> Split --> Target Encoding --> Scaling --> Training Model
```

## 2. Implementasi Target Encoding yang Aman

Kode awal melakukan encoding langsung pada seluruh dataset sebelum proses split. Hal ini menyebabkan informasi dari kolom target bocor ke dalam data test, sehingga hasil evaluasi menjadi tidak valid secara metodologis.

### Prompt yang Digunakan

*"Bagaimana cara melakukan target encoding tanpa data leakage?"*

### Implementasi

```
train_temp = X_train.copy()
train_temp["target"] = y_train
brand_mean = train_temp.groupby("device_brand")["target"].mean()
```

## 3. Standarisasi Evaluasi Model

Menambahkan Evaluasi sesuai persyaratan tugas

### Prompt yang Digunakan

*"Buatkan fungsi evaluasi model regresi dengan semua metric sekaligus."*

## Implementasi

```
def evaluate_model(y_true, y_pred, model_name):
```

## Metrik yang Digunakan

- R2 Score
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

## 4. Transformasi Target (Log Transformation)

### Prompt yang Digunakan

*"Bagaimana cara meningkatkan performa regresi?"*

## Implementasi

```
y_train_log = np.log1p(y_train)
pred = np.expml(pred_log)
```

## 5. Feature Engineering: Polynomial Features

Menambahkan beberapa model alternatif untuk melakukan pengujian model terbaik.

### Prompt yang Digunakan

*"Tambahkan polynomial regression sebagai improvement dari linear regression."*

### Model yang Ditambahkan

- Polynomial Linear Regression
- Ridge Regression
- Polynomial Ridge Regression

## 6. K-Fold Cross Validation

Evaluasi model hanya bergantung pada satu train-test split tunggal, sehingga hasil evaluasi rentan terhadap bias akibat distribusi data pada split tertentu.

### Prompt yang Digunakan

*"Tambahkan K-fold validation untuk evaluasi regresi multi-metric."*

## Implementasi

```
|KFold(n_splits=5)
```

## **7. Stress Testing Model**

### **Prompt yang Digunakan**

*"Bagaimana cara membuat stress test untuk model regresi agar menguji robustness?"*

### **Tujuan**

Menilai stabilitas dan ketahanan model pada kondisi data yang mencerminkan situasi dunia nyata, di mana kualitas dan distribusi data tidak selalu ideal.