



# **PREDIKSI KONSUMSI LISTRIK HARIAN BERBASIS CUACA MENGGUNAKAN XGBOOST**

# LATAR BELAKANG

Konsumsi Listrik = Dinamis & Tergantung Cuaca

- Dipengaruhi suhu, kelembapan, angin, sinar matahari
- Fluktuasi cuaca berdampak pada beban listrik harian

Kebutuhan Prediksi yang Akurat

- Untuk efisiensi distribusi
- Menghindari beban puncak & potensi pemadaman

# TUJAN PROYEK

- 01** Memprediksi konsumsi listrik harian (GWh)
- 02** Menjadi dasar tarif listrik dinamis atau strategi subsidi
- 03** Membantu manajemen permintaan musiman (panas, hujan, hari libur)
- 04** Menentukan kapan perlu naikkan kapasitas

# DATASET

Dataset bersifat publik diambil dari kaggle dengan data target electricity\_consumption dalam satuan GWh

Kolom	Tipe	Deskripsi
ID	string	ID unik <>cluster_id><>YYYY-MM-DD>>
date	string	Tanggal (YYYY-MM-DD)
cluster_id	string	ID cluster ( cluster_1 hingga cluster_4 )
electricity_consumption	float	Konsumsi listrik harian (GWh)
temperature_2m_max	float	Suhu maksimum 2m (°C)
temperature_2m_min	float	Suhu minimum 2m (°C)
apparent_temperature_max	float	Suhu terasa maksimum (°C)
apparent_temperature_min	float	Suhu terasa minimum (°C)
sunshine_duration	float	Durasi sinar matahari (jam)
daylight_duration	float	Durasi siang hari (detik)
wind_speed_10m_max	float	Kecepatan angin maksimum 10m (m/s)
wind_gusts_10m_max	float	Kecepatan embusan maksimum 10m (m/s)
wind_direction_10m_dominant	int	Arah angin dominan 10m (° dari utara)
shortwave_radiation_sum	float	Radiasi gelombang pendek (MJ/m <sup>2</sup> )
et0_fao_evapotranspiration	float	Evapotranspirasi FAO (mm)

# **EXPLORATORY DATA ANALYTICS (EDA)**

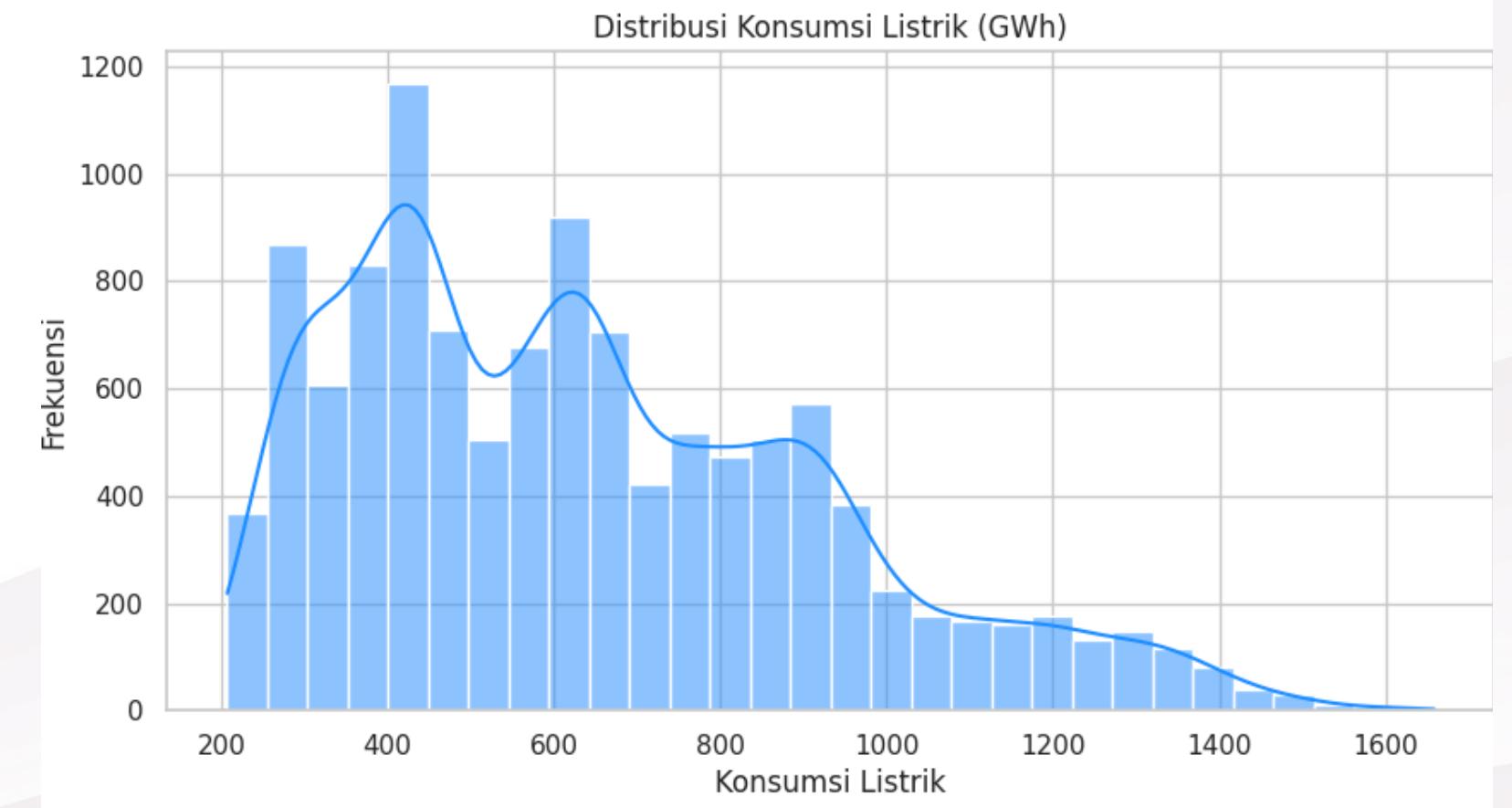
# CEK INFORMASI DATA DAN MISSING VALUE

data tersebut berjumlah 11688 baris dengan total 15 column, tidak terdapat missing value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11688 entries, 0 to 11687
Data columns (total 15 columns):
 #   Column           Non-Null Count Dtype  
 ---  -- 
 0   ID               11688 non-null  object  
 1   date              11688 non-null  object  
 2   cluster_id        11688 non-null  object  
 3   electricity_consumption 11688 non-null  float64 
 4   temperature_2m_max 11688 non-null  float64 
 5   temperature_2m_min 11688 non-null  float64 
 6   apparent_temperature_max 11688 non-null  float64 
 7   apparent_temperature_min 11688 non-null  float64 
 8   sunshine_duration 11688 non-null  float64 
 9   daylight_duration 11688 non-null  float64 
 10  wind_speed_10m_max 11688 non-null  float64 
 11  wind_gusts_10m_max 11688 non-null  float64 
 12  wind_direction_10m_dominant 11688 non-null  float64 
 13  shortwave_radiation_sum 11688 non-null  float64 
 14  et0_fao_evapotranspiration 11688 non-null  float64 
dtypes: float64(12), object(3)
memory usage: 1.3+ MB
```

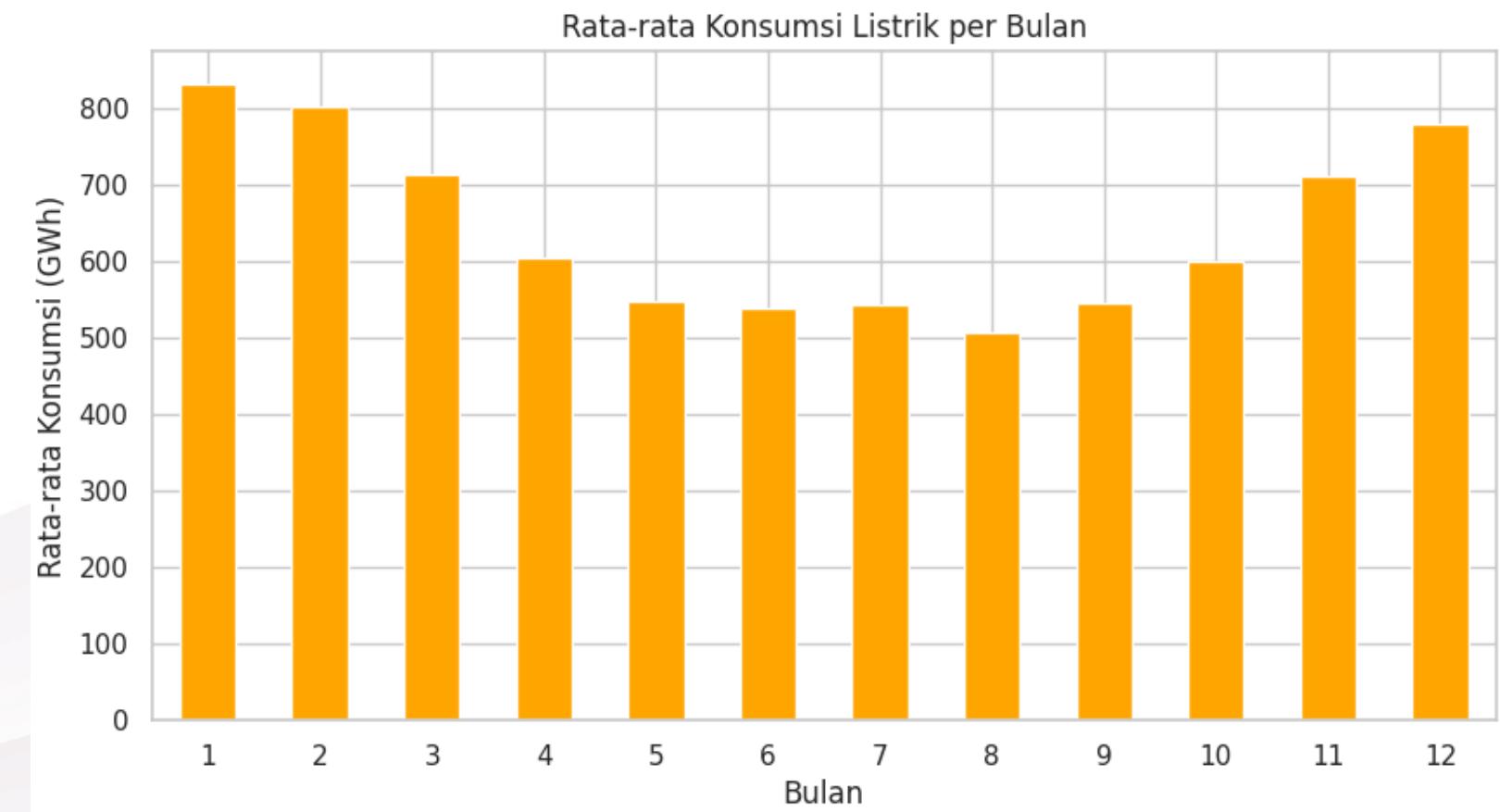
# DISTRIBUSI KONSUMSI LISTRIK

konsumsi listrik lebih banyak terdistribusi pada rentang 200 - 1000 Gwh, ini berarti sebagian besar konsumen merupakan kelas menengah berasal dari kondisi geografis yang berbeda beda, bisa dari pedesaan maupun perkotaan



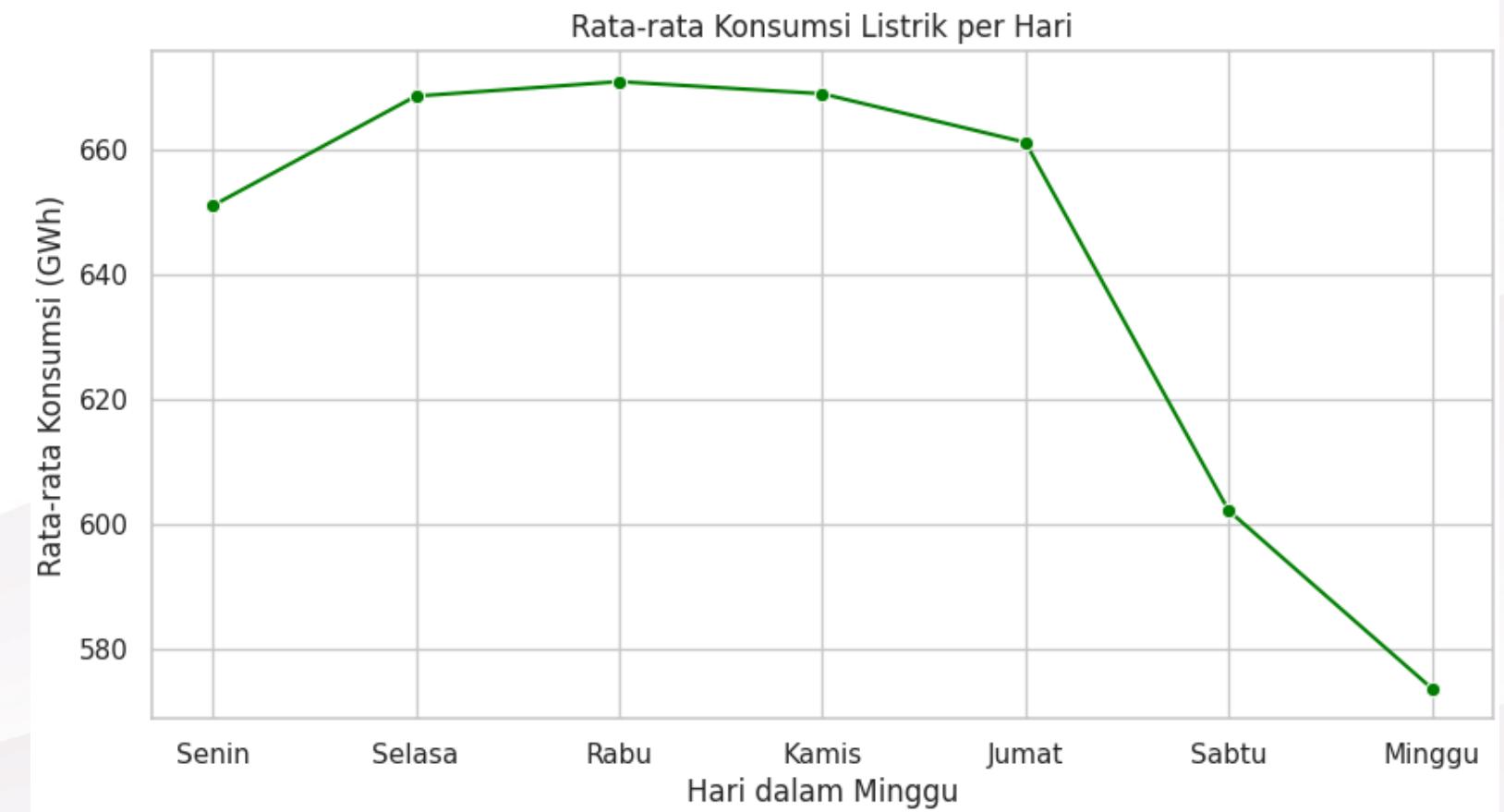
# RATA RATA KONSUMSI PERBULAN

konsumsi meningkat pada awal tahun hingga menurun sampai pertengahan bulan kemudian naik lagi pada akhir tahun, ini dipengaruhi oleh beberapa faktor seperti libur harian yang banyak terjadi pada awal dan akhir tahun, tidak menutup kemungkinan juga karena regulasi dan lain lain



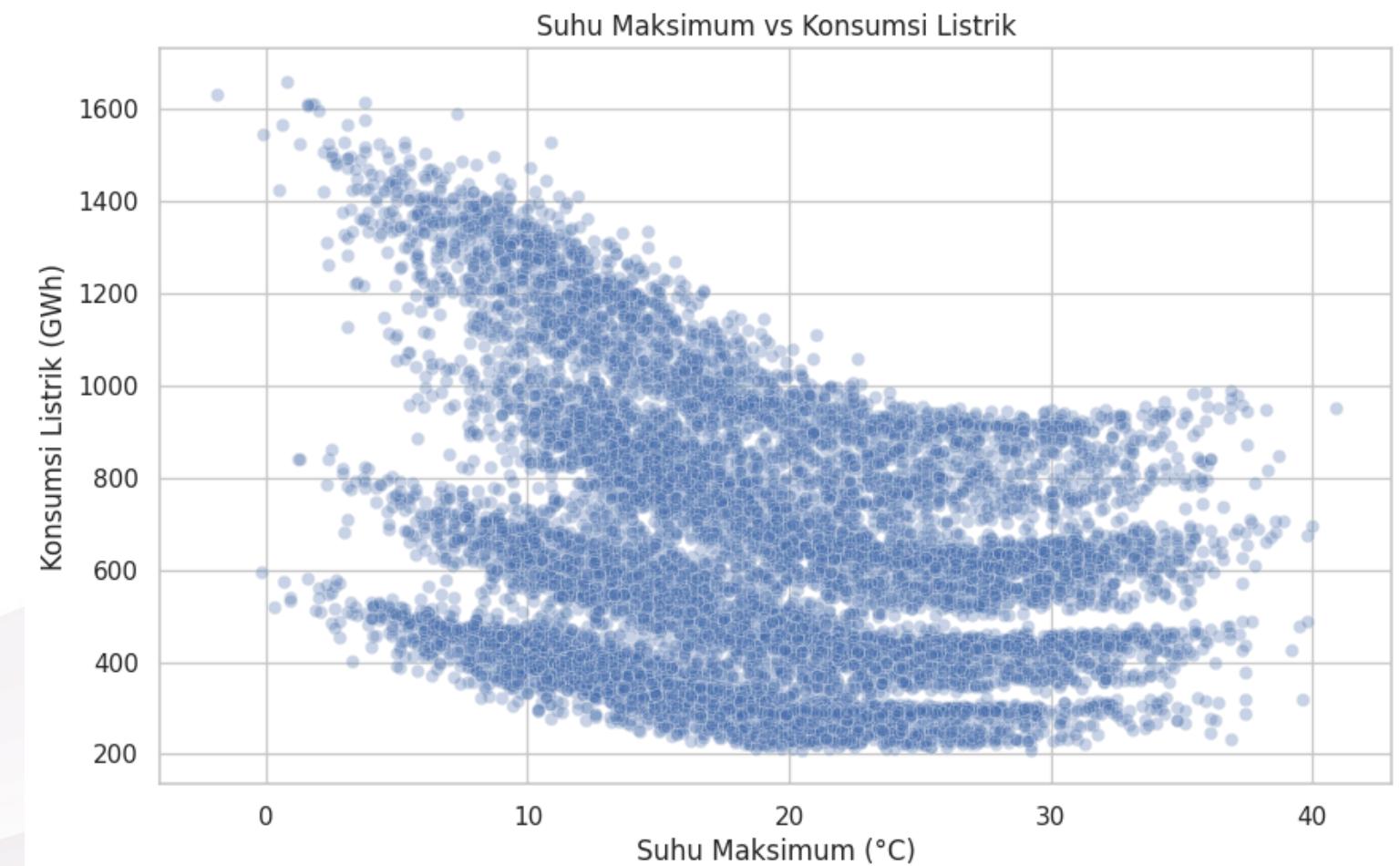
# RATA RATA KONSUMSI PERHARI

pada akhir pekan terutama pada sabtu sampai minggu konsumsi listrik cenderung menurun, bisa jadi pada akhir pekan orang-orang lebih memilih istirahat atau liburan ke luar dari pada kegiatan yang dapat memengaruhi konsumsi listrik



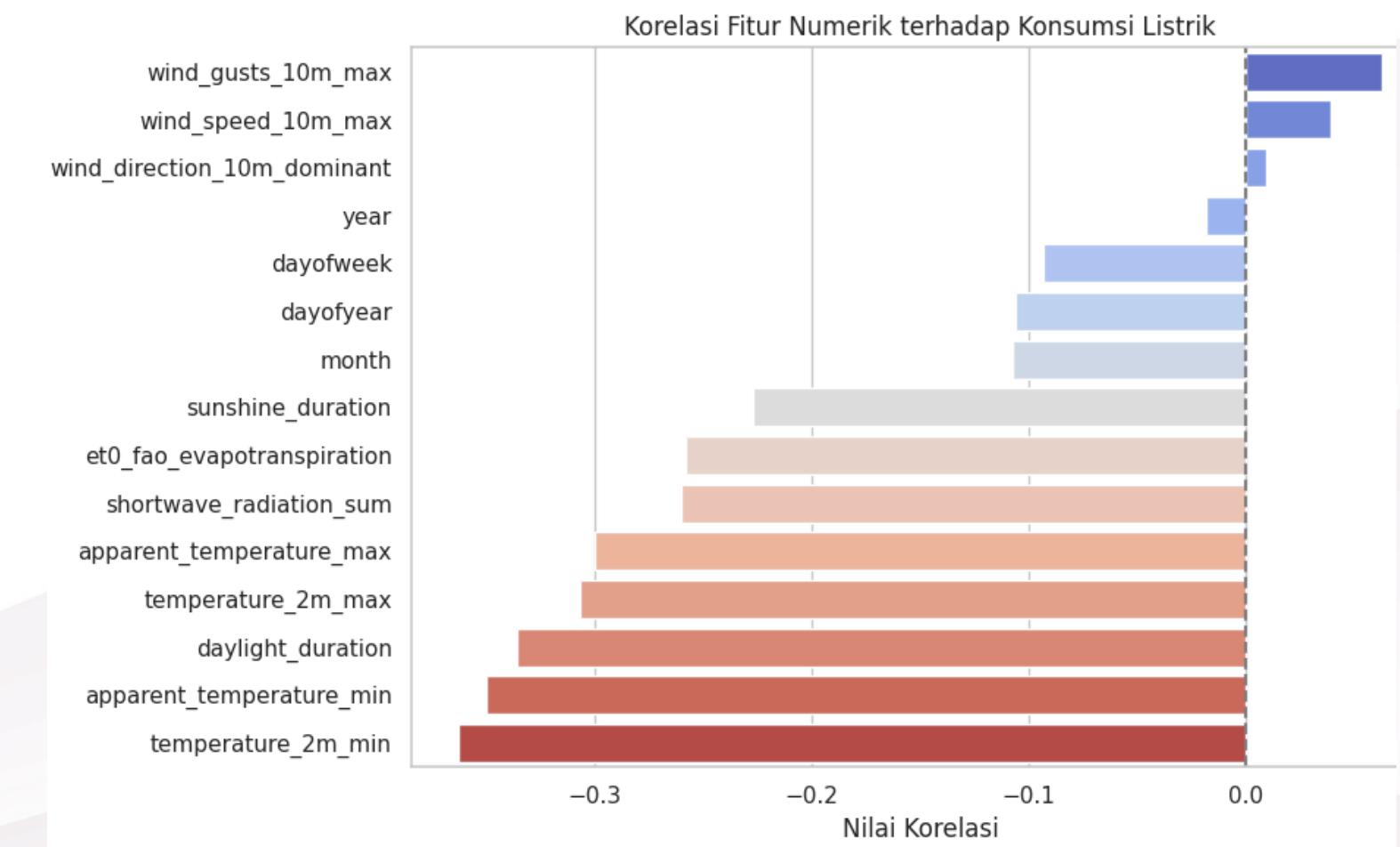
# KORELASI SUHU VS KONSUMSI

korelasi antara suhu dan konsumsi listrik cenderung negatif yang mana semakin kecil suhu, konsumsi listrik cenderung besar, mungkin saja orang-orang lebih suka berada dirumah saat suhu dingin, serta bisa dilihat pola pada data terdapat cluster yang saling terpisah.



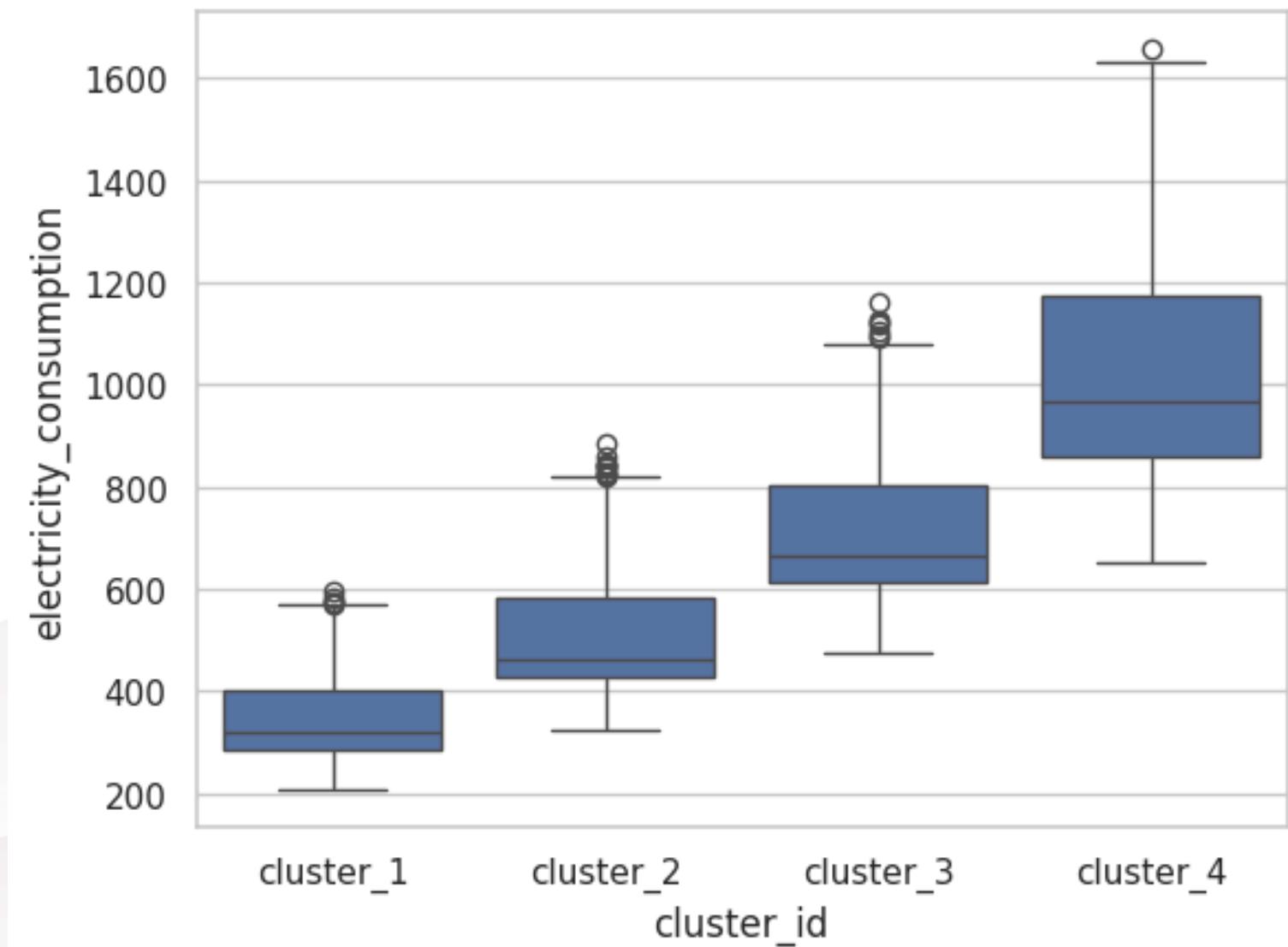
# KORELASI FITUR NUMERIK VS KONSUMSI

pengaruh angin sangat berpengaruh pada konsumsi listrik, untuk semua fitur angin berkorelasi positif terhadap konsumsi, selain itu korelasi negatif



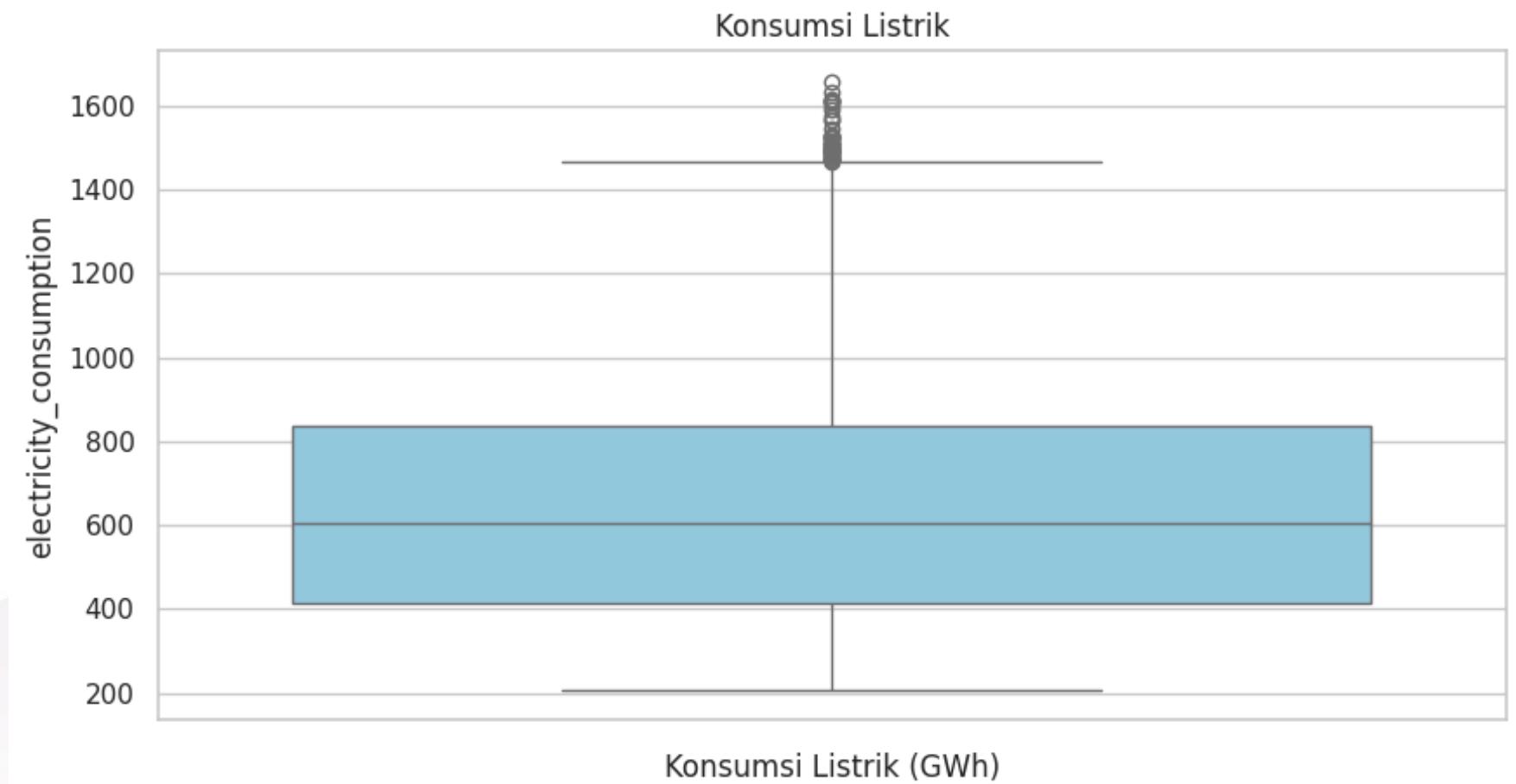
# KONSUMSI LISTRIK PER CLUSTER

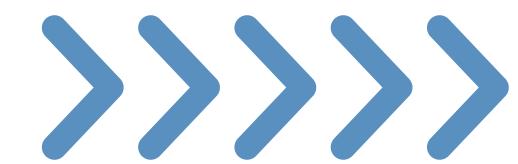
seperti yang dilihat pada scatterplot tadi bahwa setiap cluster menghasilkan konsumsi listrik yang berbeda beda, mungkin saja cluster 1 berasal dari daerah pedesaan atau wilayah yang minim distribusi listrik, sedangkan cluster 4 berasal dari perumahan elite



# CEK OUTLIER PADA KOLOM TARGET

ada beberapa data yang tidak normal, konsumsi listrik bisa mencapai 1400 keatas, tetapi mungkin saja itu memang berasal dari wilayah yang lebih banyak membutuhkan listrik, apakah perlu hapus outlier, kita lihat nanti





# PREPROCESSING

# HAPUS OUTLIER

Setelah mempertimbangkan banyak hal, kami memutuskan untuk menghapus outlier untuk kolom target dan beberapa kolom numerik, hal ini bertujuan untuk performa model, dan data outlier sangat berpengaruh terutama regresi

Jumlah data sebelum: 11642

Jumlah data setelah bersih dari outlier: 11341

Total data yang dihapus: 301

# FEATURE ENGINEERING

## Sebelum

ID	temperature_2m_min	wind_gusts_10m_max
date	apparent_temperature_max	wind_direction_10m_dominant
cluster_id	apparent_temperature_min	shortwave_radiation_sum
electricity_consumption	sunshine_duration	et0_fao_evapotranspiration
temperature_2m_max	daylight_duration	
wind_speed_10m_max		

## Sesudah

electricity_consumption	apparent_temperature_min	wind_gusts_10m_max	is_weekend
temperature_2m_max	sunshine_duration	wind_direction_10m_dominant	avg_temperature
temperature_2m_min	daylight_duration	shortwave_radiation_sum	heat_index
apparent_temperature_max	wind_speed_10m_max	et0_fao_evapotranspiration	cluster_id_encoded
month	dayofweek	dayofyear	

# MODELLING

# PEMBAGIAN TRAIN & TEST

kami menggunakan 80:20, yaitu data train sebesar 80% dan data test 20%, dengan target electricity\_consumption

```
from sklearn.model_selection import train_test_split

# Target
y = df['electricity_consumption']

# Fitur
X = df.drop(columns=['electricity_consumption'])

# Split 80-20
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

# SELEKSI MODEL

kami menggunakan skor RMSE sebagai acuan utama dan membandingkan beberapa model seperti linear regression, random forest, XGboost, Gradient boosting dan KNN, hasil terbaik adalah XGboost

Linear Regression RMSE: 66.0123  
Random Forest RMSE: 31.9767  
XGBoost RMSE: 29.9052  
Gradient Boosting RMSE: 33.6366  
KNN RMSE: 282.3638

# TUNING (OPTUNA)

Untuk mencari performa terbaik, kami melakukan fine tuning menggunakan optuna, dengan beberapa parameter, kemudian hasil yang didapatkan juga tidak beda jauh yaitu sebesar 28.3

# TRAIN ULANG MODEL

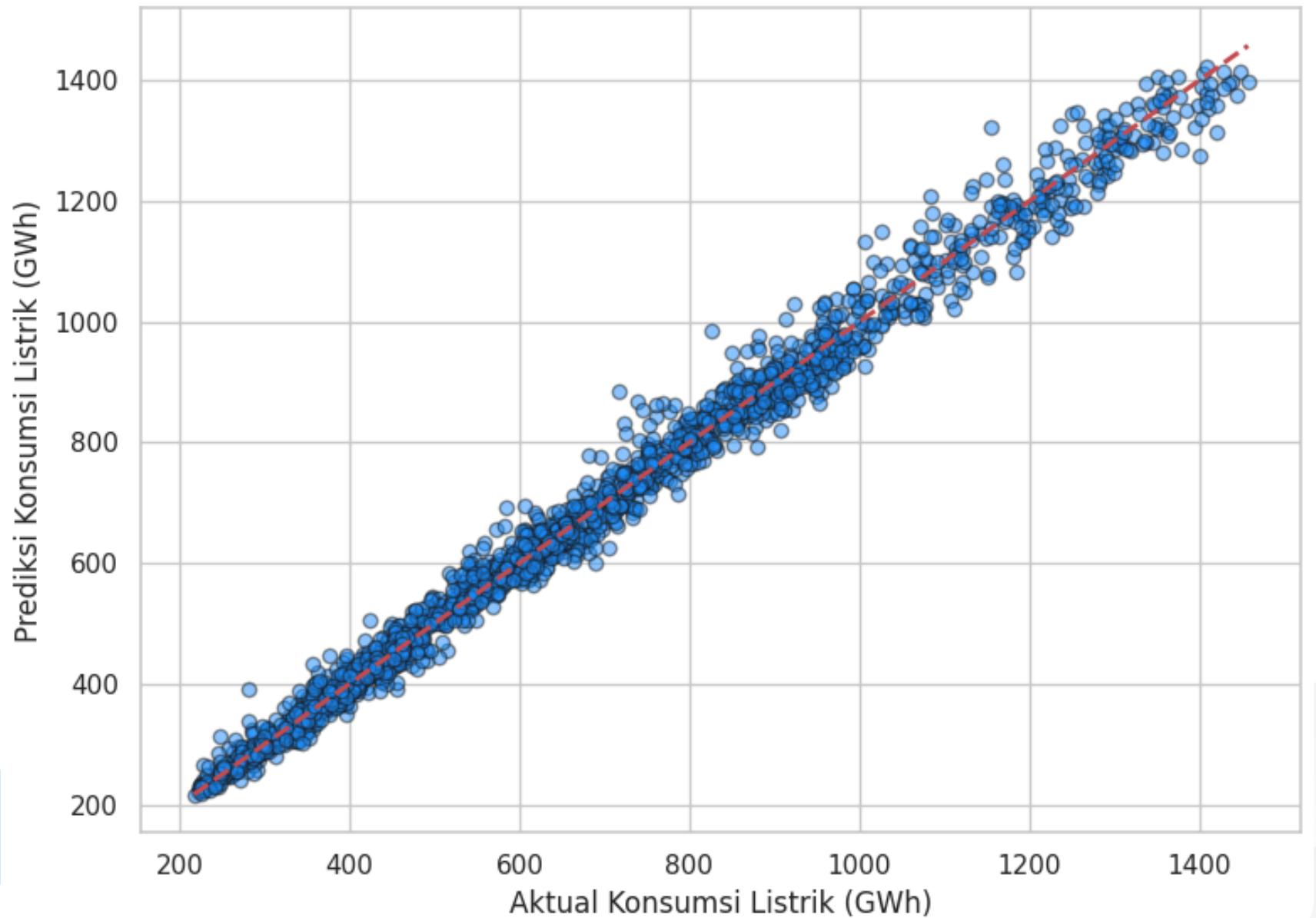
```
XGBRegressor(base_score=None, booster=None, callbacks=None,  
            colsample_bylevel=None, colsample_bynode=None,  
            colsample_bytree=0.6, device=None, early_stopping_rounds=None,  
            enable_categorical=False, eval_metric=None, feature_types=None,  
            gamma=None, grow_policy=None, importance_type=None,  
            interaction_constraints=None, learning_rate=0.1, max_bin=None,  
            max_cat_threshold=None, max_cat_to_onehot=None,  
            max_delta_step=None, max_depth=5, max_leaves=None,  
            min_child_weight=None, missing=nan, monotone_constraints=None,  
            multi_strategy=None, n_estimators=498, n_jobs=-1,  
            num_parallel_tree=None, random_state=42, ...)
```

# EVALUASI

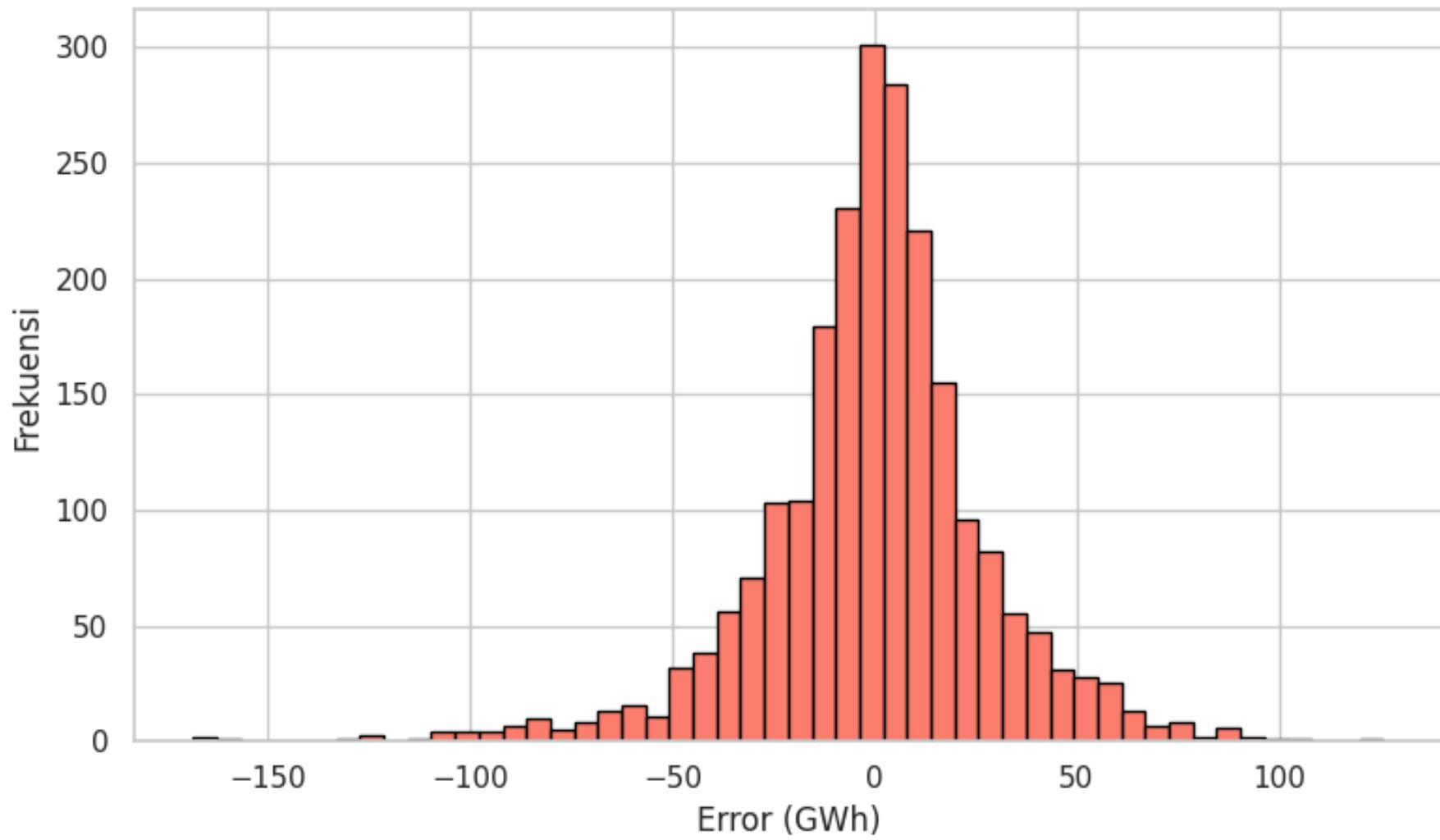
# CEK DENGAN METRIK LAIN

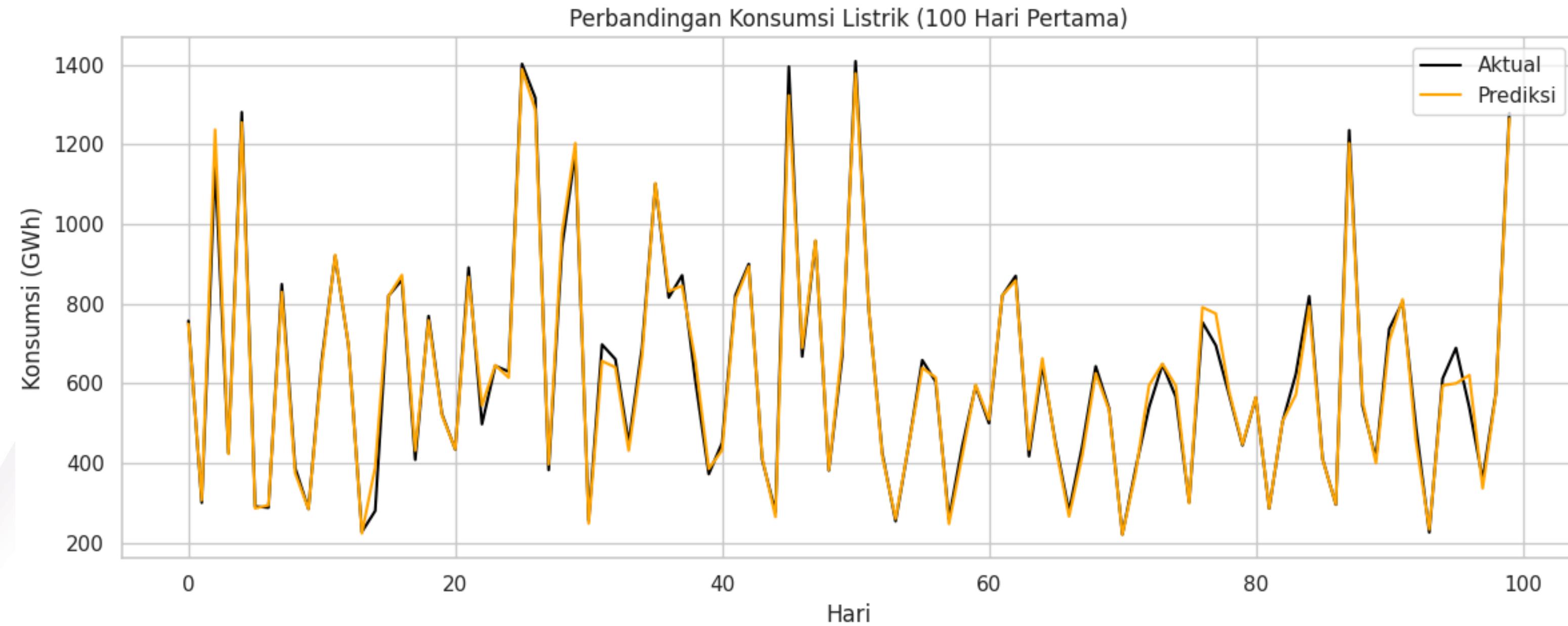
RMSE : 28.3902  
MAE : 19.8276  
 $R^2$  : 0.9903

Prediksi vs Aktual



Distribusi Error (Aktual - Prediksi)

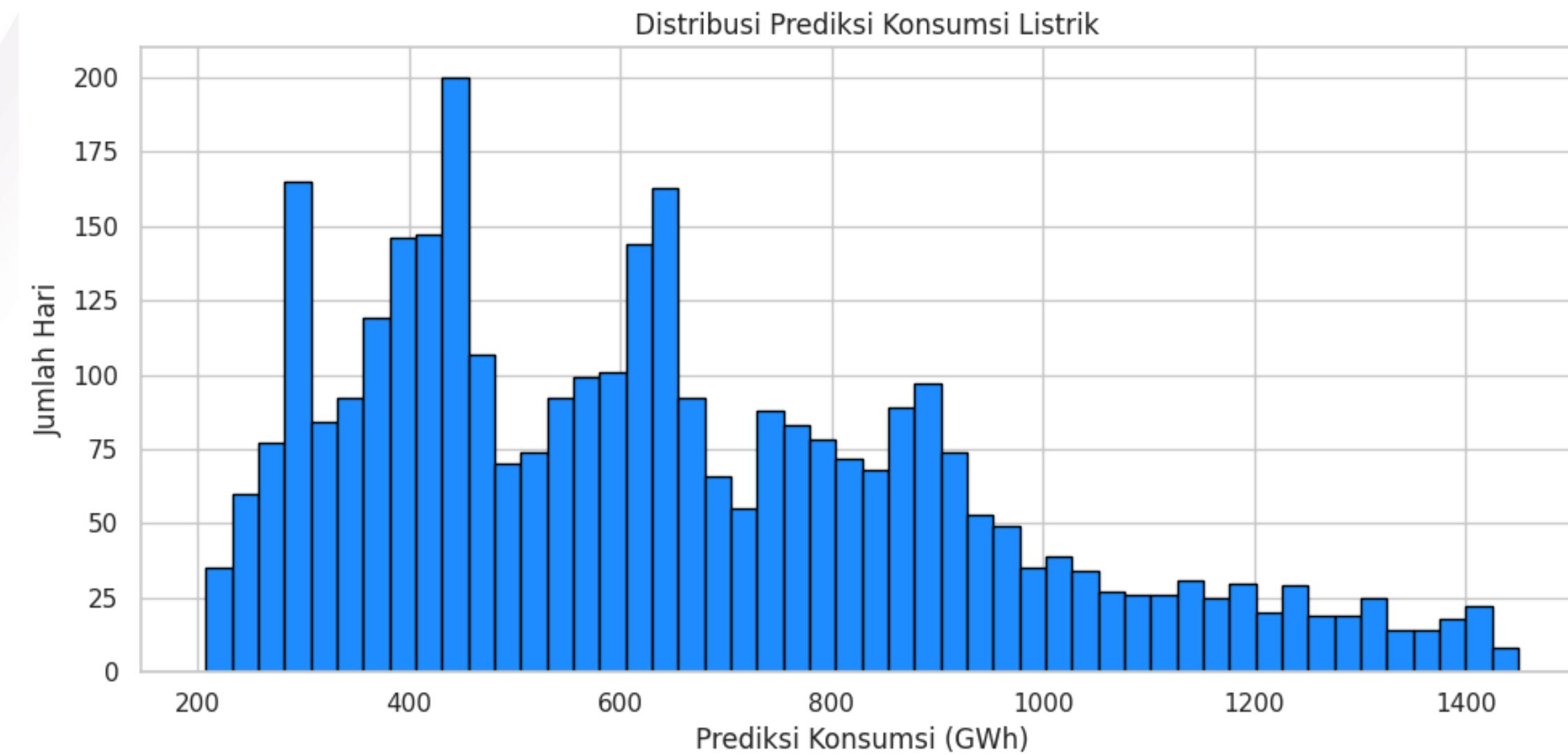




# PREDIKSI AKTUAL

dalam dataset keggle tadi, selain data train, terdapat data test yang mempunyai fitur sama tetapi tanpa label, untuk keperluan test data tersebut harus di preprocessing seperti data train agar sama fiturnya

# DISTRIBUSI PREDIKSI KONSUMSI LISTRIK



# KESIMPULAN

Prediksi konsumsi listrik harian berbasis cuaca berhasil dilakukan dengan model XGBoost, yang menunjukkan performa terbaik dibanding model lain. Data dari Kaggle dianalisis dan dibersihkan, termasuk outlier removal dan feature engineering. Hasil menunjukkan bahwa faktor cuaca seperti suhu dan angin berpengaruh signifikan terhadap konsumsi listrik. Model yang dibangun mampu memberikan prediksi yang stabil dan dapat digunakan untuk mendukung manajemen distribusi dan strategi energi yang lebih efisien.

# SOURCE

DATASET: <https://www.kaggle.com/competitions/seleksi-dsa-compfest-17/data>

CODE:

<https://colab.research.google.com/drive/1UP7TWX09htnjbmZfALGKgF-3pPorRBPp?usp=sharing>