SATRIA DATA ITB 2022

Prediksi Harga Berdasarkan Data Penjualan Mobil Bekas dengan Model Berbasis *Tree*

Biomateformatika

Hanif Muhammad Zhafran (13521157)
Ang, Johan Nicholas (18321003)
Habiburrohman (10121089)

Latar belakang

Mobil sebagai moda transportasi utama

Masyarakat mencari alternatif dengan membeli mobil bekas

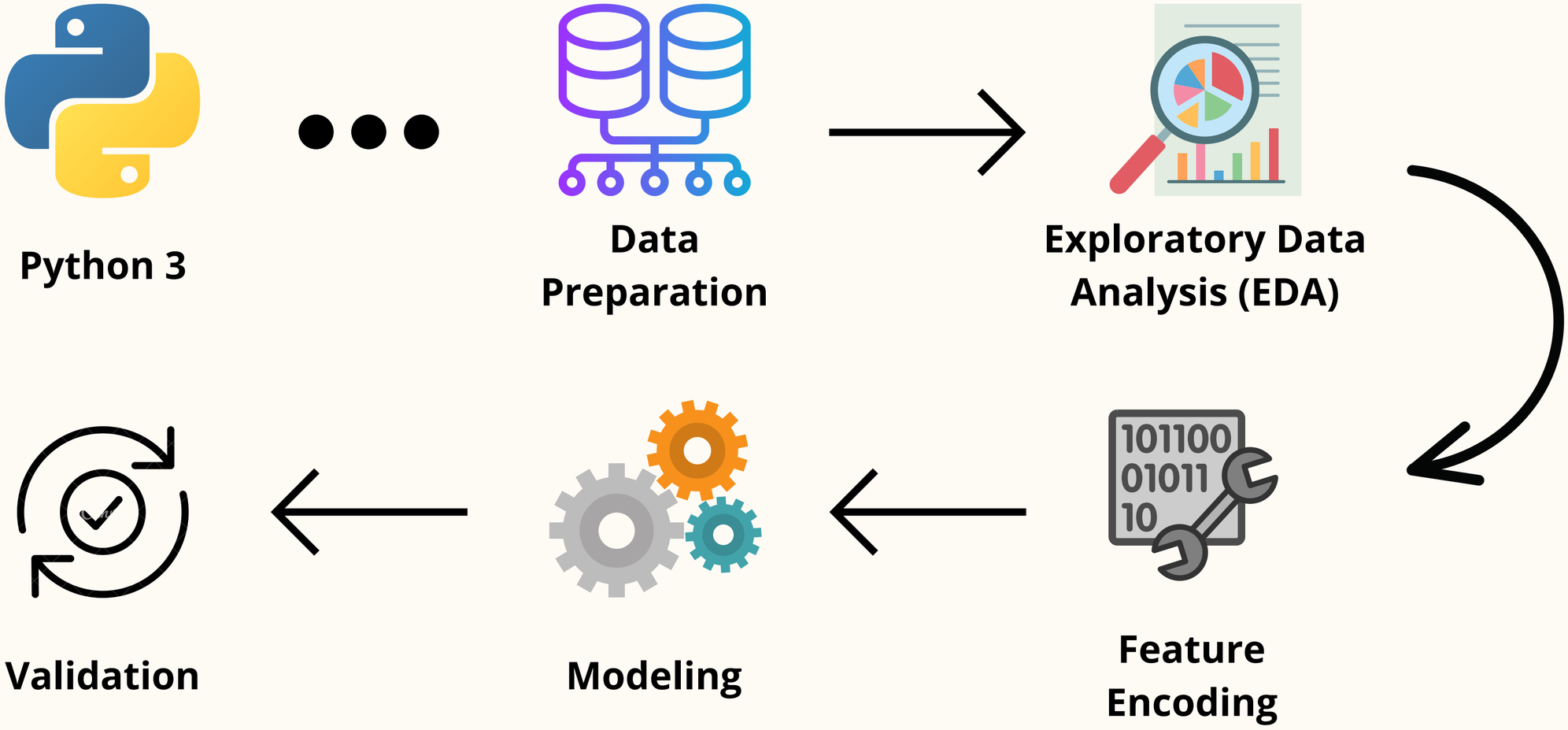
Banyak showroom mobil bekas

Tidak mudah untuk menentukan harga jual yang tepat

Analisis dan prediksi big data

Membantu pengusaha merekomendasikan harga jual yang tepat

Metodologi


Python 3
Data Preparation
Exploratory Data Analysis (EDA)
Feature Encoding
Modeling
Validation

Data Preparation

Dimensi

- Train data : 2000 baris dan 7 kolom
- Test data: 397 baris dan 7 kolom
- Fitur 'id' sebagai primary key dan fitur 'harga' sebagai target

Missing value dan duplicate data

- Ditemukan **missing value** pada fitur 'mesin'
- Ditemukan **dua jenis data duplikat** sebanyak 518 record pada train data

Feature engineering

- Dihasilkan **tiga fitur baru** dari fitur 'listing'
- **Penyeragaman** fitur 'odo' dan 'mesin'

Overview train data

column	dtypes	nunique	missing %
id	int64	2000	0.0
listing	object	800	0.0
bahan_bakar	object	5	0.0
odo	object	312	0.0
transmisi	object	4	0.0
penjual	object	6	0.0
mesin	object	60	3.0
harga	float64	473	0.0

mean imputation

Overview test data

column	dtypes	nunique	missing %
id	int64	397	0.0
listing	object	268	0.0
bahan_bakar	object	2	0.0
odo	object	121	0.0
transmisi	object	4	0.0
penjual	object	5	0.0
mesin	object	31	5.0
harga	float64	0	100.0

mean imputation

Duplicate in train data

	id	listing	bahan_bakar	odo	transmisi	penjual	mesin	harga
827	828	chevrolet trax (2016)	Bensin	75.000-80.000 Km	Automatic	Individu	>1.000 - 1.500 cc	16.8
828	829	chevrolet trax (2016)	Bensin	75.000-80.000 Km	Automatic	Individu	>1.000 - 1.500 cc	16.8
895	896	chevrolet trax (2016)	Bensin	75.000-80.000 Km	Automatic	Individu	>1.000 - 1.500 cc	15.9
330	331	daihatsu ayla (2017)	Bensin	50.000-55.000 Km	Automatic	Individu	>1.000 - 1.500 cc	10.8
339	340	daihatsu ayla (2017)	Bensin	50.000-55.000 Km	Automatic	Individu	>1.000 - 1.500 cc	10.8
...	...	...	...	...	...	...	...	...
1589	1590	wuling confero (2017)	Bensin	169449 Km	Manual	Pertama	1485 cc	9.5
994	995	wuling cortez (2018)	Bensin	100759 Km	Otomatis	Pertama	1798 cc	16.9
1107	1108	wuling cortez (2018)	Bensin	100759 Km	Otomatis	Pertama	1798 cc	16.9
1132	1133	wuling cortez (2018)	Bensin	100759 Km	Otomatis	Pertama	1798 cc	16.9
1134	1135	wuling cortez (2018)	Bensin	100759 Km	Otomatis	Pertama	1798 cc	16.9

Before-after feature engineering

	id	listing	bahan_bakar	odo	transmisi	penjual	mesin	harga
0	1	renault koleos (2019)	Bensin	50.000-55.000 Km	Automatic	Diler	>2.000 - 3.000 cc	32.2
1	2	toyota lexus (2014)	Bensin	45.000-50.000 Km	Automatic Triptonic	Diler	>2.000 - 3.000 cc	39.9
2	3	mercedes-benz e230 (1996)	Bensin	135.000-140.000 Km	Automatic	Individu	>2.000 - 3.000 cc	7.9
3	4	toyota yaris (2018)	Bensin	40.000-45.000 Km	Automatic	Individu	>1.500 - 2.000 cc	22.9
4	5	toyota fortuner (2018)	Diesel	67594 Km	Otomatis	Kedua	2393 cc	46.2

	id	produsen	listing	tahun	bahan_bakar	odo_mean	transmisi	penjual	mesin_mean	harga
0	1	renault	koleos	2019	Bensin	52500.0	Automatic	Diler	2500.0	32.2
1	2	toyota	lexus	2014	Bensin	47500.0	Automatic Triptonic	Diler	2500.0	39.9
2	3	mercedes-benz	e230	1996	Bensin	137500.0	Automatic	Individu	2500.0	7.9
3	4	toyota	yaris	2018	Bensin	42500.0	Automatic	Individu	1750.0	22.9
4	5	toyota	fortuner	2018	Diesel	67594.0	Otomatis	Kedua	2393.0	46.2

Delete kelas eksklusif

train

	penjual	count
0	Individu	715
1	Diler	514
2	Pertama	225
3	Kedua	149
4	Keempat	16
5	Ketiga	4

test

	penjual	count
0	Individu	168
1	Diler	142
2	Pertama	41
3	Kedua	38
4	Keempat	8

train

	bahan_bakar	count
0	Bensin	1493
1	Diesel	123
2	Hybrid	5
3	Bensin/LPG	1
4	LPG	1

test

	bahan_bakar	count
0	Bensin	369
1	Diesel	28

Penggabungan value yang sama

train

	transmisi	count
0	Automatic	835
1	Manual	715
2	Otomatis	330
3	Automatic Triptonic	120

test

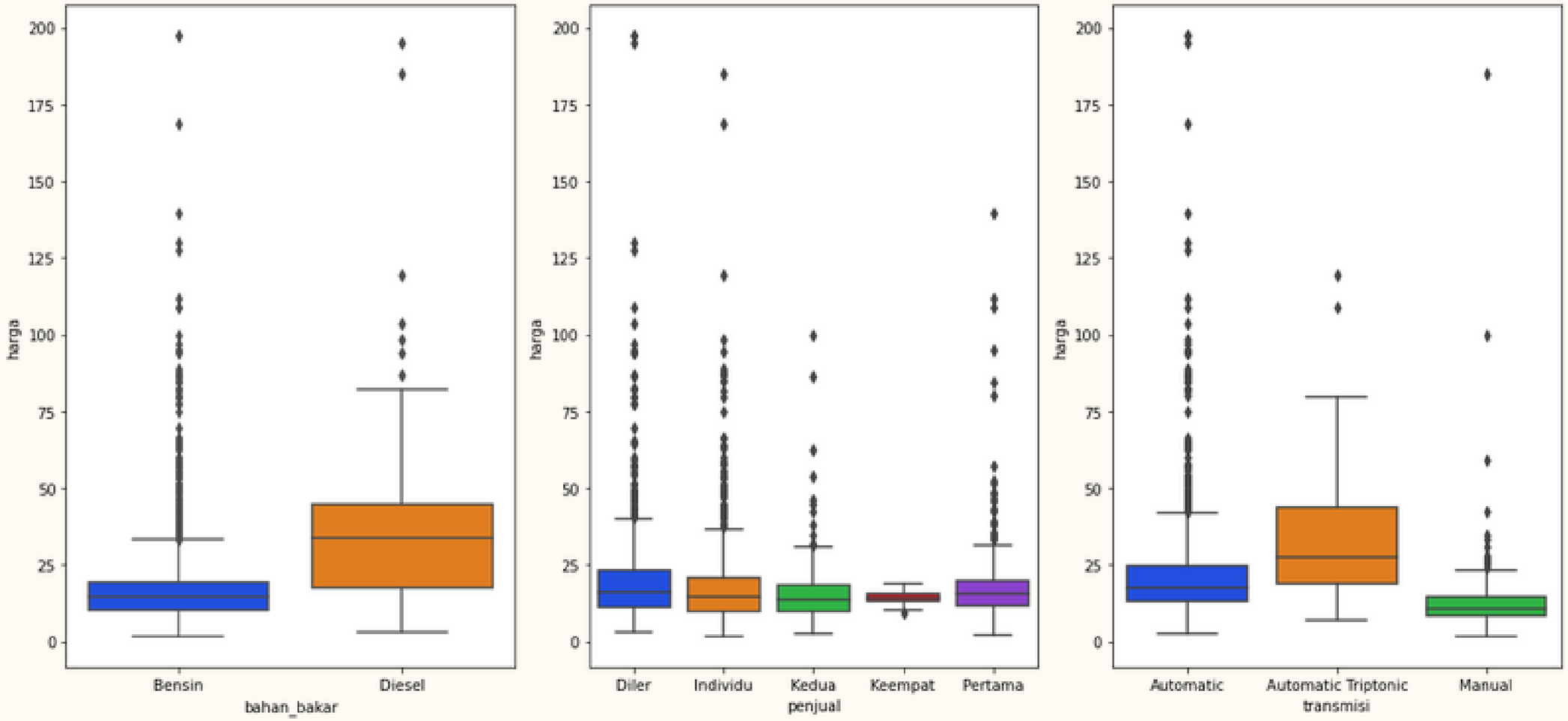
	transmisi	count
0	Automatic	159
1	Manual	153
2	Otomatis	53
3	Automatic Triptonic	32

EDA

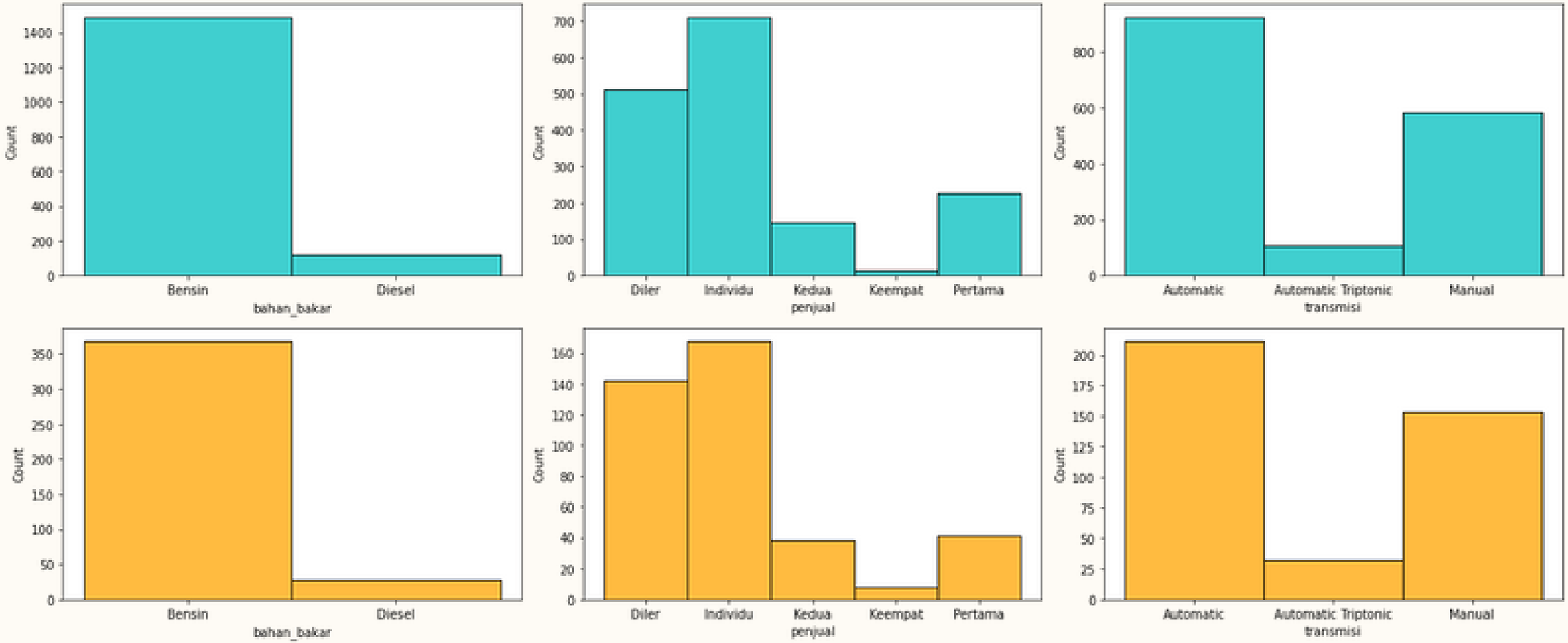
- Distribusi target menurut fitur categorical
- Distribusi fitur categotical dan numerik
- Deteksi outlier fitur numerik
- Korelasi antar fitur numerik dan target

Exploratory Data Analysis

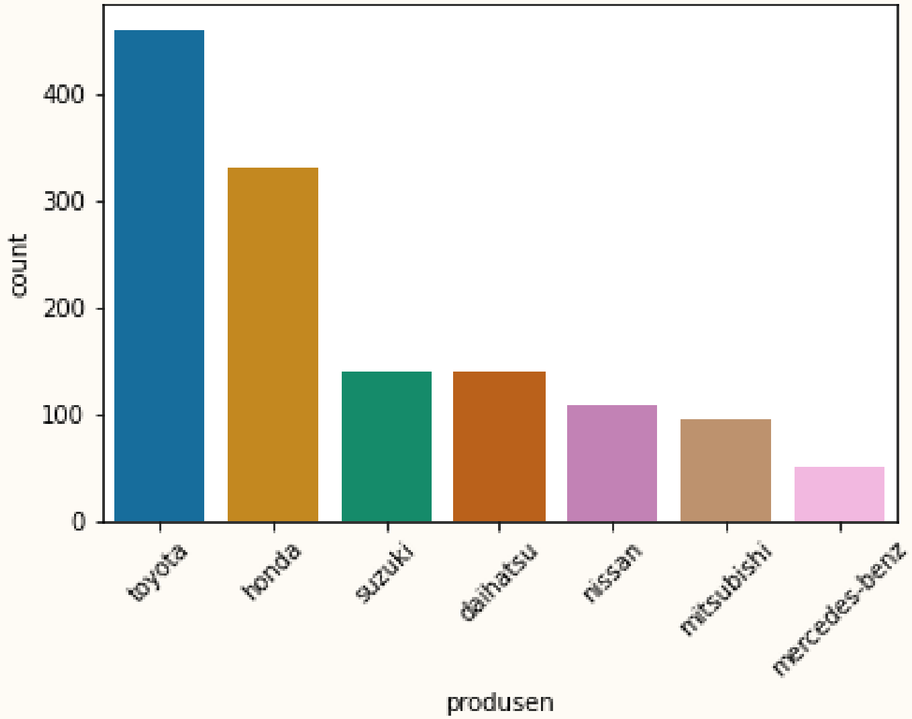
Distribusi target menurut fitur categorical

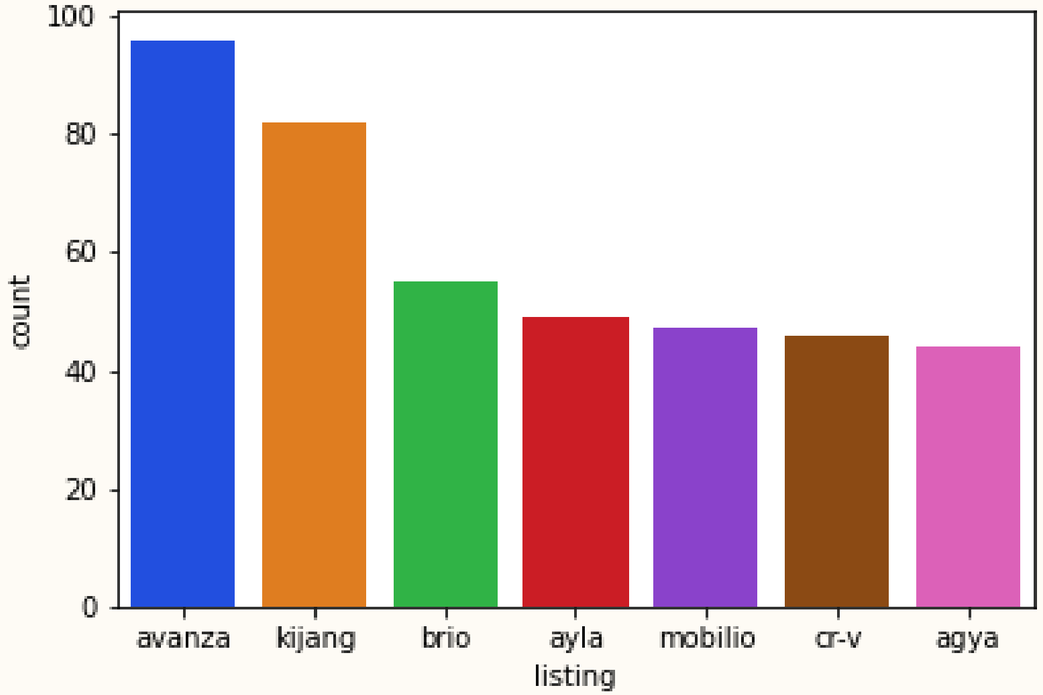


Distribusi fitur categorical

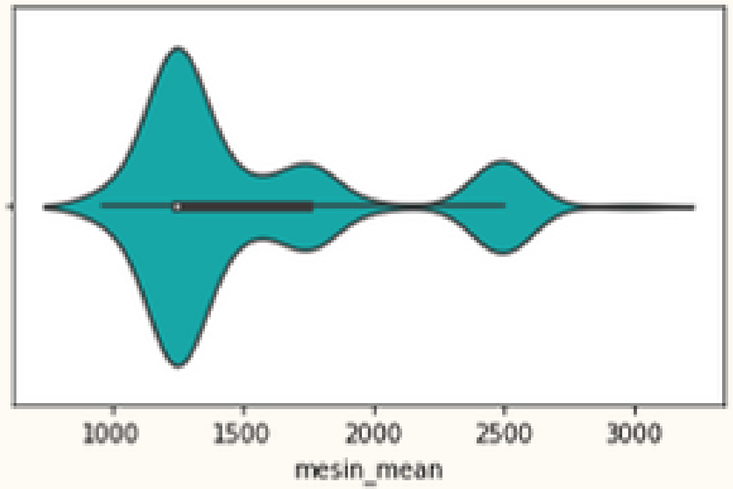


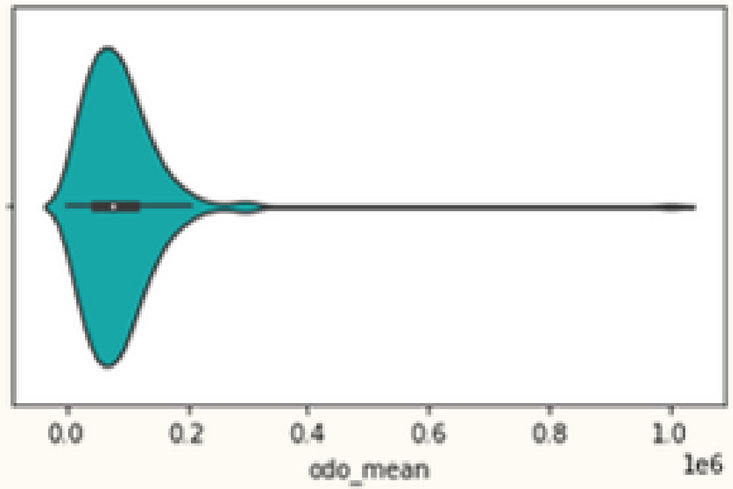
Distribusi fitur categorical (2)

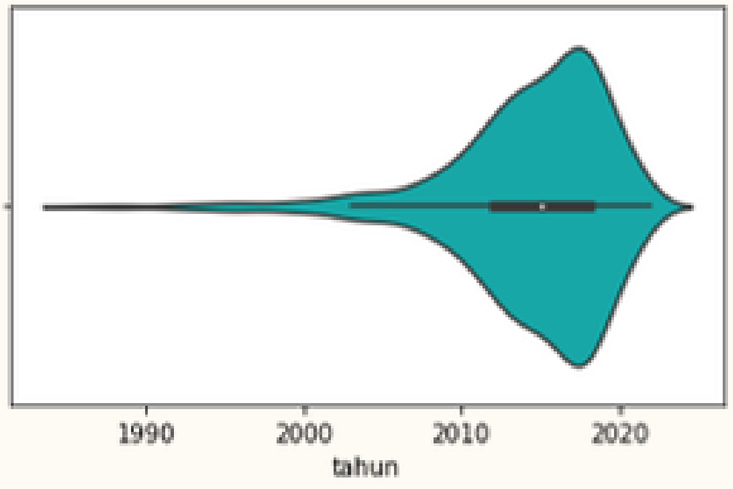


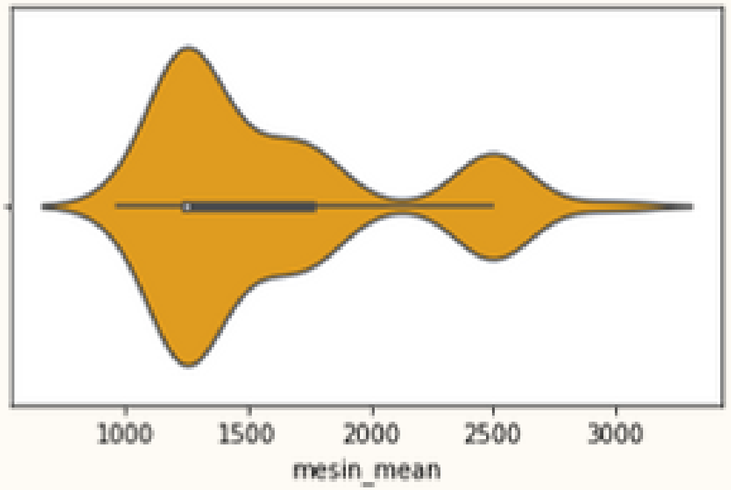


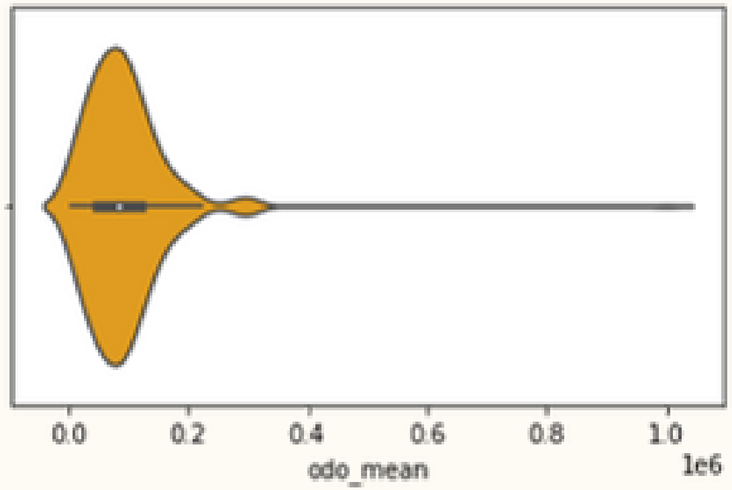
Distribusi fitur numerik

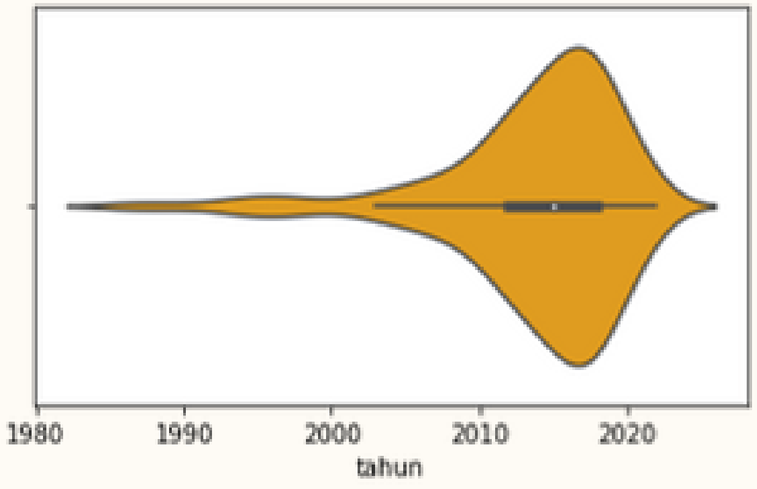












Outlier

Deteksi pencilan dengan metode z-score

- Ditemukan outlier di fitur 'tahun' dan 'odo_mean'
- Jenis natural pada fitur 'tahun' dan error pada fitur 'odo_mean'
- Pengabaian fitur 'tahun' dan imputasi mean fitur 'odo_mean'

train

	id	tahun	z_score_tahun
614	615	1997	-3.236410
1313	1314	1997	-3.236410
1375	1376	1998	-3.048921
158	159	1998	-3.048921
615	616	1998	-3.048921

test

	id	tahun	z_score_tahun
96	97	1989	-3.757972
300	301	1989	-3.757972
279	280	1990	-3.604546
14	15	1993	-3.144268
17	18	1993	-3.144268

train

	id	odo_mean	z_score_odo_mean
188	189	999999.0	11.371178
729	730	999999.0	11.371178
804	805	999999.0	11.371178
1180	1181	999999.0	11.371178
1316	1317	999999.0	11.371178
1392	1393	999999.0	11.371178

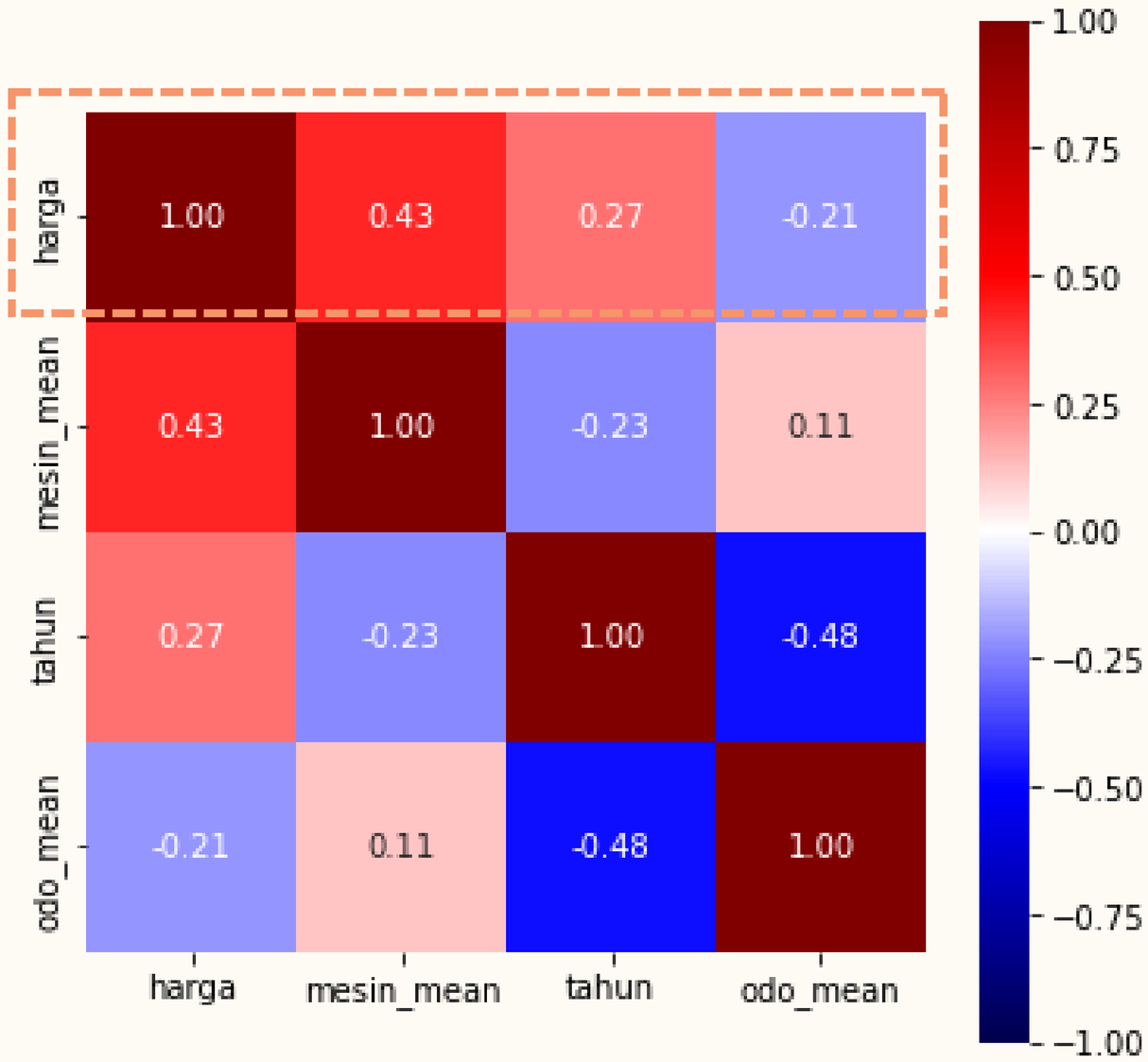
test

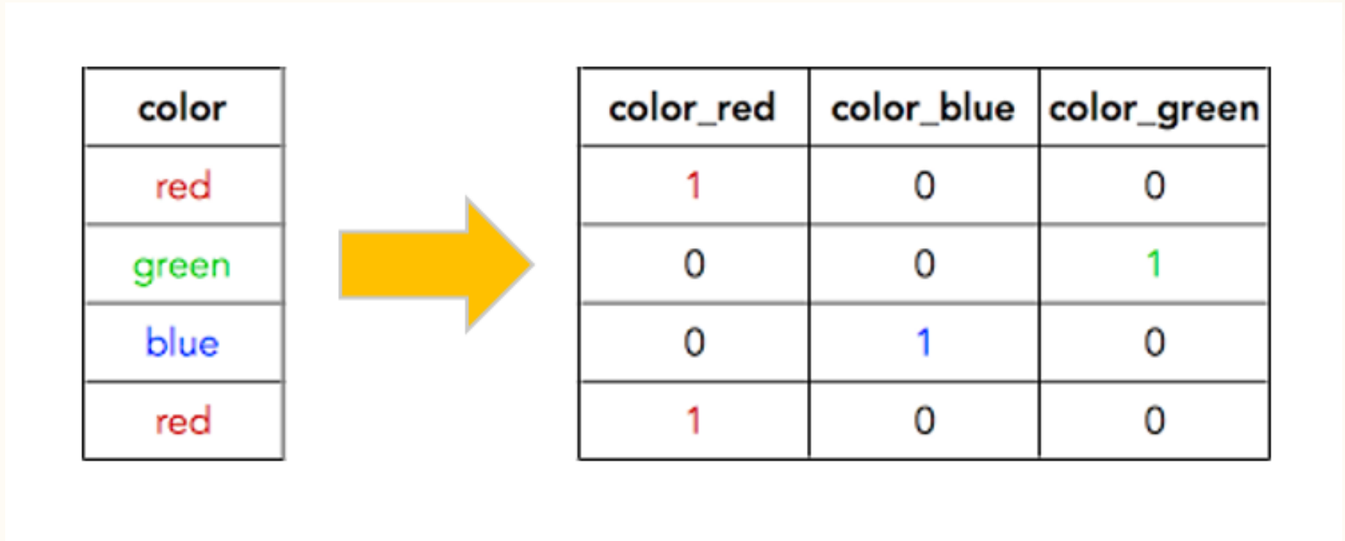
	id	odo_mean	z_score_odo_mean
127	128	999999.0	12.370998

Korelasi

antar fitur dan target

- Pearson correlation
- Multicollinearity





color
red
green
blue
red

color_red	color_blue	color_green
1	0	0
0	0	1
0	1	0
1	0	0

Feature encoding

Encoding fitur categorical

- Metode one-hot encoding
- Pola lebih mudah ditangkap oleh model

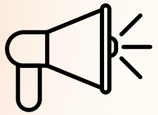
Modeling

Pembagian data

Data train dibagi menjadi set train dan set validasi dengan proporsi 8 : 2

Uji model

Dicari model terbaik dan digunakan uji Friedman untuk mengevaluasi perbedaan error

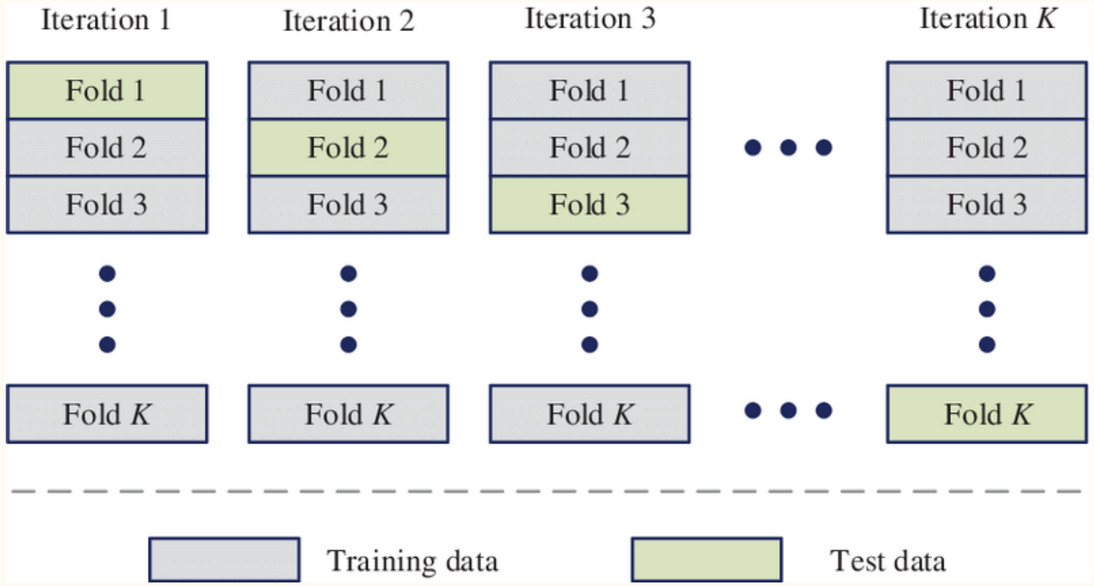


Hyperparameter tuning

Dilakukan hyperparameter tuning dengan library Hyperopt

Validation

K-Fold Cross Validation K=5



Metrics : MAPE

$$\text{MAPE} = \sum_{t=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\%$$

Uji model

Hasil MAPE

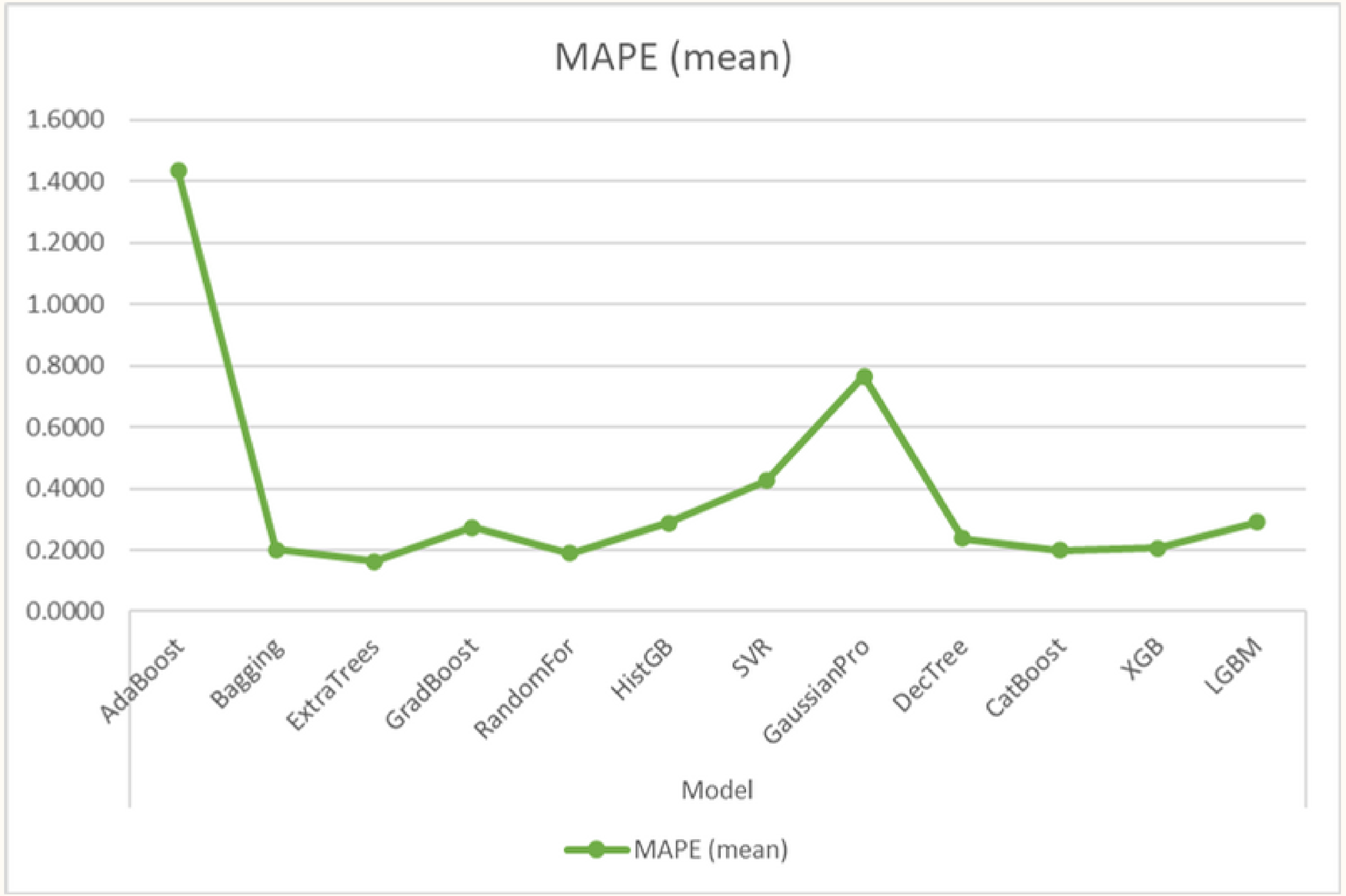
Fold	Model											
	AdaBoost	Bagging	ExtraTrees	GradBoost	RandomFor	HistGB	SVR	GaussianPrc	DecTree	CatBoost	XGB	LGBM
1	1.1567	0.1986	0.1550	0.2665	0.1732	0.2441	0.3985	0.7701	0.2211	0.1838	0.1684	0.2487
2	1.5586	0.2147	0.1641	0.2861	0.2009	0.3137	0.4696	0.7575	0.2222	0.2134	0.2062	0.3136
3	1.5887	0.1939	0.1515	0.2773	0.1754	0.3131	0.4466	0.7528	0.2008	0.2009	0.1884	0.3244
4	1.4334	0.1949	0.1584	0.2693	0.1932	0.2787	0.4371	0.7670	0.2337	0.1911	0.2134	0.2821
5	1.4341	0.2062	0.1811	0.2676	0.2067	0.2906	0.3794	0.7770	0.3108	0.2091	0.2547	0.2864
mean	1.4343	0.2017	0.1620	0.2733	0.1899	0.2880	0.4262	0.7649	0.2377	0.1996	0.2062	0.2910

- H0: Tidak ada perbedaan signifikan antara hasil error dari tiap model
- H1: Ada perbedaan signifikan antara hasil error dari tiap model

Peringkat MAPE

Fold	Model											
	AdaBoost	Bagging	ExtraTrees	GradBoost	RandomFor	HistGB	SVR	GaussianPro	DecTree	CatBoost	XGB	LGBM
1	12	5	1	9	3	7	10	11	6	4	2	8
2	12	5	1	7	2	9	10	11	6	4	3	8
3	12	4	1	7	2	8	10	11	5	6	3	9
4	12	4	1	7	3	8	10	11	6	2	5	9
5	12	2	1	6	3	8	10	11	9	4	5	7

- Diperoleh FM = 51.8307692 dan p = 0.000000293.
- Karena p < 0.05, ditolak hipotesis null yang berarti **ada perbedaan signifikan antara hasil error tiap model.**


MAPE (mean)
1.6000
1.4000
1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
AdaBoost
Bagging
ExtraTrees
GradBoost
RandomFor
HistGB
SVR
GaussianPro
DecTree
CatBoost
XGB
LGBM
Model
MAPE (mean)

Hyperparameter Tuning

Library Hyperopt

- Memanfaatkan optimasi Bayesian sehingga pencarian parameter tidak random

Nilai parameter hasil tuning

- n_estimators : 500
- min_samples_split : 3
- min_samples_leaf : 1

Fold	ExtraTrees	
	Before	After
1	0,1550	0,1369
2	0,1641	0,1712
3	0,1515	0,1614
4	0,1584	0,1641
5	0,1811	0,1610
mean	0,1620	0,1589

Kesimpulan

Model ExtraTreesRegressor yang telah di-tune mampu memprediksi harga jual mobil berdasarkan data penjualan mobil bekas dengan MAPE rata-rata sebesar 15.894%.