Final Project Presentation

Nomor Kelompok: 1
Nama Mentor: Ramdhan Hidayat
Nama:

- Elisabet Indriani
- Dita Wahyuni

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Latar Belakang

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Latar Belakang Project

Sumber Data:
https://www.kaggle.com/datasets/hellbuoy/car-price-prediction
Problem: **Regression**

Tujuan:

- Memprediksi harga mobil berdasarkan variable-variable terkait
- Memperoleh insight mengenai faktor-faktor apa saja yang membedakan mobil ‘murah’ dan mobil ‘mahal’
- Menentukan Faktor/Variable apa yang berpengaruh terhadap harga mobil

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Explorasi Data dan Visualisasi

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Business Understanding

Mobil adalah salah satu jenis transportasi yang sudah digunakan sejak zaman dahulu. Maka dari itu, tidak heran jika mobil menjadi transportasi yang perkembangannya sangat pesat. Mobil merupakan transportasi yang digunakan untuk membantu kita melakukan mobilisasi ke berbagai tempat dengan cepat, nyaman, dan juga nyaman. Saat ini ada berbagai jenis mobil yang dijual di pasaran dan tentunya memiliki spesifikasi serta keunikannya masing-masing.

Oleh karena itu, keterbutuhan akan hal tersebut menjadi peluang bagi para perusahaan mobil untuk semakin menggencarkan produksi mobil.

Bagi para perusahaan mobil, diperlukan pemahaman mengenai faktor-faktor yang mempengaruhi harga mobil di pasar agar produknya mampu bersaing dengan perusahaan mobil lainnya. Karena dengan begitu, mampu membantu perusahaan dalam menentukan harga berdasarkan variable-variable yang bervariasi

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

About Dataset

Perusahaan mobil XYZ dari Jepang yang bercita-cita untuk memasuki pasar AS dengan mendirikan unit manufaktur mereka di sana dan memproduksi mobil secara lokal untuk memberikan persaingan dengan rekan-rekan mereka di AS dan Eropa.Mereka ingin memahami faktor-faktor yang mempengaruhi harga mobil di pasar Amerika, karena mungkin sangat berbeda dengan pasar Jepang.
Pada dasarnya, perusahaan ingin mengetahui:

- **Variabel mana yang signifikan dalam memprediksi harga mobil?**
- **Seberapa baik variabel tersebut menggambarkan harga mobil?**

Sebagai seorang Data scientist dituntut untuk menerapkan beberapa teknik data science untuk harga mobil dengan variabel independen yang tersedia. Itu akan membantu manajemen untuk memahami bagaimana tepatnya harga bervariasi dengan variabel independen. Mereka dapat memanipulasi desain mobil, strategi bisnis, dll. untuk memenuhi tingkat harga tertentu.

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Data Cleansing

- Profile Data

```
data.shape
```

```
(205, 26)
```

```
data.duplicated().sum()
```

```
0
```

```
data.isnull().any()
data.isnull().sum()/ data.shape[0]
```

```
car_ID              0.0
symboling           0.0
CarName             0.0
fueltype            0.0
aspiration          0.0
doornumber          0.0
carbody             0.0
drivewheel          0.0
enginelocation      0.0
wheelbase           0.0
carlength           0.0
carwidth            0.0
carheight           0.0
curbweight          0.0
enginetype          0.0
cylindernumber      0.0
enginesize          0.0
fuelsystem          0.0
boreratio           0.0
stroke              0.0
compressionratio    0.0
horsepower          0.0
peakrpm             0.0
citympg             0.0
highwaympg          0.0
price               0.0
dtype: float64
```

- Terdapat 205 baris dan 26 kolom pada dataset tersebut
- Tidak ada data yang duplicate pada dataset “Car Price Prediction”
- Tidak ada *missing value* pada dataset yang dipilih

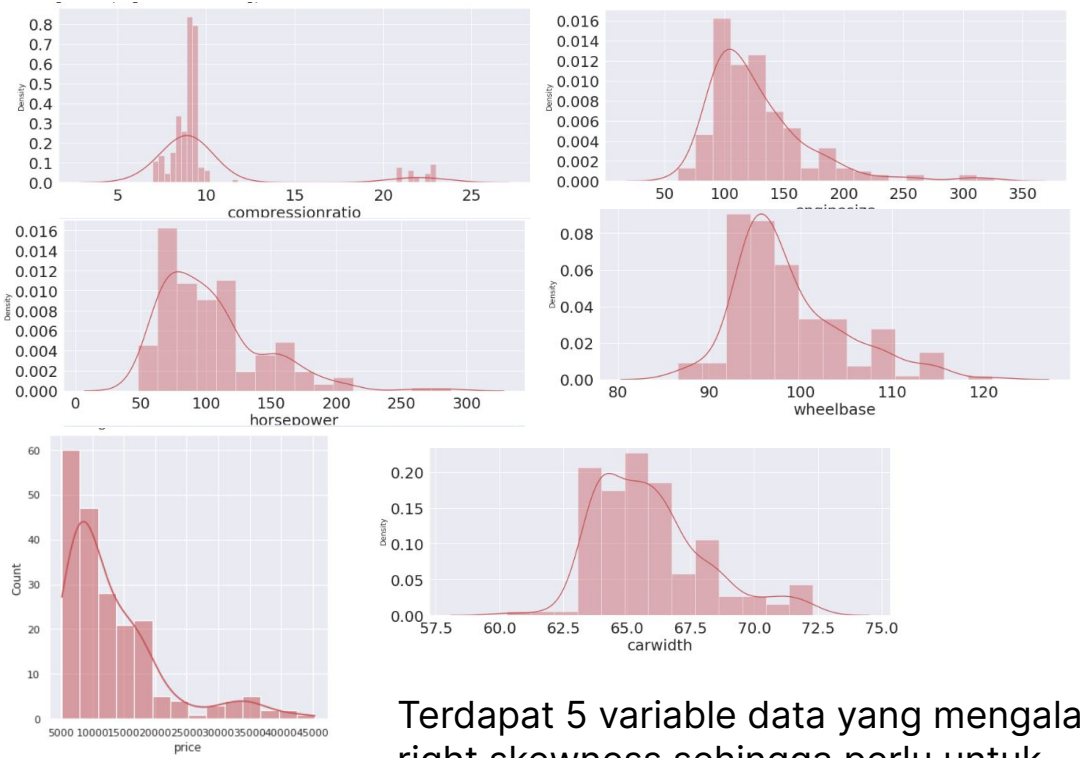
© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Data Cleansing

- Profile Data

```
data.skew()
```

car_ID	0.000000
symboling	0.211072
wheelbase	1.050214
carlength	0.155954
carwidth	0.904003
carheight	0.063123
curbweight	0.681398
enginesize	1.947655
boreratio	0.020156
stroke	-0.689705
compressionratio	2.610862
horsepower	1.405310
peakrpm	0.075159
citympg	0.663704
highwaympg	0.539997
price	1.777678



Terdapat 5 variable data yang mengalami right skewness sehingga perlu untuk ditangani agar asumsi normalitas terpenuhi dan mampu menghasilkan model machine learning yang baik

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Data Cleansing

- Profile Data

```
'mitsubishi pajero', 'Nissan versa', 'nissan gt-r', 'nissan rogue',
'nissan latio', 'nissan titan', 'nissan leaf', 'nissan juke',
'nissan note', 'nissan clipper', 'nissan nv200', 'nissan dayz',
'nissan fuga', 'nissan otti', 'nissan teana', 'nissan kicks',
```

```
'subaru tribeca', 'toyota corona mark ii', 'toyota corona',
'toyota corolla 1200', 'toyota corona hardtop',
'toyota corolla 1600 (sw)', 'toyota carina', 'toyota mark ii',
'toyota corolla', 'toyota corolla liftback',
'toyota celica gt liftback', 'toyota corolla tercel',
'toyota corona liftback', 'toyota starlet', 'toyota tercel',
'toyota cressida', 'toyota celica gt', 'toyouta tercel',
```

```
'vokswagen rabbit', 'volkswagen 1131 deluxe sedan',
'volkswagen model 111', 'volkswagen type 3', 'volkswagen 411 (sw)',
'volkswagen super beetle', 'volkswagen dasher', 'vw dasher',
'vw rabbit', 'volkswagen rabbit', 'volkswagen rabbit custom',
```

```
'maxda rx3', 'maxda glc deluxe', 'mazda rx2 coupe', 'mazda rx-4',
'mazda glc deluxe', 'mazda 626', 'mazda glc', 'mazda rx-7 gs',
'mazda glc 4', 'mazda glc custom l', 'mazda glc custom',
```

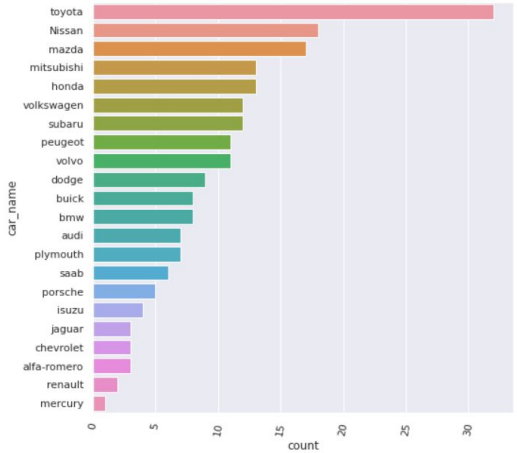
```
'porsche macan', 'porcshce panamera', 'porsche cayenne',
'porsche boxter',
```

Terdapat beberapa kesalahan/perbedaan dalam penamaan serta kapitalisasi variable “CarName” sehingga perlu diberlakukan normalisasi agar dapat dikelompokkan dan diolah ke proses yang berikutnya.

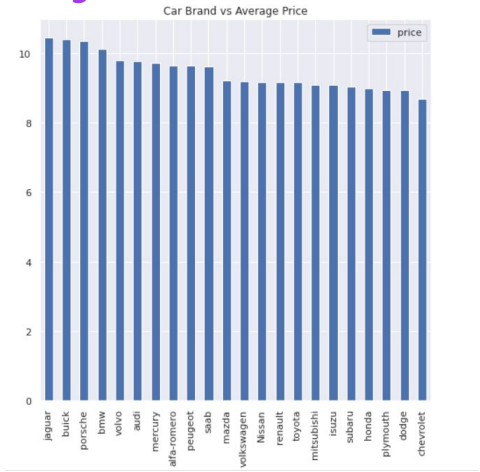
Bentuk Normalisasi :
- nissan dan Nissan menjadi **nissan**
- toyota dan toyouta menjadi **toyota**
- volkswagen dan vw menjadi **volkswagen**
- mazda dan maxda menjadi **mazda**
- porsche dan porcshce menjadi **prosche**

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Exploratory Data Analysis



- Toyota adalah brand mobil yang memiliki jumlah paling banyak pada dataset
- Mercury adalah brand mobil yang memiliki jumlah paling banyak pada dataset



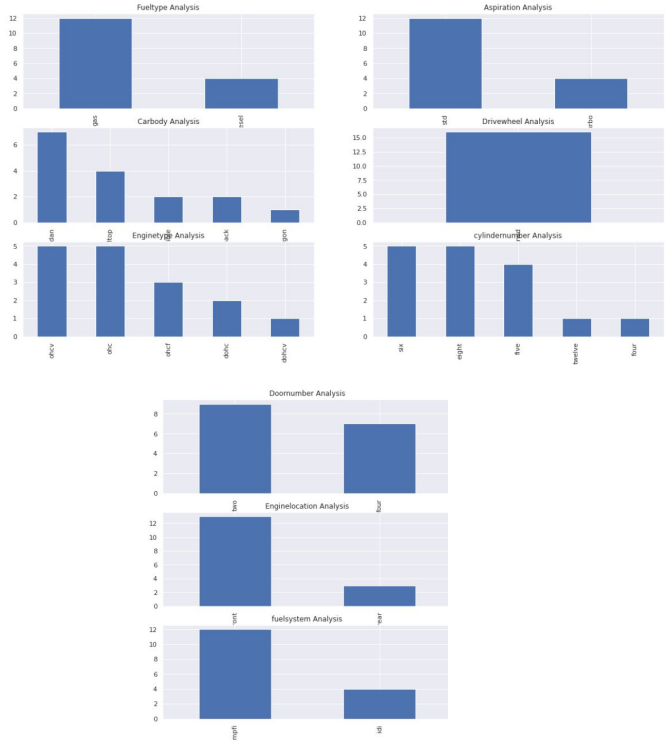
-Jaguar,Buick dan porsche memiliki rata-rata harga yang tinggi
-Chevrolet merupakan brand yang memiliki rata-rata harga termurah

count	205.000000
mean	13276.710571
std	7988.852332
min	5118.000000
25%	7788.000000
50%	10295.000000
75%	16503.000000
max	45400.000000

- Rata-rata harga mobil adalah 13276,7

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Brand with High Price Analysis

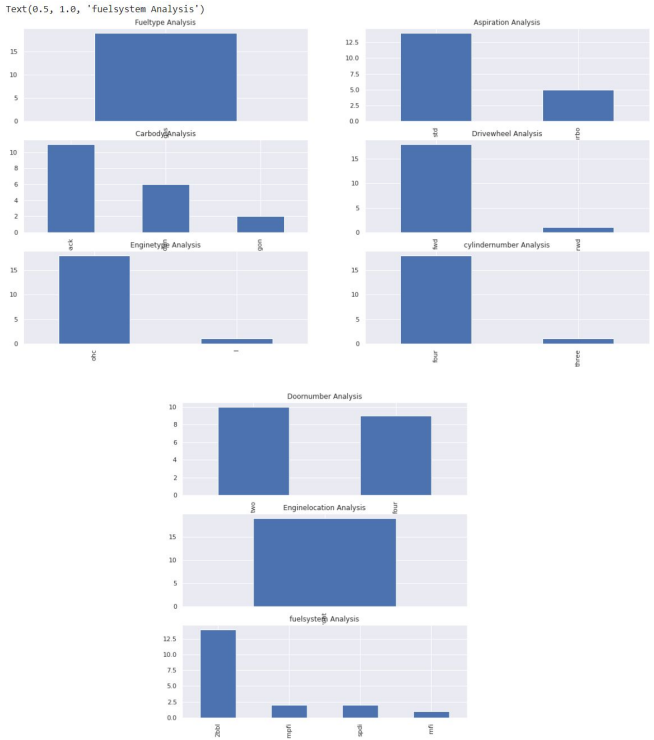


Dengan mengambil 3 brand yang memiliki average price tertinggi, maka dapat terlihat bahwa kebanyakan mobil yang memiliki average price tertinggi adalah mobil dengan spesifikasi :

- Fueltype : gas
- Aspiration : std
- Doornumber : two atau four
- Carbody : sedan
- Drivewheel : rwd
- Enginelocation : front
- Enginetype : ohcv atau ohc
- cylinder number : six atau eight
- fuelsystem : mpfi

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Brand with Low Price Analysis

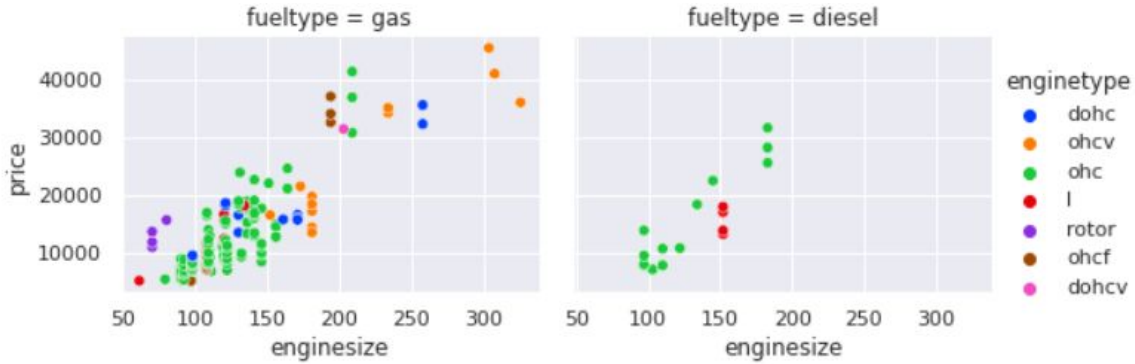


Dengan mengambil 3 brand yang memiliki average price terendah, maka dapat terlihat bahwa kebanyakan mobil yang memiliki average price terendah adalah mobil dengan spesifikasi :

- Fueltype : gas
- Aspiration : std
- Doornumber : two atau four
- Carbody : hatchback
- Drivewheel : fwd
- Enginelocation : front
- Fuelsystem : ohc
- cylinder number : four
- fuelsystem : 2bbl

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Exploratory Data Analysis

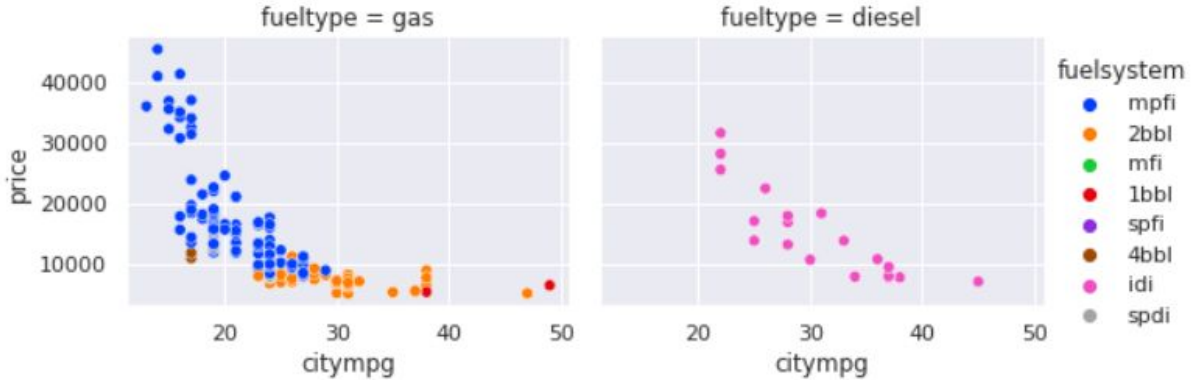


- **Enginesize VS Enginetype VS Price**

- Semakin besar enginesizenya, semakin mahal harganya. Enginetype bertipe ohcv hanya berada pada mobil dengan fueltype bertipe gas. Enginetype bertipe ohcv berada pada mobil dengan enginesize diatas 150

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Exploratory Data Analysis

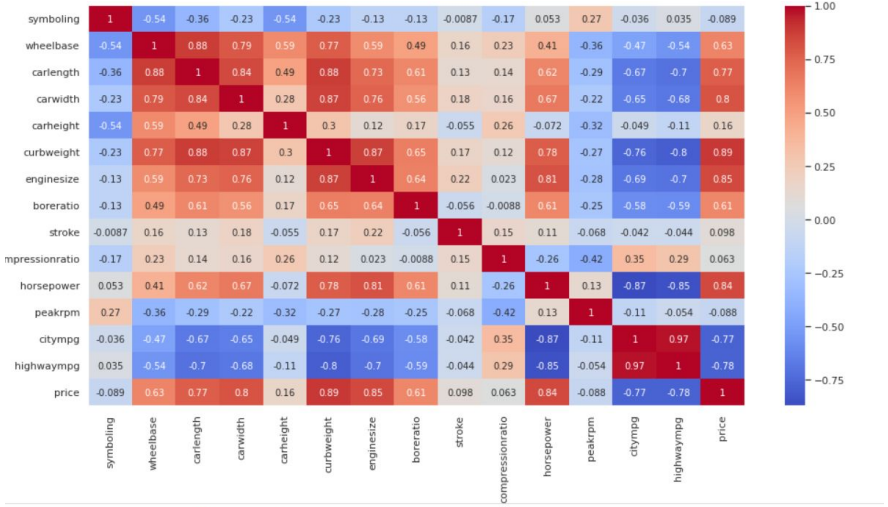


- **Citympg VS Fuelsystem VS Price**

- Semakin besar citympg, semakin murah harganya. Fuelsystem bertipe mpfi hanya berada pada mobil dengan fueltype bertipe gas dan fuelsystem bertipe idi hanya berada pada mobil dengan fueltype bertipe diesel. Fuelsystem bertipe mpfi memberikan nilai citympg yang rendah sehingga berdampak pula pada harga jual mobil

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Exploratory Data Analysis



- carwidth , carlength , curbweight , enginesize , horsepower, wheelbase, dan boreratio memiliki korelasi positif yang cukup kuat dengan price.
- carheight tidak menunjukkan tren yang signifikan dengan price.
- citympg , highwaympg, peakrpm dan symboling memiliki korelasi negatif yang cukup kuat terhadap dengan price.
- Variabel yang memiliki korelasi kuat: **wheelbase, carlength, carwidth, curbweight,enginesize,boreratio,horsepower,citymp,highwaympg**

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Modelling

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Feature Selection

- **Using Recursive Feature Elimination (RFE)**

Method	Model Accuration
Linear Regression	0,91958
Ridge Regression	**0,93218**
Decision Tree Regressor	0.87852
Random Forest Regressor	0.93067
LASSO	0,86227
Elastic Net	0,86175

- **Based on Feature Correlation**

('price','enginetype','fueltype', 'aspiration','carbody','cylindernumber', 'drivewheel','citympg', ' highwaympg','curbweight', 'enginesize','horsepower', 'carlength','carwidth', 'enginelocation')

Method	Model Accuration
Linear Regression	0.92228
Ridge Regression	**0,92458**
Decision Tree Regressor	0,88018
Random Forest Regressor	0,92334
LASSO	0.84810
Elastic Net	0,84777

Melihat Nilai Akurasi Model antara Metode Feature Selection RFE dan Feature Correlation, dapat dihighlight bahwa **RFE memiliki nilai akurasi lebih baik dari pada Metode Feature Correlation**, sehingga selanjutnya kita akan menggunakan feature yang didapat dari metode RFE. Feature yang dipilih adalah **'carbody', 'wheelbase', 'carlength', 'carwidth', 'carheight', 'curbweight', 'enginesize', 'fuelsystem', 'stroke', 'compressionratio', 'horsepower', 'peakrpm', 'citympg', 'highwaympg', 'Brand' dan 'price'**

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Train Test Split

Variable Dependent : **Price**
Variabele Independent : **carbody,wheelbase, carlength, carwidth, carheight, curbweight, enginesize, fuelsystem, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, Brand**
Train Data : **60%**
Testing Data : **40%**

Evaluation Metrics

- **RMSE : Mengukur kesalahan rata-rata yang dilakukan oleh model dalam memprediksi hasil untuk suatu model**
- **MSE : Dikenal sebagai model sigma, merupakan varian dari RMSE yang disesuaikan dengan jumlah predictor dalam model**
- **MAE : Mengukur kesalahan prediksi dengan menghitung perbedaan absolut rata-rata antara hasil yang diamati dengan prediksi**
- **R2 : Metrik yang memberi tahu kita proporsi varians dalam variabel respons dari model regresi yang dapat dijelaskan oleh variabel prediktor**

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Data Modelling

Hasil dibawah ini didapatkan dengan menggunakan metode **cross validation (K-Fold) dan feature selection (RFE)** yang diaplikasikan kepada beberapa metode regression.

Method	MAE	MSE	RMSE	R^2
Linear Regression	0,11	0,0202	0,142	0,91958
Ridge Regression	**0,1027**	**0,017**	**0,1304**	**0,93218**
Decision Tree Regressor	0,1233	0,0304	0,1745	0,87853
Random Forest Regressor	0,102	0,0174	0,1318	0,93067
LASSO	0,1461	0,0345	0,1858	0,86227
Elastic Net	0,1464	0,0346	0,1861	0,86175

Berdasarkan hasil yang diperoleh, dapat diputuskan bahwa model terbaik adalah model yang dibuat dengan menggunakan **metode Ridge Regression** yang mana disimpulkan berdasarkan **nilai dari metriks evaluasinya yang paling rendah** dengan **nilai R^2 paling besar dibandingkan dengan metode yang lainnya**

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Hyperparameter Tuning

Untuk mencoba menambah akurasi model, dilakukan proses hyperparameter tuning untuk melakukan eksperimen terhadap model guna menambah tingkat akurasi. Didapatkan hasil sebagai berikut :

Method	MAE Before	MAE After	MSE Before	MSE After	RMSE Before	RMSE After	R^2 Before	R^2 After
Ridge Regression	**0,1027**	**0,1022**	**0,017**	**0,017**	**0,1304**	**0,1302**	**0,9321**	**0,9324**

Improvement : 0,12%

Dapat terlihat bahwa terjadi improvement pada model akurasi ketika melakukan proses hyperparameter, sehingga dapat disimpulkan bahwa model mengalami kenaikan akurasi dengan menggunakan proses hyperparameter tuning

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Final Model

Method	MAE	MSE	RMSE	R^2
Ridge Regression	**0,1022**	**0,017**	**0,1302**	**0,9324**

Setelah dilakukan pemilihan model terbaik dan proses peningkatan model akurasi dengan metode hyperparameter tuning, diperoleh final model yang dipilih sebagai model terbaik adalah yang dibangun dengan metode Ridge Regression dengan tingkat akurasi model sebesar 93,24%

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Conclusion

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Conclusion

- Mobil yang memiliki harga mahal memiliki spesifikasi khusus yang berbeda dari mobil yang memiliki harga murah yaitu

1. Mobil mahal kebanyakan memiliki **carbody bertipe sedan** sedangkan mobil murah memiliki **carbody bertipe hatchback**
2. Mobil mahal kebanyakan memiliki **drivewheel bertipe rwd** sedangkan mobil murah memiliki **drivewheel bertipe fwd**
3. Mobil mahal kebanyakan memiliki **cylindernumber yaitu six atau eight** sedangkan mobil murah memiliki **cylindernumber yaitu four**
4. Mobil mahal kebanyakan memiliki **fuelsystem bertipe mpfi** sedangkan mobil murah memiliki **fuelsystem bertipe 2bbl**
5. Semakin tinggi nilai **carlength,carwidth,curbweight,enginesize,horsepower** maka nilai harga mobil semakin mahal atau sebaliknya
6. Semakin tinggi nilai **citympg dan highwaympg** maka nilai harga mobil semakin murah atau sebaliknya.

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Conclusion

- Dilakukan 2 metode Feature Selection yaitu RFE dan Feature Correlation dan diperoleh hasil bahwa akurasi model lebih besar dengan menggunakan RFE. Metode Cross validation dg menggunakan metode **K-Fold**, dan model terbaik yang dihasilkan adalah **Ridge Regression**
- Hyperparameter Tuning berhasil meningkatkan akurasi model sebesar **0,12%**
- Variabel2 yang bisa dijadikan variable predictor adalah **carbody,wheelbase, carlength, carwidth, carheight, curbweight, enginesize, fuelsystem, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, Brand**
- Perusahaan disarankan untuk menentukan harga jual mobil berdasarkan variabel yang telah dipilih untuk dijadikan variable prediktor, karena memiliki relevansi yang kuat

© 2022 Program Studi Independen Bersertifikat Zenius Bersama Kampus Merdeka

Terima kasih!

Ada pertanyaan?