**ANALISIS KEBAIKAN MODEL REGRESI PADA DATA AUTO MPG DENGAN MELAKUKAN *PRE-PROCESSING* DATA DAN TRANSFORMASI DATA**

Pada Saat ini kendaraan yang paling banyak digunakan oleh masyarakat adalah *city cycle. City cycle* tentu saja memerlukan bahan bakar untuk dapat berjalan. Penggunaan bahan bakar ini tentunya cukup beragam dari masing-masing masyarakat. Penggunaan bahan bakar dari setiap kendaraan tentu saja beragam. Hal ini ditentukan dari beberapa faktor yang dapat menyebabkan penggunaan bahan bakar menjadi banyak atau sedikit. Adapun faktor-faktor yang mempengaruhi dari penggunaan bahan bakar adalah faktor *displacement, horsepower,* dan *weight*. Faktor-faktor ini akan mempengaruhi penggunaan bahan bakar kendaraan per mil per galonnya.

Data dengan karakteristik seperti ini dapat dianalisis seberapa besar pengaruh yang diberikan setiap faktor terhadap predictor. Dan dalam hal ini akan dianalisis dengan menggunakan metode regresi. Data yang digunakan dalam penelitian ini terdapat *missing value*, sehingga *missing value* tersebut akan dideteksi dan diatasi, selanjutnya dilakukan regresi dan didapatkan model regresinya. Dari hasil yang didapatkan, akan dilakukan perbandingan dengan model regresi lainnya yang didapatkan setelah mengatasi *outlier* dan melakukan transformasi data.

Berikut merupakan hasil analisis regresi dengan menggunakan bantuan *software Pyton.*

1. **Model Regresi dengan Melakukan *Missing Value***

Hasil *Output:*

OLS Regression Results ==========================================================================Dep. Variable: mpg R-squared: 0.703

Model: OLS Adj. R-squared: 0.701

Method: Least Squares F-statistic: 311.0

Date: Sun, 10 Mar 2019 Prob (F-statistic): 1.68e-103 Time: 03:05:47 Log-Likelihood: -1140.8

No. Observations: 398 AIC: 2290.

Df Residuals: 394 BIC: 2305.

Df Model: 3

Covariance Type: nonrobust ========================================================================== coef std err t P>|t| [0.025 0.975] --------------------------------------------------------------------------displacement -0.0068 0.007 -1.028 0.305 -0.020 0.006

Horsepower -0.0372 0.013 -2.933 0.004 -0.062 -0.012

weight -0.0054 0.001 -7.640 0.000 -0.007 -0.004

b 44.9344 1.200 37.435 0.000 42.575 47.294 ==========================================================================

Omnibus: 37.768 Durbin-Watson: 0.853

Prob(Omnibus): 0.000 Jarque-Bera (JB): 49.930

Skew: 0.705 Prob(JB): 1.44e-11

Kurtosis: 4.012 Cond. No. 1.74e+04 ==========================================================================

Berdasarkan hasil output diatas terlihat bahwa nilai *prob (F-Statistic)* sebesar 1,68e-103 lebih kecil < dari α (0,05), sehingga dapat disimpulkan faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight* secara simultan berpengaruh terhadap penggunaan bahan bakar. Serta nilai *R-Squared* atau koefisien determinasi dengan menggunakan model regresi OLS yaitu sebesar 0,703 atau 70,3%. Nilai ini bermakna bahwa faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight*, berpengaruh terhadap penggunaan bahan bakar yaitu sebesar 70,3%. Sedangkan sisanya 29,7% dipengaruhi oleh faktor yang lainnya.

1. **Model Regresi dengan Melakukan *Missing Value* dan *Outlier***

Hasil *Output:*

OLS Regression Results ==========================================================================Dep. Variable: mpg R-squared: 0.707

Model: OLS Adj. R-squared: 0.705

Method: Least Squares F-statistic: 312.0

Date: Sun, 10 Mar 2019 Prob (F-statistic): 5.20e-1033 Time: 18:33:52 Log-Likelihood: -1140.8

No. Observations: 392 AIC: 2239.

Df Residuals: 388 BIC: 2255.

Df Model: 3

Covariance Type: nonrobust ========================================================================== coef std err t P>|t| [0.025 0.975] --------------------------------------------------------------------------displacement -0.0066 0.007 -0.995 0.320 -0.020 0.006

horsepower -0.0444 0.013 -3.369 0.001 -0.070 -0.018

weight -0.0053 0.001 -6.931 0.000 -0.007 -0.004

b 44.9922 1.205 37.324 0.000 42.622 47.362

==========================================================================

Omnibus: 28.268 Durbin-Watson: 0.842

Prob(Omnibus): 0.000 Jarque-Bera (JB): 33.744

Skew: 0.618 Prob(JB): 4.70e-08

Kurtosis: 3.734 Cond. No. 1.76e+04

==========================================================================

Berdasarkan hasil output diatas terlihat bahwa nilai *prob (F-Statistic)* sebesar 5,20e-1033 lebih kecil < dari α (0,05), sehingga dapat disimpulkan faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight* secara simultan berpengaruh terhadap penggunaan bahan bakar. Serta nilai *R-Squared* atau koefisien determinasi dengan menggunakan model regresi OLS yaitu sebesar 0,707 atau 70,7%. Nilai ini bermakna bahwa faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight*, berpengaruh terhadap penggunaan bahan bakar yaitu sebesar 70,7%. Sedangkan sisanya 29,3% dipengaruhi oleh faktor yang lainnya.

1. **Model Regresi dengan Melakukan *Missing Value* dan Transformasi**

Hasil *Output:*

OLS Regression Results ==========================================================================Dep. Variable: mpg R-squared: 0.891

Model: OLS Adj. R-squared: 0.890

Method: Least Squares F-statistic: 1071.

Date: Sun, 10 Mar 2019 Prob (F-statistic): 2.68e-189 Time: 18:35:30 Log-Likelihood: 1687.7

No. Observations: 398 AIC: -3369.

Df Residuals: 395 BIC: -3357.

Df Model: 3

Covariance Type: nonrobust ========================================================================== coef std err t P>|t| [0.025 0.975] -------------------------------------------------------------------------- 1 -0.2178 0.010 -20.866 0.000 -0.238 -0.197

2 0.0838 0.031 2.724 0.007 0.023 0.144

3 0.0196 0.001 18.358 0.000 0.018 0.022

==========================================================================

Omnibus: 23.617 Durbin-Watson: 1.114

Prob(Omnibus): 0.000 Jarque-Bera (JB): 27.440

Skew: 0.546 Prob(JB): 1.10e-06

Kurtosis: 3.681 Cond. No. 177.

Berdasarkan hasil output diatas terlihat bahwa nilai *prob (F-Statistic)* sebesar 2,68e-189 lebih kecil < dari α (0,05), sehingga dapat disimpulkan faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight* secara simultan berpengaruh terhadap penggunaan bahan bakar. Serta nilai *R-Squared* atau koefisien determinasi dengan menggunakan model regresi OLS yaitu sebesar 0,891 atau 89,1%. Nilai ini bermakna bahwa faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight*, berpengaruh terhadap penggunaan bahan bakar yaitu sebesar 89,1%. Sedangkan sisanya 10,9% dipengaruhi oleh faktor yang lainnya.

1. **Model Regresi dengan Melakukan *Missing Value, Outlier* dan Transformasi**

Hasil *Output:*

OLS Regression Results ==========================================================================Dep. Variable: mpg R-squared: 0.894

Model: OLS Adj. R-squared: 0.893

Method: Least Squares F-statistic: 1096.

Date: Sun, 10 Mar 2019 Prob (F-statistic): 2.51e-189

Time: 18:33:52 Log-Likelihood: 1668.5

No. Observations: 392 AIC: -3331.

Df Residuals: 389 BIC: -3319.

Df Model: 3

Covariance Type: nonrobust ========================================================================== coef std err t P>|t| [0.025 0.975] -------------------------------------------------------------------------- 1 -0.2232 0.011 -21.175 0.000 -0.244 -0.202

2 0.0531 0.032 1.651 0.100 -0.010 0.116

3 0.0210 0.001 18.001 0.000 0.019 0.023 ==========================================================================

Omnibus: 17.194 Durbin-Watson: 1.119

Prob(Omnibus): 0.000 Jarque-Bera (JB): 18,947

Skew: 0.460 Prob(JB): 7.68e-05

Kurtosis: 3.562 Cond. No. 185.

Berdasarkan hasil output diatas terlihat bahwa nilai *prob (F-Statistic)* sebesar 2,51e-189 lebih kecil < dari α (0,05), sehingga dapat disimpulkan faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight* secara simultan berpengaruh terhadap penggunaan bahan bakar. nilai *R-Squared* atau koefisien determinasi dengan menggunakan model regresi OLS yaitu sebesar 0,894 atau 89,4%. Nilai ini bermakna bahwa faktor-faktor yang terkait yaitu *displacement, horsepower,* dan *weight*, berpengaruh terhadap penggunaan bahan bakar yaitu sebesar 89,4%. Sedangkan sisanya 10,6% dipengaruhi oleh faktor yang lainnya.

1. **Perbandingan Hasil Model Regresi**

Berikut merupakan hasil perbandingan dari model regresi yang didapatkan.

|  |  |
| --- | --- |
| **Deteksi** | **Nilai *R-Squared*** |
| *Missing Value* | 0,703 |
| *Missing Value* dan *Outlier* | 0,707 |
| *Missing Value* dan *Transformasi* | 0,891 |
| *Missing Value*, *Outlier*, dan *Transformasi* | 0,894 |

Berdasarkan tabel tersebut terlihat bahwa nilai *R-Squared* semakin besar apabila data melalui tahap-tahap *pre-processing* data dan transformasi data yang lebih komplek. Hal ini dibuktikan nilai *R-Squared* yang hanya dideteksi *missing value*nya saja senilai 0,703, nilai ini lebih kecil dibandingkan dengan nilai *R-Squared* yang lainnya. Nilai *R-Squared* berada diantara kisaran 0 – 1. Semakin nilai mendekati 1, maka model regresi tersebut akan semakin baik. Dan pada penelitian ini, model regresi yang palin baik adalah model regresi pada data yang telah melalui proses perbaikan *missing value, outlier,* dan transformasi.

Sehingga pada saat melakukan analisis data, sebaiknya data yang akan diolah harus melalui proses pre-processing data dan transformasi data, hal ini akan berpengaruh terhadap hasil analisis yang akan semakin baik.