**TUGAS 3 DATA MINING**

**MEMBANDINGKAN MODEL REGRESI DARI BERBAGAI TAHAP *PREPOCESSING DATA***

Sumber data yang digunakan dalam analisis ini adalah data sekunder yang diperoleh melalui website https://www.kaggle.com yaitu data cuaca di Australia pada jam 9 pagi. Variabel yang digunakan dalam analisis ini berjumlah 4 variabel dengan jumlah observasi sebesar 142192. Berikut merupakan variabel penelitian yang digunakan dalam analisis kali ini.

|  |  |
| --- | --- |
| **Variabel** | **Keterangan** |
| Win Speed () | Kecepatan angin |
| Humidity () | Kandungan uap air dalam udara |
| Pressure () | Tekanan udara |
| Temperatur () | Suhu udara |

Sebelum dilakukan analisis lebih lanjut, perlu dilakukannya *preprocessing data*. Ada beberapa tahap dalam melakukan *preprocessing data*. Pada analisis kali ini, tahap *preprocessing data* yang dilakukan diantaranya yaitu *data cleaning* dan transformasi data. Kemudian dilakukan analisis regresi untuk menentukan model terbaik terhadap data yang dianalisis. Analisis kali ini, data diolah menggunakan program *Python.* Metode yang digunakan dalam mentransformasi data yaitu menggunakan *softmax*. Berikut hasil analisis regresi data cuaca di Australia pada jam 9 pagi.

1. Regresi menggunakan data yang telah diatasi *missing value* didalamnya

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **R-squared:** | 0.899 |
| **Model:** | OLS | **Adj. R-squared:** | 0.899 |
| **Method:** | Least Squares | **F-statistic:** | 4.212e+05 |
| **Date:** | Sat, 09 Mar 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:26:56 | **Log-Likelihood:** | -4.5126e+05 |
| **No. Observations:** | 142193 | **AIC:** | 9.025e+05 |
| **Df Residuals:** | 142190 | **BIC:** | 9.026e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **x1** | 0.0105 | 0.002 | 5.836 | 0.000 | 0.007 | 0.014 |
| **x2** | -0.1580 | 0.001 | -187.621 | 0.000 | -0.160 | -0.156 |
| **x3** | 0.0272 | 6.95e-05 | 391.857 | 0.000 | 0.027 | 0.027 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1623.513 | **Durbin-Watson:** | 0.252 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1411.270 |
| **Skew:** | 0.190 | **Prob(JB):** | 3.52e-307 |
| **Kurtosis:** | 2.694 | **Cond. No.** | 120. |

1. Regresi menggunakan data yang telah diatasi *missing value* dan *outlier* didalamnya

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **R-squared:** | 0.899 |
| **Model:** | OLS | **Adj. R-squared:** | 0.899 |
| **Method:** | Least Squares | **F-statistic:** | 4.134e+05 |
| **Date:** | Sat, 09 Mar 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:32:58 | **Log-Likelihood:** | -4.4230e+05 |
| **No. Observations:** | 139471 | **AIC:** | 8.846e+05 |
| **Df Residuals:** | 139468 | **BIC:** | 8.846e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **x1** | 0.0169 | 0.002 | 8.763 | 0.000 | 0.013 | 0.021 |
| **x2** | -0.1552 | 0.001 | -178.950 | 0.000 | -0.157 | -0.154 |
| **x3** | 0.0270 | 7.21e-05 | 374.311 | 0.000 | 0.027 | 0.027 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1999.133 | **Durbin-Watson:** | 0.250 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1682.203 |
| **Skew:** | 0.206 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 2.653 | **Cond. No.** | 129. |

1. Regresi menggunakan data yang telah diatasi *missing value* didalamnya dan kemudian ditransformasikan

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **R-squared:** | 0.900 |
| **Model:** | OLS | **Adj. R-squared:** | 0.900 |
| **Method:** | Least Squares | **F-statistic:** | 4.287e+05 |
| **Date:** | Sun, 10 Mar 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 15:46:55 | **Log-Likelihood:** | -4.5013e+05 |
| **No. Observations:** | 142193 | **AIC:** | 9.003e+05 |
| **Df Residuals:** | 142190 | **BIC:** | 9.003e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **x1** | 1.279e+06 | 2.52e+05 | 5.069 | 0.000 | 7.84e+05 | 1.77e+06 |
| **x2** | -2.165e+07 | 1.12e+05 | -193.494 | 0.000 | -2.19e+07 | -2.14e+07 |
| **x3** | 2.278e+07 | 2.89e+05 | 78.861 | 0.000 | 2.22e+07 | 2.33e+07 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1609.777 | **Durbin-Watson:** | 0.253 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1384.837 |
| **Skew:** | 0.185 | **Prob(JB):** | 1.93e-301 |
| **Kurtosis:** | 2.689 | **Cond. No.** | 303. |

1. Regresi menggunakan data yang telah diatasi *missing value* dan *outlier* didalamnya kemudian ditransformasikan

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **R-squared:** | 0.901 |
| **Model:** | OLS | **Adj. R-squared:** | 0.901 |
| **Method:** | Least Squares | **F-statistic:** | 4.210e+05 |
| **Date:** | Sun, 10 Mar 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 15:47:37 | **Log-Likelihood:** | -4.4115e+05 |
| **No. Observations:** | 139471 | **AIC:** | 8.823e+05 |
| **Df Residuals:** | 139468 | **BIC:** | 8.823e+05 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **x1** | 2.101e+06 | 2.68e+05 | 7.852 | 0.000 | 1.58e+06 | 2.63e+06 |
| **x2** | -2.089e+07 | 1.13e+05 | -185.460 | 0.000 | -2.11e+07 | -2.07e+07 |
| **x3** | 2.116e+07 | 3.03e+05 | 69.755 | 0.000 | 2.06e+07 | 2.18e+07 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1983.135 | **Durbin-Watson:** | 0.250 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 1652.098 |
| **Skew:** | 0.201 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 2.650 | **Cond. No.** | 324. |

Perbandingan model regresi

|  |  |
| --- | --- |
| **Tahap Prepocessing Data** | **R-sq** |
| *Missing Value* | 0.899 |
| *Missing Value* dan *Outlier* | 0.899 |
| *Missing Value* dan Transformasi Data | 0.900 |
| *Missing Value, Outlier,* dan Transformasi Data | 0.901 |

Pada tabel diatas dapat dilihat bahwa nilai *R square* dari keempat model mendekati nilai 1. Hal ini menunjukkan bahwa hubungan antara variabel prediktor dan variabel respon sangat kuat dan positif. Dan nilai *R square* terbesar yaitu pada analisis regresi pada data yang telah diatasi *missing value, outlier*, kemudian ditransformasikan, dengan nilai *R square* sebesar 0.901. Sehingga dapat disimpulkan bahwa model terbaik didapatkan dari data yang berkualitas yaitu data yang telah diatasi *missing value* dan *outlier-*nya*,* dan telah ditansformasikan.