Sumber Data : Real Estate (UCI Machine Learning)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **House age (X1)** | **Distance to the nearest MRT station (X2)** | **Number of convenience stores (X3)** | **Latitude (X4)** | **Longitude (X5)** | **House price of unit area (Y)** |
| 1 | 32 | 84.87882 | 10 | 24.98298 | 121.54024 | 37.9 |
| 2 | 19.5 | 306.5947 | 9 | 24.98034 | 121.53951 | 42.2 |
| 3 | 13.3 | 561.9845 | 5 | 24.98746 | 121.54391 | 47.3 |
| 4 | 13.3 | 561.9845 | 5 | 24.98746 | 121.54391 | 54.8 |
| 5 | 5 | 390.5684 | 5 | 24.97937 | 121.54245 | 43.1 |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| 410 | 13.7 | 4082.015 | 0 | 24.94155 | 121.5038 | 15.4 |
| 411 | 5.6 | 90.45606 | 9 | 24.97433 | 121.5431 | 50 |
| 412 | 18.8 | 390.9696 | 7 | 24.97923 | 121.5399 | 40.6 |
| 413 | 8.1 | 104.8101 | 5 | 24.96674 | 121.5407 | 52.5 |
| 414 | 6.5 | 90.45606 | 9 | 24.97433 | 121.5431 | 63.9 |

**Model Regresi sebelum *Pre-Processing Data* dan Transformasi**

Berikut merupakan *output* model regresi dan *R2* dengan bantuan *software* Minitab.

Y = - 6034 - 0.217 X1 - 0.00390 X2 + 1.26 X3 + 224 X4 + 3.9 X5

**R-Sq = 56.9%**

Dari *output* Minitab dapat diketahui bahwa *R2* adalah sebesar 56,9% yang berarti bahwa variabel Y (*house price of unit area)* dapat dijelaskan oleh variabel-variabel X sebesar 56,9%, sisanya yaitu 43,1% dijelaskan oleh variabel diluar model.

1. **Model Regresi setelah *Pre-Processing Data* (*Missing Value)***

Berikut merupakan *output* model regresi dan *R2* dengan bantuan *software* Minitab, setelah dilakukan imputasi pada data yang *missing*

Y = - 3860 - 0.218 X1 - 0.00456 X2 + 1.10 X3 + 230 X4 - 15.2 X5

**R-Sq = 55.1%**

Dari *output* Minitab dapat diketahui bahwa *R2* adalah sebesar 55,1% yang berarti bahwa variabel Y (*house price of unit area)* dapat dijelaskan oleh variabel-variabel X sebesar 55,1%, sisanya yaitu 44,9% dijelaskan oleh variabel diluar model.

1. **Model Regresi setelah *Pre-Processing Data* (*Missing Value*  dan *Outlier)***

Berikut merupakan *output* model regresi dan *R2* dengan bantuan *software* Minitab, setelah dilakukan imputasi pada data yang *missing*  dan membuang pengamatan yang *outlier*

Y = - 7881 - 0.213 X1 - 0.00436 X2 + 1.18 X3 + 224 X4 + 19.2 X5

**R-Sq = 58.3%**

Dari *output* Minitab dapat diketahui bahwa *R2* adalah sebesar 58,3% yang berarti bahwa variabel Y (*house price of unit area)* dapat dijelaskan oleh variabel-variabel X sebesar 58,3%, sisanya yaitu 41,7% dijelaskan oleh variabel diluar model.

1. **Model Regresi setelah *Pre-Processing Data* (*Missing Value*) dan Transformasi Data**

Berikut merupakan *output* model regresi dan *R2* dengan bantuan *software* Minitab, setelah dilakukan imputasi untuk data yang *missing* dan setelah dilakukan transformasi data.

Y = 0.0968 - 0.198 X1 - 0.103 X2 + 0.515 X3 + 196 X4 - 39.9 X5

**R-Sq = 93.1%**

Dari *output* Minitab dapat diketahui bahwa *R2* adalah sebesar 93,1% yang berarti bahwa variabel Y (*house price of unit area)* dapat dijelaskan oleh variabel-variabel X sebesar 93,1%, sisanya yaitu 6,9% dijelaskan oleh variabel diluar model.

1. **Model Regresi setelah *Pre-Processing Data* (*Missing Value, Outlier*) dan Transformasi Data**

Berikut merupakan *output* model regresi dan *R2* dengan bantuan *software* Minitab, setelah dilakukan imputasi untuk data yang missing, penghapusan pengamatan yang *outlier,* dan setelah dilakukan transformasi data.

Y = 0.0803 - 0.193 X1 - 0.0870 X2 + 0.760 X3 + 153 X4 - 31.0 X5

**R-Sq = 94.3%**

Dari *output* Minitab dapat diketahui bahwa *R2* adalah sebesar 94,3% yang berarti bahwa variabel Y (*house price of unit area)* dapat dijelaskan oleh variabel-variabel X sebesar 94,3%, sisanya yaitu 5,7% dijelaskan oleh variabel diluar model.

**KESIMPULAN**

Berdasarkan *R2* yang dihasilkan masing-masing model dapat disimpulkan bahwa data yang sudah melalui tahap *pre-processing* (imputasi data *missing* dan penghapusan pengamatan *outlier*) dan transformasi akan menghasilkan model regresi yang lebih baik. Hal tersebut terbukti dari *R2* data yang sudah melalui tahap *pre-processing* dan transformasi lebih besar dari *R2* data yang belum melalui kedua tahap tersebut.