**KEBAIKAN MODEL REGRESI YANG DIHASILKAN PADA PENGOLAHAN DATA *BREAST CANCER COMBRIA***

*Breast Cancer Combria* atau kanker payudara menjadi pembunuh wanita terbanyak di dunia. Sebagian besar kanker payudara berasal dari sel-sel duktus 86%, lobubar 12%, dan sisanya berasal dari jaringan lain. Terdapat beberapa faktor yang memengaruhi terjadinya kanker payudara, yaitu besar kadar glukosa, insulin, leptin, dan adiponectin, BMI (*Body Mass Index)*, umur, resistin, dan homa. *Data set Breast Cancer Combria* diperoleh dari *website University of California at Irvine (UCI) Machine Learning Respository*. Berdasarkan data tersebut akan dilakukan pemilihan model regresi terbaik dengan melakukan *preprocessing data* terlebih dahulu dan terdapat empat kriteria *preprocessing data* sebagai berikut.

1. *Missing value*
2. *Missing value* dan *outlier*
3. *Missing value* dan transformasi
4. *Missing value*, *outlier*, dan transformasi

Variabel yang terdapat pada data *Breast Cancer Combria* yang digunakan untuk mendapatkan model regresi terbaik adalah sebagai berikut.

Tabel 1.Variabel Penelitian

|  |  |
| --- | --- |
| **Variabel** | **Keterangan** |
| Y | *Glucose* |
| *X1* | *Age* |
| *X2* | BMI |
| *X3* | Insulin |
| *X4* | Leptin |
| *X5* | *Adiponectin* |

Regresi yang didapatkan dengan bantuan *Software Minitab 16* adalah sebagai berikut.

1. Regresi dengan data yang telah diatasi *Missing Value*nya

**Regression Analysis: Glucose versus Age, BMI, ...**

The regression equation is

Glucose = 70.1 + 0.268 Age - 0.031 BMI + 1.00 Insulin + 0.188 Leptin

- 0.181 Adiponectin

Predictor Coef SE Coef T P

Constant 70.15 14.48 4.84 0.000

Age 0.2679 0.1128 2.38 0.019

BMI -0.0308 0.4490 -0.07 0.945

Insulin 1.0032 0.1854 5.41 0.000

Leptin 0.1882 0.1197 1.57 0.119

Adiponectin -0.1811 0.2769 -0.65 0.515

S = 18.9361 R-Sq = 32.4% R-Sq(adj) = 29.3%

1. Regresi dengan data yang telah diatasi *Missing Value* dan *outlier*nya

**Regression Analysis: Glucose versus Age, BMI, ...**

The regression equation is

Glucose = 69.5 + 0.187 Age + 0.257 BMI + 0.529 Insulin + 0.061 Leptin

+ 0.115 Adiponectin

Predictor Coef SE Coef T P

Constant 69.47 11.07 6.28 0.000

Age 0.18663 0.09121 2.05 0.043

BMI 0.2571 0.3550 0.72 0.471

Insulin 0.5293 0.2083 2.54 0.013

Leptin 0.0613 0.1074 0.57 0.569

Adiponectin 0.1150 0.2628 0.44 0.663

S = 14.1522 R-Sq = 12.9% R-Sq(adj) = 8.5%

1. Regresi dengan data yang telah diatasi *Missing Value*nya dan ditransformasikan

**Regression Analysis: Glucose versus Age, BMI, ...**

The regression equation is

Glucose = 0.0235 - 0.594 Age - 0.237 BMI - 0.165 Insulin - 0.309 Leptin

- 0.182 Adiponectin

Predictor Coef SE Coef T P

Constant 0.0234564 0.0004644 50.50 0.000

Age -0.59368 0.01240 -47.88 0.000

BMI -0.23671 0.03054 -7.75 0.000

Insulin -0.16540 0.01757 -9.41 0.000

Leptin -0.30853 0.01155 -26.71 0.000

Adiponectin -0.18205 0.01990 -9.15 0.000

S = 0.000122382 R-Sq = 96.4% R-Sq(adj) = 96.3%

1. Regresi dengan data yang telah diatasi *Missing Value* dan *outlier*nya serta telah ditransformasikan

**Regression Analysis: Glucose versus Age, BMI, ...**

The regression equation is

Glucose = 0.0233 - 0.613 Age - 0.257 BMI - 0.127 Insulin - 0.322 Leptin

- 0.154 Adiponectin

Predictor Coef SE Coef T P

Constant 0.0233281 0.0004474 52.14 0.000

Age -0.61316 0.01138 -53.90 0.000

BMI -0.25655 0.02767 -9.27 0.000

Insulin -0.12705 0.01900 -6.69 0.000

Leptin -0.321999 0.009577 -33.62 0.000

Adiponectin -0.15402 0.02170 -7.10 0.000

S = 0.000105373 R-Sq = 97.2% R-Sq(adj) = 97.0%

Nilai R-Sq terbesar mengindikasikan bahwa model regresi yang dihasilkan semakin baik. Berikut nilai R-Sq yang didapatkan dari hasil regresi yang telah dilakukan.

Tabel 2**.** Nilai R-Sq

|  |  |
| --- | --- |
| **Perlakuan** | R-Sq |
| *Missing value* | 32,4% |
| *Missing value* dan *outlier* | 12,9% |
| *Missing value* dan transformasi | 96,4% |
| ***Missing value, outlier* dan transformasi** | **97,2%** |

Berdasarkan Tabel 2. dapat diketahui bahwa *preprocessing data* dengan penanganan *missing value*, *outlier* dan dilakukannya transformasi data memiliki nilai R-Sq terbesar dibandingkan dengan perlakuan yang lain sehingga diperoleh keputusan bahwa model regresi terbaik yaitu **model regresi dengan penanganan *missing value*, *outlier* dan dilakukannya transformasi data.** Kesimpulannya yaitu semakin baik proses pengolahan data dengan dilakukannya *preprocessing data* yang mencakup *missing value, oulier* dan transformasi akan menghasilkan model regresi yang semakin baik juga.