**MEMBANDINGKAN KEBAIKAN MODEL REGRESI PADA DATA *LIFE EXPEXTANCY***

Data *Life Expectancy* merupakan dataset yang terkait dengan usia harapan hidup dimana terdapat faktor-faktor yang diduga berpengaruh terhadap besarnya harapan hidup. Data tersebut dikumpulkan oleh *World Health Organization (WHO)* dan terdapat dalam web kaggle.com. Berdasarkan data tersebut akan dilakukan pemilihan model regresi terbaik dengan melakukan *preprocessing data* terlebih dahulu dan terdapat empat kriteria *preprocessing data* yaitu :

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

Adapun variabel yang terdapat pada data *Life Expectancy* yang dipilih untuk mendapatkan model regresi terbaik yaitu:

Y: *life expectancy* 🡪 harapan hidup dalam usia

X1 : *adult mortality* 🡪 angka kematian dewasa dari jenis kelamin

X2 : *infant deaths* 🡪 jumlah kematian bayi per 1000 populasi

X3 : *income composition of resources* 🡪 indeks pembangunan manusia dalam hal komposisi pendapatan sumber daya

X4 : *schooling* 🡪 jumlah tahun sekolah

Dengan menggunakan *Minitab* didapatkan model regresi sebagai berikut.

1. ***Missing value***

Life expectancy = 59.7 - 0.0357 Adult Mortality - 0.00641 infant deaths + 16.5 Income composition of resources + 0.438 Schooling

1. ***Missing value* dan *outlier***

Life expectancy = 48.2 - 0.0234 Adult Mortality - 0.0124 infant deaths + 45.7 Income composition of resources - 0.375 Schooling

1. ***Missing value* dantransformasi**

Life expectancy = 0.212 - 0.301 Adult Mortality - 0.169 infant deaths + 32.7 Income composition of resources + 1.48 Schooling

1. ***Missing value, outlier* dantransformasi**

Life expectancy = 0.131 - 0.180 Adult Mortality - 0.00567 infant deaths + 68.0 Income composition of resources + 0.222 Schooling

Membandingkan model regresi untuk mendapatkan model regresi terbaik dapat dilakukan dengan melihat nilai  dari masing-masing model yang ada. Nilai  terbesar mengindikasikan bahwa model regresi yang dihasilkan semakin baik. Berikut nilai yang didapatkan.

**Tabel 1.** Nilai 

|  |  |
| --- | --- |
| **Kriteria** |  |
| *Missing value* | 69,5% |
| *Missing value* dan *outlier* | 79,0% |
| *Missing value* dan transformasi | 90,7% |
| *Missing value, outlier* dan transformasi | **94,9%** |

Berdasarkan Tabel 1. diatas dapat dilihat bahwa *preprocessing data* dengan kriteria penanganan *missing value*, *outlier* dan dilakukannya transformasi data memiliki nilai  terbesar. Sehingga diperoleh keputusan bahwa model regresi terbaik yaitu pada model regresi dengan penanganan *missing value*, *outlier* dan dilakukannya transformasi data. Maka dapat disimpulkan bahwa semakin baik data yang ada (dilakukannya *preprocessing data* yang mencakup *missing value, oulier* dan transformasi) akan menghasilkan model regresi yang semakin baik pula.