Model Matematika yang ada

1. Korelasi Pearson dihitung dengan rumus:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Dimana:

- r_{xy} adalah koefisien korelasi Pearson antara variabel xx x dan yy y
- x_i dan y_i adalah nilai individu
- x bar dan y bar adalah nilai rata-rata

Korelasi berkisar dari -1 hingga 1, dimana:

- 1 menunjukkan korelasi positif sempurna
- -1 menunjukkan korelasi negatif sempurna
- 0 menunjukkan tidak ada korelasi

2. Skewness (kemiringan) adalah ukuran asimetri distribusi data dan dihitung dengan:

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3/n}{(\sum_{i=1}^{n}(x_i - \bar{x})^2/n)^{3/2}}$$

Dimana:

- Skewness positif: distribusi memiliki ekor yang lebih panjang ke kanan
- Skewness negatif: distribusi memiliki ekor yang lebih panjang ke kiri
- Skewness = 0: distribusi simetris sempurna (seperti distribusi normal)

3. Transformasi dilakukan untuk mengurangi skewness (kemiringan distribusi data) agar lebih mendekati distribusi normal.

a. Transformasi Log : Digunakan untuk mengurangi skewness positif dan Penambahan +1 (log1p) dilakukan agar nilai 0 tetap dapat ditransformasikan.

$$x' = \log(x + 1)$$

b. Transformasi Yeo-Johnson

Transformasi ini merupakan perluasan dari Box-Cox, yang bisa digunakan untuk data dengan nilai negatif.

$$y_\lambda(x) = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & \text{jika } \lambda \neq 0, x \geq 0 \\ \log(x+1), & \text{jika } \lambda = 0, x \geq 0 \\ -\frac{(-x+1)^{2-\lambda} - 1}{2-\lambda}, & \text{jika } \lambda \neq 2, x < 0 \\ -\log(-x+1), & \text{jika } \lambda = 2, x < 0 \end{cases}$$

4. Model Regresi Linear Multivariat
Model regresi linear digunakan untuk memprediksi variabel target berdasarkan beberapa fitur.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Dimana

- y: Variabel target (misalnya, harga rumah - medv).
- $x_1, x_2, \ldots, x_n$: Variabel prediktor (fitur input).
- $\beta_0, \beta_1, \ldots, \beta_n$: Koefisien regresi.
- ϵ: Error term (kesalahan prediksi).

5. Estimasi Koefisien dengan Ordinary Least Squares (OLS) digunakan untuk mencari koefisien terbaik yang meminimalkan error:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

OLS meminimalkan jumlah kuadrat residual:

$$\min_\beta \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_\beta \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

6. Evaluasi Model: MSE mengukur rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Dimana semakin kecil MSE, semakin baik modelnya.

7. Residual dan Distribusinya

Residual adalah selisih antara nilai aktual dan prediksi:

$$e_i = y_i - \hat{y}_i$$

Distribusi residual yang baik dalam model regresi linear seharusnya mendekati distribusi normal dengan mean 0. Hal ini bisa diuji menggunakan:

- Shapiro-Wilk Test
- QQ-Plot