PROGRAM STUDI
TEKNIK INFORMATIKA – S1
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO

MATA KULIAH
DATA MINING

Technology vector created by sentavio - www.freepik.com

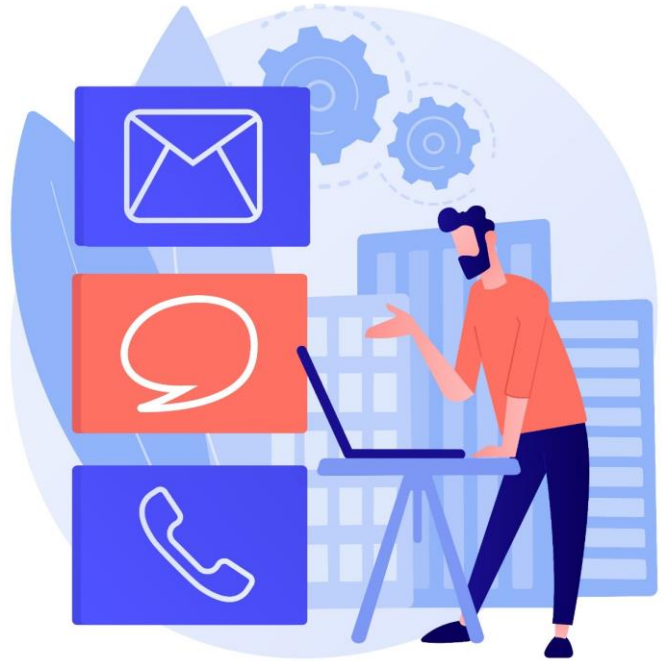
DATA MINING
"Regresi"

TIM PENGAMPU DOSEN DATA MINING

2023

Kontak Dosen

- Junta Zeniarja, M.Kom
- Email: junta@dsn.dinus.ac.id
- Youtube : https://www.youtube.com/JuntaZeniarja
- Scholar : http://bit.do/JuntaScholar



Regresi

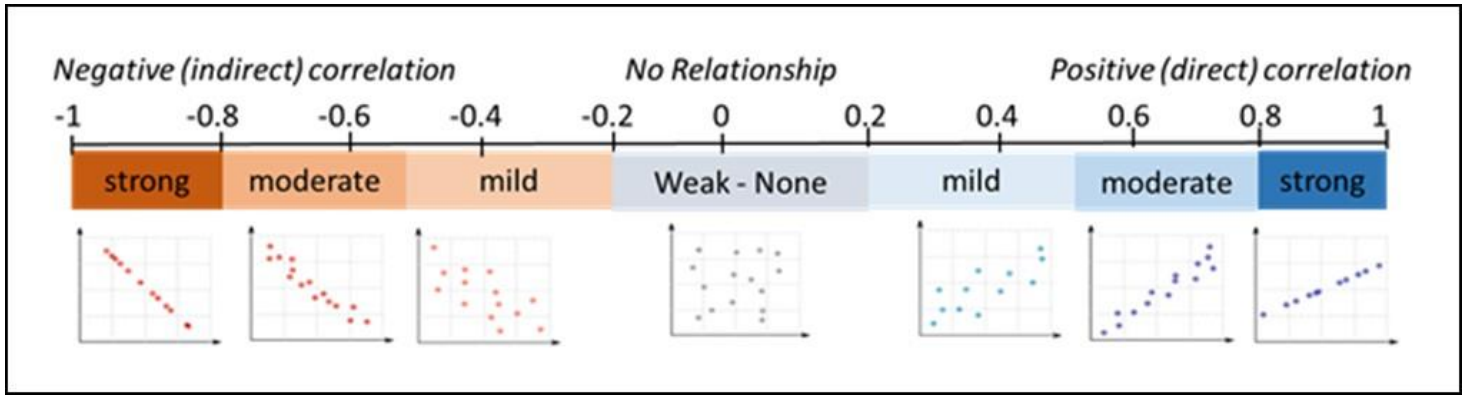
- Regresi dapat digunakan untuk memprediksi nilai dari variabel terikat apabila terdapat kenaikan atau penurunan pada variabel bebas.

- Korelasi digunakan untuk mengukur hubungan antara dua variabel.

Korelasi

- Besaran koefisien -1 & 1 adalah hubungan yang sempurna, sedangkan apabila kedua variabel tidak memiliki hubungan maka nilai korelasi akan bernilai 0.
- Nilai positif mengindikasikan hubungan yang positif, dan sebaliknya.
- Nilai negative menunjukkan bahwa nilai variabel terikat akan meningkat apabila nilai pada variabel bebas mengalami penurunan.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$



Contoh Korelasi

Disini tinggi badan merupakan variabel independen (x), dan berat badan merupakan variabel dependen (y).

Data Tinggi Badan(cm) : 151,174,138,186,128,136,179,163,152,131

Data Berat Badan (Kg) : 63, 81, 56, 91, 47, 57, 76, 72, 62, 48

mean x = (151+174+138+186+128+136+179+163+152+131)/10 = 153.8
mean y = (63+81+56+91+47+57+76+72+62+48)/10 = 65.3

r = (((151-153.8)*(63-65.3)) + ... + ((131-153.8)*(48-65.3))) /
√(((151-153.3)² + ... + (131-153.8)²)*((63-65.3)² + ... +
(48-65,3)²))
= 0.9771295961897941

Catatan : **korelasi 0.97** adalah nilai yang **sangat tinggi**, artinya nilai y benar-benar sangat dipengaruhi oleh nilai x, karena korelasi tinggi maka algoritma Regresi Linier ini cocok digunakan untuk data tersebut.

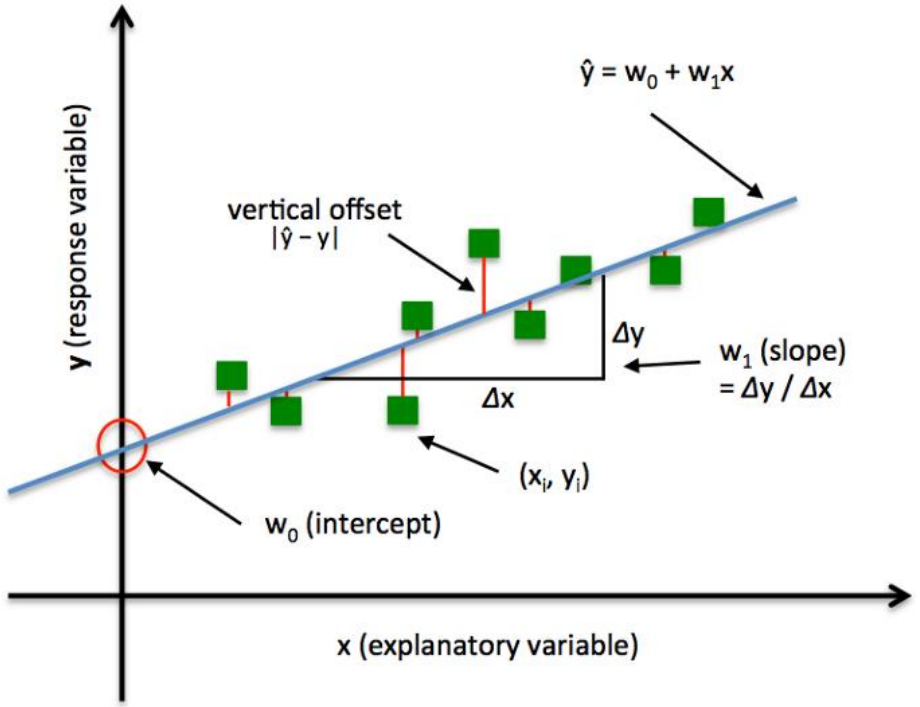
Contoh Korelasi

```
mean x = (151+174+138+186+128+136+179+163+152+131)/10 = 153.8
mean y = (63+81+56+91+47+57+76+72+62+48)/10 = 65.3
a = (((151-153.8)*(63-65.3)) + ... + ((131-153.8)*(48-65.3))) /
((151-153.3) + ... + (131-153.8))²
  = 0.67461045
b = 65.3 - 0.67461045 * 153.8
  = -38.45508707607701
```

maka persamaan garis :

y = 0.67461045 x - 38.45508707607701

Catatan :
Jadi persamaan garis diatas dapat digunakan untuk melakukan prediksi apabila kita memiliki data tinggi badan yang baru, berat badan dapat diperkirakan dengan rumus tersebut, masukkan nilai tinggi baru ke x, maka perkiraan nilai y (berat badan) akan didapat.



$$Slope \quad a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$intercept \quad b = \bar{y} - a\bar{x}$$

Slope adalah tingkat kemiringan garis, intercept adalah jarak titik y pada garis dari titik 0.

Contoh #1

1	10	0.5
2	25	1.25

$$y = c_0 + c_1 x_1$$
$$1.25 = c_0 + c_1 25$$
$$0.5 = c_0 + c_1 10$$

Then, elimination

$$c_0 = 0$$
$$0.75 = c_1 15$$
$$c_1 = 0.05$$

Thus, the general formula is

$$y = 0.05 x_1$$

Contoh #2

Observation, i	Shear Strength, y_i (psi)	Age of Propellant, x_i (weeks)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

$$\beta = \frac{n(\sum_{i=1} x_i y_i) - (\sum_{i=1} x_i)(\sum_{i=1} y_i)}{n(\sum_{i=1} x_i{}^2) - (\sum_{i=1} x_i)^2}$$

$$= \frac{20(528{,}492.64) - ((267.25)(42{,}627.15))}{20(4677.69) - (71{,}422.56)}$$

$$= \frac{-822{,}253}{22{,}131.24} = -37.15$$

$$\alpha = \bar{y} - b\bar{x} = 2{,}131.36 - (-37.15\ )(13.3625) = 2627.821$$

$$y = \alpha + \beta x = 2627.821 - 37.15x$$

Regresi Linier Berganda

Dalam suatu penelitian yang dilakukan terhadap 10 rumah tangga yang dipilih secara acak, diperoleh data pengeluaran untuk pembelian barang-barang tahan lama per minggu (Y), pendapatan per minggu (X1), dan jumlah anggota rumah tangga (X2).

Seandainya suatu rumah tangga mempunyai X_1 dan X_2, masing-masing 11 dan 8. Berapa besarnya nilai Y. Artinya, berapa ratus rupiah rumah tangga yang bersangkutan akan mengeluarkan biaya untuk pembelian barang-barang tahan lama?

Y(Ratusan Rupiah)	23	7	15	17	23	22	10	14	20	19
X_1(Ribuan Rupiah)	10	2	4	6	8	7	4	6	7	6
X_2(Orang)	7	3	2	4	6	5	3	3	4	3

Penyelesaian

Y	X_1	X_2	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
23	10	7	100	49	70	230	161
7	2	3	4	9	6	14	21
15	4	2	16	4	8	60	30
17	6	4	36	16	24	102	68
23	8	6	64	36	48	184	138
22	7	5	49	25	35	154	110
10	4	3	16	9	12	40	30
14	6	3	36	9	18	84	42
20	7	4	49	16	28	140	80
19	6	3	36	9	18	114	57
$\sum Y = 170$	$\sum X_1 = 60$	$\sum X_2 = 40$	$\sum X_1^2 = 406$	$\sum X_2^2 = 182$	$\sum X_1X_2 = 267$	$\sum X_1Y = 1122$	$\sum X_2Y = 737$

$$A = \begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 \\ \sum X_2 & \sum X_2X_1 & \sum X_2^2 \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad H = \begin{bmatrix} \sum Y \\ \sum X_1Y \\ \sum X_2Y \end{bmatrix}$$

$$A = \begin{bmatrix} 10 & 60 & 40 \\ 60 & 406 & 267 \\ 40 & 267 & 182 \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad H = \begin{bmatrix} 170 \\ 1122 \\ 737 \end{bmatrix}$$

$$b = A^{-1}H$$

$$b = \begin{bmatrix} 0{,}919 & -0{,}084 & -0{,}077 \\ -0{,}084 & 0{,}078 & -0{,}095 \\ -0{,}077 & -0{,}095 & 0{,}163 \end{bmatrix} \begin{bmatrix} 170 \\ 1122 \\ 737 \end{bmatrix}$$

$$b = \begin{bmatrix} 5{,}233 \\ 3{,}221 \\ 0{,}451 \end{bmatrix}$$

Dari hasil penghitungan diatas model regresi linier berganda dapat dituliskan sebagai berikut:

$$\hat{Y} = 5{,}233 + 3{,}221X_1 + 0{,}451X_2$$

$\hat{Y} = 5{,}233 + 3{,}221X(11) + 0{,}451X(8)$

$\hat{Y} = 44{,}272$

Ketika suatu rumah tangga memiliki pendapatan perminggu sebesar Rp11.000 dengan anggota rumah tangga sebanyak 8 orang maka pengeluaran untuk pembelian barang-barang tahan lama per minggu sebesar Rp4.427,2 (nilai $\hat{Y}$ dikali 100).

Implementasi Python
(Regresi)

Pertemuan 12

Studi Kasus 1 => Regresi Linear di Python (Data Tinggi & Berat Badan)

```
[1]  # Import Library yang diperlukan Scipy
     from scipy import stats
```

```
[2]  # x = data tinggi badan (cm)
     x = [151,174,138,186,128,136,179,163,152,131]
```

```
[3]  # y = data berat badan (kg)
     y = [63,81,56,91,47,57,76,72,62,48]
```

```
[4]  slope, intercept, r, p, std_err = stats.linregress(x,y)
```

```
[5]  # Buat Fungsi Linear Regresi => y = 0.67461045 x - 38.45508707607701
     def myfunc(x):
       return slope * x + intercept
```

```
[6]  berat_badan = myfunc(151)
```

```
     print(berat_badan)
```

```
63.41109074243812
```

Studi Kasus 2 => Regresi Linear di Python (Data Penjualan.xlsx) [1]

Mempersiapkan Library yang diperlukan :

1. **Numpy** => library yang akan digunakan untuk kebutuhan scientific dan matematika.
2. **Pandas** => library yang digunakan untuk manipulasi data seperti membuat tabel, mengubah dimensi data, mengecek data, dsb. Pandas mampu membaca berbagai format file seperti file .txt, .csv, .tsv, .xls dll.
3. **Matplotlib** => Library yang digunakan untuk membuat grafik plot sesuai kebutuhan.
4. **Sklearn** => library untuk berbagai metode dan algoritma yang digunakan dalam machine learning.

```
[1]  # Mempersiapkan library
     import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import sklearn
```

Studi Kasus 2 => Regresi Linear di Python (Data Penjualan.xlsx) [2]

```
[2]  # Memanggil Dataset
     dataset = pd.read_excel('Data Penjualan.xlsx')
     x = dataset.iloc[:, :-1].values
     y = dataset.iloc[:, 1].values
```

```
[5]  dataset.keys()
```

```
[6]  dataset.shape
```

```
[7]  # Menampilkan isi sebagian dataset
     dataku = pd.DataFrame(dataset)
     dataku.head()
```

```
[8]  # Split Dataset menjadi Training Set dan Testing Set
     from sklearn.model_selection import train_test_split
     x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state=0)
```

	A	B
1	BiayaProduksi	NilaiPenjualan
2	1500	90500
3	1800	89500
4	1900	105000
5	2050	102000
6	2050	90500
7	2100	104500
8	2200	109500
9	2400	150000
10	3050	152000
11	3200	173000
12	3200	153000
13	3500	174500
14	3500	150000
15	3750	198000
16	3750	187000
17	3900	194500
18	4000	200500
19	4000	170500
20	4100	204500
21	4500	224500

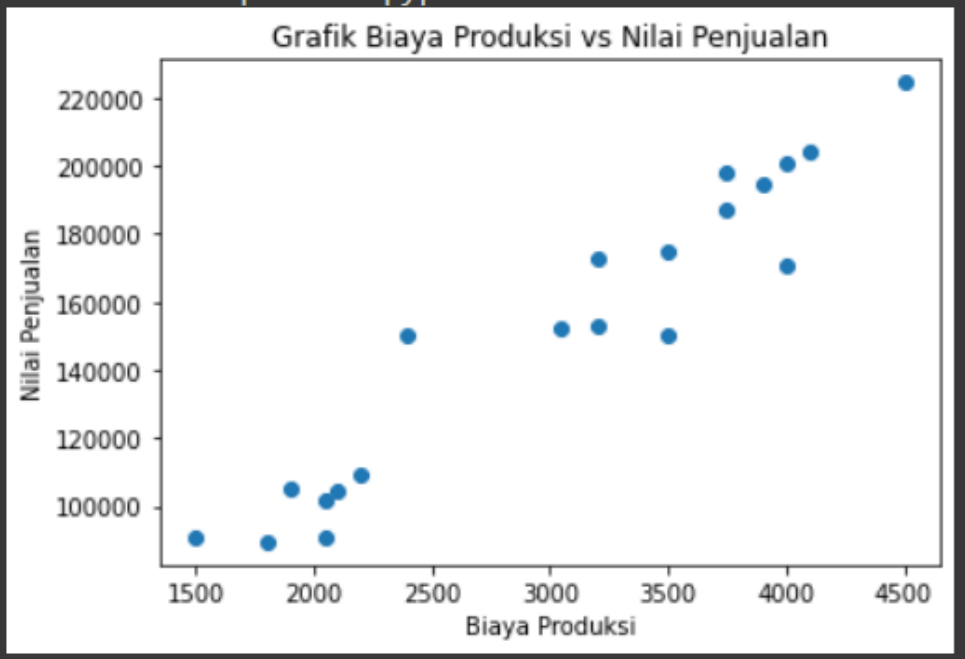
Studi Kasus 2 => Regresi Linear di Python (Data Penjualan.xlsx) [3]

```
[9]  # Melakukan Fitting Simple Linear Regression pada Training Set
     from sklearn.linear_model import LinearRegression
     regressor = LinearRegression()
     regressor.fit(x_train, y_train)
```

```
[10] # Memprediksi hasil Test-Set
     y_pred = regressor.predict(x_test)
```

```
[12] # Visualisasi Data
     plt.scatter(dataku.BiayaProduksi, dataku.NilaiPenjualan)
     plt.xlabel("Biaya Produksi")
     plt.ylabel("Nilai Penjualan")
     plt.title("Grafik Biaya Produksi vs Nilai Penjualan")
     plt.show
```

```
<function matplotlib.pyplot.show>
```



Studi Kasus 2 => Regresi Linear di Python (Data Penjualan.xlsx) [4]

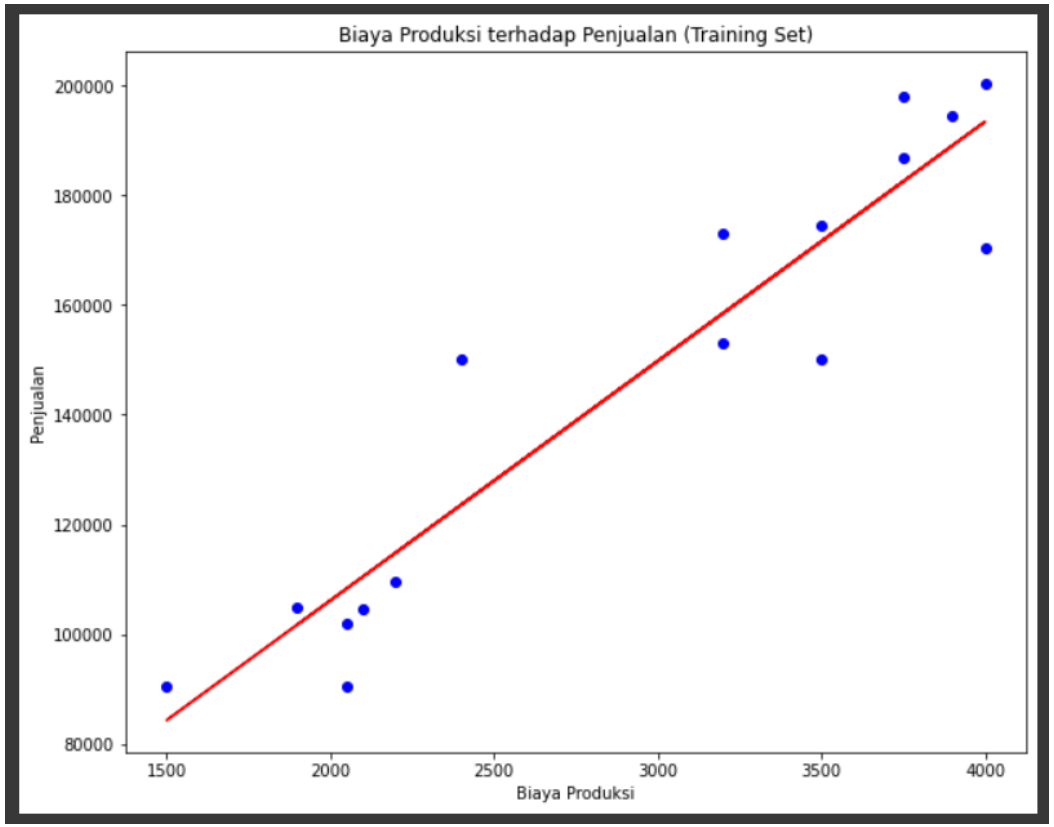
```
[13] # Visualisasi Hasil Prediksi Pada Training-Set
     # Ukuran Plot
     plt.figure(figsize=(10,8))

     # Biru adalah data observasi
     plt.scatter(x_train, y_train, color = 'blue')

     # Garis merah adalah hasil prediksi dari machine learning
     plt.plot(x_train, regressor.predict(x_train), color = 'red')

     # Memberi Judul dan Label
     plt.title('Biaya Produksi terhadap Penjualan (Training Set)')
     plt.xlabel('Biaya Produksi')
     plt.ylabel('Penjualan')

     # Menampilkan plot
     plt.show()
```



Studi Kasus 2 => Regresi Linear di Python (Data Penjualan.xlsx) [5]

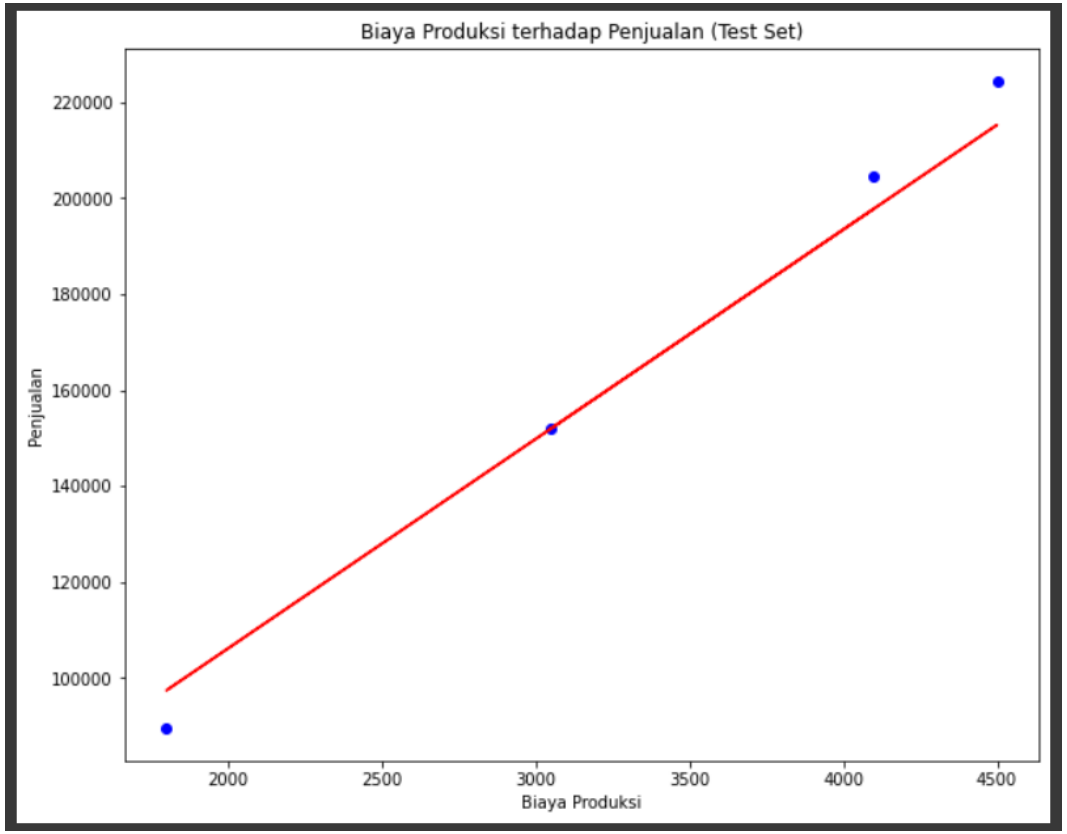
```
[14] # Visualisasi Hasil Prediksi Pada Test-Set
     # Ukuran Plot
     plt.figure(figsize=(10,8))

     # Biru adalah data observasi
     plt.scatter(x_test, y_test, color = 'blue')

     # Garis merah adalah hasil prediksi dari machine learning
     plt.plot(x_test, regressor.predict(x_test), color = 'red')

     # Memberi Judul dan Label
     plt.title('Biaya Produksi terhadap Penjualan (Test Set)')
     plt.xlabel('Biaya Produksi')
     plt.ylabel('Penjualan')

     # Menampilkan plot
     plt.show()
```



Latihan Soal (Kuis)

Tentukan persamaan regresi berikut:

	Usia Mobil (tahun)	Harga Mobil ($100)
	X	y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

Referensi

1. Kusrini, Taufiq Emha, Algoritma Data Mining, *Penerbit Andi*, 2009.
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 4th Edition, *Elsevier*, 2017.
3. https://medium.com/@jrendz/regresi-linier-dengan-r-dan-python-ebb80662c6da.
4. https://statmat.id/regresi-linier-berganda/.
5. Teguh Wahyono, Fundamental of Python for Machine Learning, Penerbit Gava Media, 2018.
6. Sumber gambar: www.freepik.com.

THANKS

ANY QUESTIONS?