

Tugas MK. DATA MINING DAN INFORMATION RETRIEVAL TA 2024-2025

ESTIMASI

LINEAR REGRESSION

Disusun untuk memenuhi salah satu tugas mata kuliah IFB- 307 yang diberikan oleh: Jasman Pardede,
Dr., S.Si., M.T.



Disusun oleh :

Deden Fahrul Roziqin	152022182
Muhammad Yazid	152022192
Hamzah Sabahedthin	152022204
Budi Amin	152022213

Kelas DD

INSTITUT TEKNOLOGI NASIONAL BANDUNG
FAKULTAS TEKNOLOGI INDUSTRI INFORMATIKA
2024/2025

Kata Pengantar

Puji syukur kami panjatkan kehadirat Allah SWT yang Maha Kuasa yang telah melimpahkan karunia dan rahmat-Nya sehingga kami dapat menyusun laporan tugas Estimasi dengan menggunakan regression linear sederhana untuk memenuhi tugas mata kuliah Data Mining dan Information Retrieval.

Dalam penyusunan laporan ini, tidak lepas dari adanya dukungan dan bantuan dari berbagai pihak. Oleh karena itu, kami mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan bantuan, terutama kepada:

1. Bapak Jasman Pardede, Dr., S.Si., M.T. selaku Dosen mata kuliah Data Mining Dan Information Retrieval.
2. Dan teman-teman semua yang telah membantu dalam memberikan saran dalam penyusunan laporan akhir ini.

Kami sebagai penulis dan penyusun menyadari bahwa laporan ini masih jauh dari kata sempurna.

Oleh karena itu kritik dan saran yang bersifat membangun dari semua pihak sangat diharapkan. Agar kedepannya laporan ini dapat menjadi lebih baik lagi. Kami berharap, semoga laporan ini dapat bermanfaat baik bagi kami pada khususnya dan para pembaca pada umumnya.

Bandung, 8 Oktober 2024

Penyusun

DAFTAR ISI

Kata Pengantar.....	1
BAB I PENDAHULUAN.....	3
1.1. Latar Belakang.....	3
1.2. Tujuan.....	3
1.3. Rumusan Masalah.....	3
1.4. Batasan Masalah.....	4
BAB II LANDASAN TEORI.....	5
2.1. Regresi Linear.....	5
2.1.1. Sejarah.....	5
2.1.2. Formula.....	5
2.2. Python.....	7
BAB III STUDI KASUS DAN IMPLEMENTASI.....	8
3.1. Studi Kasus.....	8
3.2. Implementasi Program.....	8
BAB IV PENUTUP.....	12
4.1. Kesimpulan.....	12
4.2. Saran.....	12
DAFTAR PUSTAKA.....	13

BAB I PENDAHULUAN

1.1. Latar Belakang

Pengalaman kerja sering dianggap sebagai salah satu faktor penting yang mempengaruhi produktivitas dan kinerja individu dalam sebuah organisasi, termasuk dalam bidang penjualan. Setiap individu yang bergabung dalam sebuah organisasi memiliki tujuan dan harapan yang berbeda, salah satunya adalah untuk mendapatkan kompensasi finansial dari pekerjaan yang dilakukan. Pengalaman kerja diyakini dapat meningkatkan keterampilan dan pengetahuan karyawan, yang pada akhirnya dapat berdampak positif terhadap kinerja, seperti peningkatan omzet penjualan.

Dalam konteks ini, teori regresi linear sederhana menjadi alat analisis yang relevan untuk mengukur hubungan antara variabel bebas (pengalaman kerja) dan variabel terikat (omzet penjualan). Regresi linear sederhana adalah metode statistik yang digunakan untuk memodelkan hubungan linear antara dua variabel. Dengan menggunakan pendekatan ini, kita dapat mengevaluasi sejauh mana perubahan dalam pengalaman kerja (X) memengaruhi perubahan dalam omzet penjualan (Y). Model regresi menghasilkan persamaan yang menunjukkan besarnya pengaruh pengalaman kerja terhadap omzet penjualan dan memungkinkan prediksi omzet berdasarkan tingkat pengalaman kerja.

1.2. Tujuan

1. Mengevaluasi pengaruh pengalaman kerja terhadap omzet penjualan
2. Mengembangkan Model Estimasi Omzet Penjualan Berdasarkan Pengalaman Kerja
3. Memberikan Estimasi yang Relevan untuk Pengambilan Keputusan

1.3. Rumusan Masalah

1. Apa pengaruh pengalaman kerja terhadap omzet penjualan?
2. Bagaimana cara mengembangkan model estimasi omzet penjualan berdasarkan pengalaman kerja?
3. Sejauh mana estimasi omzet penjualan yang dihasilkan dapat digunakan dalam pengambilan keputusan bisnis?

1.4.Batasan Masalah

1. Penelitian ini menggunakan data yang terdiri pengalaman kerja karyawan dalam satuan tahun dan omzet penjualan ribuan rupiah, yang diperoleh dari enam karyawan di perusahaan.
2. Model estimasi yang dikembangkan hanya akan mempertimbangkan pengalaman kerja sebagai variabel independen dan tidak akan mempertimbangkan faktor lain seperti pendidikan, pelatihan, atau kondisi pasar yang dapat mempengaruhi omzet penjualan.
3. Data yang digunakan dalam penelitian ini terbatas pada tabel yang berisi pengalaman kerja dan omzet penjualan enam karyawan, tanpa mempertimbangkan data dari karyawan lain atau periode waktu yang lebih luas. estimasi akan dilakukan untuk karyawan baru yang memiliki pengalaman kerja selama 6,5 tahun.

BAB II LANDASAN TEORI

2.1. Regresi Linear

2.1.1. Sejarah

Regresi linear pertama kali diperkenalkan oleh Sir Francis Galton pada akhir abad ke-19. Galton menggunakan istilah "regresi" untuk menggambarkan fenomena statistik di mana keturunan dari individu yang sangat tinggi atau sangat pendek cenderung lebih mendekati rata-rata tinggi populasi.

Karl Pearson, seorang murid Galton, mengembangkan lebih lanjut konsep ini dengan memperkenalkan koefisien korelasi dan metode kuadrat terkecil (least squares method) untuk mengestimasi parameter regresi. Metode ini menjadi dasar dari analisis regresi linear yang kita kenal sekarang. dalam data mining, regresi linear digunakan untuk memprediksi nilai dari variabel dependen berdasarkan satu atau lebih variabel independen. Ini sangat berguna dalam berbagai aplikasi seperti prediksi penjualan, analisis risiko, dan pengenalan pola.

Dengan kemajuan teknologi komputer, regresi linear kini dapat diterapkan pada dataset yang sangat besar dan kompleks. Algoritma ini juga telah diintegrasikan ke dalam berbagai perangkat lunak data mining dan analisis statistik seperti RapidMiner, R, dan Python. regresi linear juga menjadi dasar bagi banyak algoritma pembelajaran mesin lainnya. Misalnya, regresi linear berganda, regresi ridge, dan regresi lasso adalah variasi dari regresi linear yang digunakan untuk menangani masalah multikolinearitas dan overfitting dalam data.

2.1.2. Formula

1. Formula untuk menghitung koefisien konstanta(a)

$$a = \frac{(\sum y) (\sum x^2) - (\sum x) (\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Keterangan :

$\sum Y$ = jumlah dari semua data Y

$\sum X^2$ = jumlah dari semua data X dikuadratkan(2)

$\sum X$ = jumlah dari semua data X

$\sum XY$ = jumlah dari hasil perkalian setiap data X dan setiap data Y

n = jumlah banyaknya data

2. Formula untuk menghitung koefisien regresi (b)

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Keterangan :

$\sum Y$ = jumlah dari semua data Y

$\sum X^2$ = jumlah dari semua data X dikuadratkan(2)

$\sum X$ = jumlah dari semua data X

$\sum XY$ = jumlah dari hasil perkalian setiap data X dan setiap data Y

n = jumlah banyaknya data

3. Persamaan Regresi

$$Y = a + bX$$

Keterangan :

a = Konstanta persamaan nilai a

b = Regresi persamaan nilai b

bX = nilai X yang diberikan

4. Koefisien Determinasi

$$R^2 = \frac{((n)(\sum XY) - (\sum X)(\sum Y))^2}{(n(\sum X^2) - (\sum X)^2)(n(\sum Y^2) - (\sum Y)^2)}$$

Keterangan :

n= Jumlah banyaknya data

$\sum XY$ = jumlah dari hasil perkalian setiap data X dan setiap data Y

$\sum X^2$ = jumlah dari semua data X dikuadratkan(2)

$\sum Y$ = Jumlah semua dari data Y

2.2. Python

Python adalah bahasa pemrograman tingkat tinggi yang diciptakan oleh Guido van Rossum dan dirilis pertama kali pada tahun 1991. Dirancang dengan fokus pada kemudahan membaca dan menulis kode, Python menjadi pilihan utama di kalangan pengembang dan ilmuwan data. Sintaksis yang sederhana dan intuitif membuatnya mudah dipahami, baik oleh pemula maupun pengembang berpengalaman. salah satu kekuatan Python terletak pada ekosistem pustakanya yang luas. Pustaka-pustaka seperti NumPy untuk komputasi numerik, Pandas untuk manipulasi data, dan Scikit-learn untuk pembelajaran mesin memberikan alat yang kuat untuk analisis data dan pengembangan model prediktif. Dengan dukungan komunitas yang besar, pengguna dapat dengan mudah menemukan sumber daya dan bantuan.

Python juga mendukung pemrograman fungsional dan modular, memungkinkan pengembang untuk menulis kode yang terorganisir dan dapat digunakan kembali. Fleksibilitas ini menjadikannya ideal untuk berbagai aplikasi, termasuk pengembangan web, otomatisasi, dan analisis data. Dengan semua keunggulannya, Python telah menjadi bahasa utama dalam bidang data science dan pembelajaran mesin, memfasilitasi eksplorasi dan pengembangan model analitis yang efektif.

BAB III STUDI KASUS DAN IMPLEMENTASI

3.1. Studi Kasus

Kami memiliki tabel data pengalaman kerja karyawan dan omzet penjualan setiap karyawan. Disini terdapat tabel yang berisikan data karyawan, pengalaman kerja karyawan dengan satuan tahun dan omzet penjualan dari setiap karyawan. Disini kami ingin mencari estimasi omzet penjualan dari karyawan baru yang bekerja selama 6.5 tahun.

Karyawan	Pengalaman	Omzet
	Kerja	Penjualan
1	10	8
2	8	6
3	7	5
4	4	3
5	3	1
6	6.5	dicari

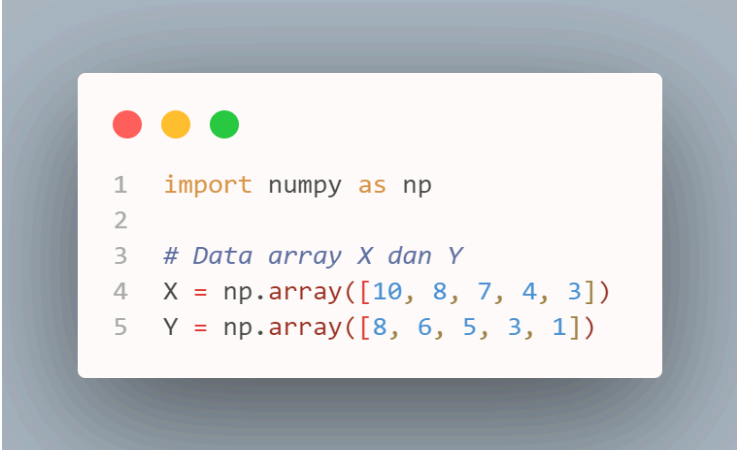
3.2. Implementasi Program

Disini kami menggunakan bahasa pemrograman python untuk mengimplementasikan solusi dari studi kasus yang kami pilih. Berikut merupakan program yang telah kami buat.

```
1 import numpy as np
2
3 # Data array X dan Y
4 X = np.array([10, 8, 7, 4, 3])
5 Y = np.array([8, 6, 5, 3, 1])
6
7 # Jumlah data
8 n = len(X)
9
10 # Menghitung sum X, sum Y, sum X^2, sum Y^2, dan sum XY
11 jumlah_X = np.sum(X)
12 jumlah_Y = np.sum(Y)
13 jumlah_X2 = np.sum(X**2)
14 jumlah_Y2 = np.sum(Y**2)
15 jumlah_XY = np.sum(X * Y)
16
17 # Menghitung mean (rata-rata) dari X dan Y
18 rata_X = np.mean(X)
19 rata_Y = np.mean(Y)
20
21 # Menghitung slope (b) dan intercept (a)
22 b_pembilang = n * jumlah_XY - jumlah_X * jumlah_Y
23 b_penyebut = n * jumlah_X2 - jumlah_X**2
24 b = b_pembilang / b_penyebut
25 a = rata_Y - b * rata_X
26
27 # Menghitung koefisien determinasi (R^2)
28 r_pembilang = (n * jumlah_XY - jumlah_X * jumlah_Y) ** 2
29 r_penyebut = (n * jumlah_X2 - jumlah_X**2) * (n * jumlah_Y2 - jumlah_Y**2)
30 R_kuadrat = r_pembilang / r_penyebut
```

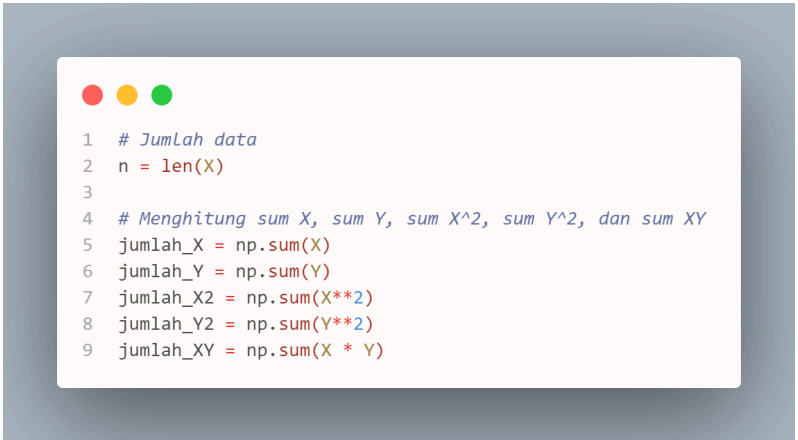
```
1 # Menghitung nilai persentase koefisien determinasi
2 persentase_R = R_kuadrat * 100
3
4 # Menambahkan nilai X input untuk menghitung Y
5 X_input = 6.5 # Nilai X yang diinput
6 Y_output = a + b * X_input # Menghitung Y berdasarkan persamaan regresi
7
8 # Menampilkan hasil rata-rata
9 print("1. Nilai rata-rata:")
10 print(f" Rata-rata X (X̄) = {rata_X:.3f}")
11 print(f" Rata-rata Y (Ȳ) = {rata_Y:.3f}")
12
13 # Menampilkan perhitungan nilai a dan b
14 print("2. Menghitung Nilai a (konstanta) dan b (koefisien regresi):")
15 print(f" Maka, b = {b:.3f} dan a = {a:.3f}")
16 print(f" Persamaan Regresi: Y = {a:.3f} + {b:.3f}X")
17
18 # Menampilkan nilai Y berdasarkan input X
19 print(f" Untuk X = {X_input}, maka Y = {Y_output:.3f}")
20
21 # Menampilkan perhitungan koefisien determinasi
22 print("3. Menghitung Koefisien Determinasi (R^2):")
23 print(
24     f" R^2 = {(n * jumlah_XY - jumlah_X * jumlah_Y) ** 2} / "
25     f" {(n * jumlah_X2 - jumlah_X**2) * (n * jumlah_Y2 - jumlah_Y**2)} "
26 )
27 print(f" = {(r_pembilang / r_penyebut):.3f}")
28 print(f" = {R_kuadrat:.5f}")
29
30 # Menampilkan nilai persentase koefisien determinasi
31 print(f" Persentase Koefisien Determinasi = {persentase_R:.2f}%")
32
```

Disini kami menggunakan library numpy dari python untuk mengolah data karyawan menggunakan array. Pertama kami mengimport library numpy lalu membuat array untuk menyimpan data pengalaman kerja yang disimpan di variabel x dan omzet penjualan yang disimpan di variabel y.



```
1 import numpy as np
2
3 # Data array X dan Y
4 X = np.array([10, 8, 7, 4, 3])
5 Y = np.array([8, 6, 5, 3, 1])
```

Lalu kami membuat variabel n untuk menyimpan panjang data. Lalu kami membuat variabel jumlah dan menggunakan rumus sum dari numpy untuk menyimpan nilai total dari setiap kolom ke dalam variabel tersebut.



```
1 # Jumlah data
2 n = len(X)
3
4 # Menghitung sum X, sum Y, sum X^2, sum Y^2, dan sum XY
5 jumlah_X = np.sum(X)
6 jumlah_Y = np.sum(Y)
7 jumlah_X2 = np.sum(X**2)
8 jumlah_Y2 = np.sum(Y**2)
9 jumlah_XY = np.sum(X * Y)
```

Lalu kami membuat variabel rata_X dan rata_Y untuk menyimpan rata-rata kolom dan kolom y menggunakan rumus mean dari numpy. Kami menghitung nilai a dan b sesuai dengan formula dari regresi linear. Untuk nilai b kami bagi menjadi dua bagian terlebih dahulu yaitu b_pembilang dan b_penyebut lalu dari kedua variabel tersebut kami gunakan untuk menghitung nilai b.

```

1 # Menghitung mean (rata-rata) dari X dan Y
2 rata_X = np.mean(X)
3 rata_Y = np.mean(Y)
4
5 # Menghitung slope (b) dan intercept (a)
6 b_pembilang = n * jumlah_XY - jumlah_X * jumlah_Y
7 b_penyebut = n * jumlah_X2 - jumlah_X**2
8 b = b_pembilang / b_penyebut
9 a = rata_Y - b * rata_X

```

Lalu kami menghitung koefisien determinasi R^2 dengan cara membaginya menjadi dua variabel yaitu $r_{\text{pembilang}}$ dan r_{penyebut} . Setelah mendapatkan nilai dari $r_{\text{pembilang}}$ dan r_{penyebut} , kami menggunakan kedua nilai dari variabel tersebut untuk menghitung koefisien determinasi yang nilainya disimpan dalam variabel R_{kuadrat} . Setelah itu kami menghitung nilai persentase koefisien determinasi menggunakan nilai dari variabel R_{kuadrat} dan nilainya disimpan di variabel persentase_R . Lalu kami menghitung output dari Y menggunakan nilai a dan nilai b yang sebelumnya sudah dihitung dan nilai x yang didefinisikan dengan variabel X_{input} yang nilainya berasal dari studi kasus kami yaitu 6.5 tahun.

```

1 # Menghitung koefisien determinasi ( $R^2$ )
2 r_pembilang = (n * jumlah_XY - jumlah_X * jumlah_Y) ** 2
3 r_penyebut = (n * jumlah_X2 - jumlah_X**2) * (n * jumlah_Y2 - jumlah_Y**2)
4 R_kuadrat = r_pembilang / r_penyebut
5
6 # Menghitung nilai persentase koefisien determinasi
7 persentase_R = R_kuadrat * 100
8
9 # Menambahkan nilai X input untuk menghitung Y
10 X_input = 6.5 # Nilai X yang diinput
11 Y_output = a + b * X_input # Menghitung Y berdasarkan persamaan regresi

```

Lalu kami menampilkan semua langkah langkah dan hasil dari operasi regresi linear kami mulai dari hasil rata-rata x dan y, perhitungan nilai a dan b, hasil dari nilai y berdasarkan input x, perhitungan koefisien determinasi dan persentase dari koefisien determinasi.

```

1 # Menampilkan hasil rata-rata
2 print("1. Nilai rata-rata:")
3 print(f" Rata-rata X ( $\bar{X}$ ) = {rata_X:.3f}")
4 print(f" Rata-rata Y ( $\bar{Y}$ ) = {rata_Y:.3f}\n")
5
6 # Menampilkan perhitungan nilai a dan b
7 print("2. Menghitung Nilai a (konstanta) dan b (koefisien regresi):")
8 print(f" Maka, b = {b:.3f} dan a = {a:.3f}\n")
9 print(f" Persamaan Regresi: Y = {a:.3f} + {b:.3f}X\n")
10
11 # Menampilkan nilai Y berdasarkan input X
12 print(f" Untuk X = {X_input}, maka Y = {Y_output:.3f}\n")
13
14 # Menampilkan perhitungan koefisien determinasi
15 print("3. Menghitung Koefisien Determinasi ( $R^2$ ):")
16 print(
17     f"  $R^2 = \frac{[(\sum \{jumlah\_XY\}) - (\sum \{jumlah\_X\})(\sum \{jumlah\_Y\})]^2}{[\sum \{(\sum \{jumlah\_X^2\}) - (\sum \{jumlah\_X\})^2][\sum \{(\sum \{jumlah\_Y^2\}) - (\sum \{jumlah\_Y\})^2\}]}$  "
18 )
19 print(f" = [{r_pembilang}] / [{r_penyebut}]\n")
20 print(f" = {R_kuadrat:.5f}\n")
21
22 # Menampilkan nilai persentase koefisien determinasi
23 print(f" Persentase Koefisien Determinasi = {persentase_R:.2f}%")

```

Berikut merupakan output dari program yang telah kami buat.

```

1. Nilai rata-rata:
Rata-rata X ( $\bar{X}$ ) = 6.400
Rata-rata Y ( $\bar{Y}$ ) = 4.600

2. Menghitung Nilai a (konstanta) dan b (koefisien regresi):
Maka, b = 0.928 dan a = -1.337

Persamaan Regresi: Y = -1.337 + 0.928X

Untuk X = 6.5, maka Y = 4.693

3. Menghitung Koefisien Determinasi ( $R^2$ ):
 $R^2 = \frac{[5(178) - (32)(23)]^2}{[(5)(238) - (32)^2][(5)(135) - (23)^2]}$ 
=  $\frac{[23716]}{[24236]}$ 
= 0.97854

Persentase Koefisien Determinasi = 97.85%

```

BAB IV PENUTUP

4.1. Kesimpulan

Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa pengalaman kerja memiliki pengaruh yang signifikan terhadap omzet penjualan karyawan. model estimasi yang dikembangkan menggunakan data pengalaman kerja dan omzet penjualan dari satu karyawan baru yang memiliki pengalaman kerja selama 6,5 tahun.

Meskipun hasil estimasi ini dapat menjadi acuan dalam pengambilan keputusan, penting untuk diingat bahwa analisis ini dibatasi pada jumlah data yang kecil dan tidak mempertimbangkan faktor lain yang mungkin mempengaruhi omzet.

4.2. Saran

1. Disarankan untuk mengumpulkan data lebih banyak, termasuk faktor-faktor lain yang mempengaruhi omzet penjualan agar model dapat menjadi lebih akurat dan representatif.
2. Mempertimbangkan variabel lain untuk memberikan pemahaman yang lebih komprehensif mengenai faktor yang mempengaruhi omzet.
3. Melakukan uji dan evaluasi model secara berkala untuk memastikan bahwa model tetap relevan dengan kondisi pasar yang berubah.

DAFTAR PUSTAKA

http://eprints.undip.ac.id/45641/1/06_DIVIANI.pdf

[Python \(bahasa pemrograman\) - Wikipedia bahasa Indonesia, ensiklopedia bebas](#)

[Algoritma Estimasi dalam Data Mining: Linear Regression - Flin Setyadi](#)

https://www.youtube.com/watch?v=k0bfMzkXTiA&ab_channel=tugixline