EVALUASI DAN VALIDASI METODE NUMERIK

Pengertian Evaluasi dan Validasi

Dalam analisis data dan metode numerik, proses evaluasi dan validasi memiliki peran yang sangat penting. Setelah suatu model atau metode diterapkan, analis perlu memastikan bahwa hasil yang diperoleh benar-benar mampu merepresentasikan kondisi data sebenarnya.

Evaluasi digunakan untuk mengukur kualitas hasil analisis, sedangkan validasi digunakan untuk memastikan bahwa model memiliki performa yang stabil ketika diterapkan pada data lain.

Pada metode Principal Component Analysis (PCA), evaluasi diperlukan karena PCA melakukan reduksi dimensi data. Ketika jumlah variabel dikurangi, terdapat kemungkinan sebagian informasi ikut hilang. Oleh sebab itu diperlukan pengukuran untuk mengetahui:

- Seberapa besar informasi yang berhasil dipertahankan.
- Berapa jumlah komponen utama terbaik.
- Apakah model stabil.
- Apakah hasil analisis dapat dipercaya.

Dalam praktik data science, evaluasi dan validasi sangat penting karena:

1. Mengurangi risiko kesalahan analisis.
2. Membantu memilih model terbaik.
3. Menghindari overfitting.
4. Memastikan hasil analisis konsisten.
5. Meningkatkan kualitas pengambilan keputusan.

HUBUNGAN EVALUASI DENGAN PCA

Pada PCA, data dengan banyak variabel disederhanakan menjadi beberapa komponen utama.

Sebagai contoh:

Sebuah supermarket memiliki data:

- Jumlah pengunjung
- Biaya promosi
- Jumlah produk
- Pendapatan harian
- Lama antrean
- Jumlah transaksi

Seluruh variabel tersebut saling berkaitan.

Jika seluruh variabel dianalisis secara langsung:

- Proses analisis menjadi lebih kompleks.
- Visualisasi sulit dilakukan.
- Komputasi lebih berat.

Karena itu digunakan PCA untuk mengubah variabel asli menjadi komponen utama.

Namun setelah reduksi dimensi dilakukan, perusahaan perlu mengetahui:

- Apakah informasi penting masih dipertahankan?
- Apakah hanya 1 atau 2 komponen sudah cukup?
- Apakah hasil PCA konsisten?

Pertanyaan tersebut dijawab melalui evaluasi dan validasi.

KONSEP PENGUKURAN KESALAHAN

1. Variance Explained

Variance explained menunjukkan seberapa besar variasi data yang berhasil dijelaskan oleh suatu komponen utama.

Semakin besar nilai variance explained, maka semakin penting komponen tersebut.

Contoh:

Komponen	Variance Explained
PC1	90%
PC2	7%
PC3	2%
PC4	1%

Interpretasi:

- PC1 menyimpan sebagian besar informasi data.
- Komponen lainnya hanya menyimpan sedikit informasi.

2. Cumulative Variance

Cumulative variance merupakan total informasi yang dijelaskan oleh beberapa komponen sekaligus.

Contoh:

Komponen	Cumulative Variance
PC1	90%

Komponen	Cumulative Variance
PC1 + PC2	97%
PC1 + PC2 + PC3	99%

Jika cumulative variance sudah lebih dari 80% atau 90%, maka jumlah komponen dianggap cukup baik.

3. Reconstruction Error

Reconstruction error digunakan untuk mengukur seberapa besar perbedaan antara:

- Data asli
- Data hasil rekonstruksi PCA

Semakin kecil error:

- PCA semakin baik.
- Informasi yang hilang semakin sedikit.

4. Cross-validation

Cross-validation digunakan untuk menguji kestabilan model.

Data dibagi menjadi:

- Data training
- Data testing

Tujuannya:

- Mengurangi overfitting
- Menguji konsistensi model
- Memastikan hasil PCA stabil

STUDI KASUS

ANALISIS EVALUASI DAN VALIDASI PCA PADA PENJUALAN SUPERMARKET

Latar Belakang Kasus

Sebuah perusahaan supermarket nasional memiliki 25 cabang yang tersebar di beberapa kota besar.

Manajemen perusahaan ingin memahami faktor utama yang memengaruhi peningkatan pendapatan cabang.

Selama ini perusahaan hanya melihat pendapatan akhir tanpa memahami hubungan antar variabel.

Padahal terdapat banyak faktor yang saling berkaitan, seperti:

- Jumlah pengunjung
- Besarnya biaya promosi
- Jumlah produk tersedia
- Pendapatan harian

Manajemen mengalami beberapa masalah:

1. Data memiliki banyak variabel.
2. Variabel saling berkorelasi.
3. Sulit mengetahui faktor paling dominan.
4. Sulit melakukan visualisasi data.
5. Analisis menjadi lebih kompleks.

Untuk mengatasi masalah tersebut digunakan Principal Component Analysis (PCA).

Namun perusahaan juga ingin memastikan bahwa hasil PCA benar-benar valid dan dapat dipercaya.

Karena itu dilakukan:

- Evaluasi PCA
- Pengukuran kesalahan
- Validasi model
- Cross-validation

DATASET

Data Penjualan Supermarket

Cabang	Pengunjung	Promosi	Produk	Pendapatan
1	230	15	420	58
2	250	16	430	61
3	245	18	440	62
4	260	20	450	65
5	280	22	470	70
6	290	24	480	72
7	300	25	490	75
8	310	27	500	77
9	320	28	510	80

Cabang	Pengunjung	Promosi	Produk	Pendapatan
10	330	29	520	82
11	340	30	530	84
12	350	31	540	86
13	360	32	550	88
14	370	34	560	90
15	380	35	570	92
16	390	36	580	94
17	400	38	590	96
18	410	39	600	98
19	420	40	610	100
20	430	42	620	102
21	440	43	630	104
22	450	45	640	106
23	460	46	650	108
24	470	47	660	110
25	480	48	670	112

LANGKAH PENYELESAIAN DENGAN R

1. Membuat Dataset

```
# Membuat data supermarket
supermarket <- data.frame(
  Pengunjung = c(230,250,245,260,280,
                 290,300,310,320,330,
                 340,350,360,370,380,
                 390,400,410,420,430,
                 440,450,460,470,480),

  Promosi = c(15,16,18,20,22,
              24,25,27,28,29,
              30,31,32,34,35,
              36,38,39,40,42,
              43,45,46,47,48),

  Produk = c(420,430,440,450,470,
             480,490,500,510,520,
             530,540,550,560,570,
             580,590,600,610,620,
             630,640,650,660,670),
```

```
  Pendapatan = c(58,61,62,65,70,
                 72,75,77,80,82,
                 84,86,88,90,92,
                 94,96,98,100,102,
                 104,106,108,110,112)
)

# Menampilkan data
supermarket
```

Penjelasan:

Data disusun dalam bentuk data frame.

Setiap kolom merepresentasikan variabel berbeda:

- Pengunjung → jumlah pelanggan harian.
- Promosi → biaya promosi.
- Produk → jumlah produk tersedia.
- Pendapatan → total pendapatan cabang.

Dataset ini akan digunakan untuk seluruh proses evaluasi dan validasi PCA.

---

2. Memahami Struktur Data

```
# Struktur data
str(supermarket)

# Ringkasan statistik
summary(supermarket)
```

Fungsi str() digunakan untuk melihat:

- Nama variabel
- Tipe data
- Jumlah observasi

Sedangkan summary() digunakan untuk melihat:

- Nilai minimum
- Nilai maksimum
- Mean
- Median
- Quartile

Tahapan ini penting untuk memahami karakteristik data sebelum analisis dilakukan.

---

3. Standardisasi Data

```
# Standardisasi data
scaled_data <- scale(supermarket)

# Menampilkan hasil
head(scaled_data)
```

Pada PCA, standardisasi sangat penting karena setiap variabel memiliki skala berbeda.

Contoh:

- Pengunjung → ratusan
- Promosi → puluhan
- Produk → ratusan
- Pendapatan → ratusan

Jika data tidak distandardisasi:

- Variabel dengan nilai besar akan mendominasi.
- Hasil PCA menjadi bias.

Fungsi scale() melakukan:

1. Mean menjadi 0.
2. Standar deviasi menjadi 1.

---

4. Membuat Matriks Korelasi

```
# Matriks korelasi
correlation_matrix <- cor(scaled_data)

# Menampilkan matriks korelasi
correlation_matrix
```

Matriks korelasi digunakan untuk melihat hubungan antar variabel.

Interpretasi korelasi:

Nilai Korelasi	Interpretasi
Mendekati 1	Hubungan sangat kuat
Mendekati 0	Tidak ada hubungan
Mendekati -1	Hubungan berlawanan

Jika variabel memiliki korelasi tinggi, maka PCA sangat cocok digunakan.

---

5. Menghitung Nilai Eigen

```
# Menghitung eigen
result_eigen <- eigen(correlation_matrix)

# Menampilkan hasil
result_eigen
```

Fungsi eigen() digunakan untuk menghitung:

- Nilai eigen
- Vektor eigen

Nilai eigen menunjukkan seberapa besar informasi yang disimpan suatu komponen.

Semakin besar nilai eigen:

- Semakin penting komponen tersebut.

---

6. Menampilkan Nilai Eigen

```
# Menampilkan nilai eigen
result_eigen$values
```

Interpretasi:

- Komponen dengan nilai eigen terbesar dianggap paling dominan.
- Komponen dengan nilai eigen kecil biasanya kurang penting.

Dalam PCA biasanya digunakan aturan:

- Nilai eigen > 1 → komponen penting.
- Nilai eigen < 1 → komponen kurang penting.

---

7. Menampilkan Vektor Eigen

```
# Menampilkan vektor eigen
result_eigen$vectors
```

Vektor eigen menunjukkan kontribusi setiap variabel terhadap komponen utama.

Interpretasi:

- Nilai besar → kontribusi tinggi.
- Nilai kecil → kontribusi rendah.

Melalui vektor eigen perusahaan dapat mengetahui variabel mana yang paling memengaruhi pendapatan.

PENERAPAN PCA

8. Menjalankan PCA

```
# PCA
pca_result <- prcomp(supermarket,
                     scale = TRUE)

# Ringkasan PCA
summary(pca_result)
```

Fungsi prcomp() digunakan untuk melakukan PCA secara otomatis.

Parameter scale = TRUE memastikan seluruh variabel distandardisasi.

Output PCA biasanya menghasilkan:

- Standard deviation
- Proportion of variance
- Cumulative proportion

EVALUASI MODEL

9. Proporsi Variansi

```
# Melihat proporsi variansi
summary(pca_result)
```

Proporsi variansi menunjukkan persentase informasi yang dijelaskan setiap komponen.

Contoh interpretasi:

Komponen	Proporsi Variansi
PC1	95%
PC2	3%
PC3	1%
PC4	1%

Interpretasi:

- PC1 menjelaskan sebagian besar variasi data.
- Komponen lain hanya memberikan tambahan kecil.

Artinya data dapat disederhanakan hanya menggunakan satu komponen utama.

---

10. Menghitung Cumulative Variance

```
# Cumulative variance
cumsum(pca_result$sdev^2 / sum(pca_result$sdev^2))
```

Cumulative variance digunakan untuk menentukan jumlah komponen terbaik.

Contoh:

Komponen	Cumulative Variance
PC1	95%
PC1 + PC2	98%
PC1 + PC2 + PC3	99%

Jika cumulative variance sudah mencapai 90% maka komponen dianggap cukup mewakili data.

PENGUKURAN KESALAHAN

11. Reconstruction Error

```
# Mengambil dua komponen utama
scores <- pca_result$x[,1:2]

# Rekonstruksi data
reconstruction <- scores %*%
  t(pca_result$rotation[,1:2])

# Menghitung error
error <- mean((scale(supermarket) - reconstruction)^2)

# Menampilkan error
error
```

Reconstruction error digunakan untuk mengukur informasi yang hilang setelah reduksi dimensi.

Penjelasan:

- Data direduksi menjadi 2 komponen.
- Data kemudian dikembalikan ke bentuk awal.
- Selisih antara data asli dan hasil rekonstruksi dihitung.

Interpretasi:

- Error kecil → PCA baik.
- Error besar → terlalu banyak informasi hilang.

---

12. Menghitung Mean Squared Error

```
# Mean squared error
mse <- mean((scale(supermarket) - reconstruction)^2)

# Menampilkan MSE
mse
```

Mean Squared Error (MSE) merupakan ukuran rata-rata kuadrat kesalahan.

Semakin kecil MSE:

- Hasil PCA semakin baik.
- Informasi data semakin terjaga.

VALIDASI MODEL

13. Cross-validation Sederhana

```
# Membagi data
set.seed(123)

index <- sample(1:nrow(supermarket),
                0.8 * nrow(supermarket))

train_data <- supermarket[index, ]
test_data <- supermarket[-index, ]
```

Data dibagi menjadi:

- 80% training
- 20% testing

Tujuan:

- Menguji kestabilan PCA.
- Menghindari overfitting.
- Memastikan model dapat digunakan pada data baru.

---

14. PCA pada Data Training

```
# PCA training
pca_train <- prcomp(train_data,
                    scale = TRUE)

# Ringkasan hasil
summary(pca_train)
```

Hasil PCA pada data training digunakan untuk melihat apakah pola data tetap konsisten.

Jika hasil training dan testing mirip, maka model dianggap stabil.

---

15. Membuat Scree Plot

```
# Scree plot
plot(pca_result$sdev^2,
     type = "b",
```

```
     main = "Scree Plot PCA",
     xlab = "Komponen",
     ylab = "Nilai Eigen")
```

Scree plot digunakan untuk melihat jumlah komponen dominan.

Interpretasi:

- Penurunan tajam menunjukkan batas komponen penting.
- Komponen setelah penurunan tajam biasanya kurang signifikan.

ANALISIS HASIL

Berdasarkan hasil evaluasi diperoleh:

1. Komponen pertama memiliki nilai eigen terbesar.
2. Proporsi variansi terbesar berada pada PC1.
3. Cumulative variance sudah sangat tinggi hanya dengan sedikit komponen.
4. Reconstruction error relatif kecil.
5. Hasil cross-validation stabil.

Hal tersebut menunjukkan bahwa:

- Variabel dalam data supermarket saling berkaitan.
- Pengunjung, promosi, dan jumlah produk sangat memengaruhi pendapatan.
- PCA berhasil menyederhanakan data tanpa kehilangan banyak informasi.

INTERPRETASI BISNIS

Hasil PCA dan evaluasi model membantu perusahaan memahami pola bisnis secara lebih sederhana.

Beberapa insight yang diperoleh:

1. Pengunjung Sangat Memengaruhi Pendapatan

Cabang dengan jumlah pengunjung tinggi cenderung memiliki pendapatan lebih besar.

Artinya:

- Perusahaan perlu meningkatkan traffic pelanggan.
- Lokasi strategis menjadi faktor penting.

---

2. Promosi Memiliki Pengaruh Positif

Cabang dengan promosi besar menunjukkan peningkatan transaksi.

Artinya:

- Strategi pemasaran efektif.
- Diskon dan iklan dapat meningkatkan penjualan.

---

3. Jumlah Produk Memengaruhi Minat Pembeli

Cabang dengan produk lebih lengkap memiliki pendapatan lebih tinggi.

Artinya:

- Ketersediaan barang sangat penting.
- Perusahaan perlu menjaga stok produk.

---

4. PCA Membantu Penyederhanaan Data

Daripada menganalisis seluruh variabel satu per satu, perusahaan cukup menggunakan beberapa komponen utama.

Keuntungan:

- Analisis lebih cepat.
- Visualisasi lebih mudah.
- Pengambilan keputusan lebih sederhana.

KELEBIHAN DAN KEKURANGAN PCA

Kelebihan PCA

1. Mengurangi dimensi data.
2. Mempercepat komputasi.
3. Mengurangi korelasi antar variabel.
4. Membantu visualisasi data.
5. Mempertahankan sebagian besar informasi.

Kekurangan PCA

1. Interpretasi komponen kadang sulit.
2. Informasi kecil dapat hilang.
3. Sensitif terhadap skala data.
4. Membutuhkan standardisasi.

KESIMPULAN

Evaluasi dan validasi metode numerik merupakan tahapan penting dalam analisis data.

Pada Principal Component Analysis (PCA), evaluasi dilakukan menggunakan:

- Nilai eigen
- Proporsi variansi
- Cumulative variance
- Reconstruction error
- Mean Squared Error
- Cross-validation

Hasil evaluasi menunjukkan bahwa PCA mampu menyederhanakan data supermarket dengan tetap mempertahankan sebagian besar informasi penting.

Komponen pertama menjadi faktor paling dominan dalam menjelaskan variasi data.

Variabel pengunjung, promosi, dan jumlah produk memiliki hubungan kuat terhadap pendapatan supermarket.

Penggunaan R mempermudah proses:

- Perhitungan nilai eigen
- Standardisasi data
- Visualisasi
- Evaluasi model
- Validasi PCA

Sehingga perusahaan dapat melakukan analisis data secara lebih cepat, akurat, dan efisien dalam pengambilan keputusan bisnis.