Nama : Fikry Fahrezy Ramadhan
NIM : 2802658263

1. Regresi Linear Sederhana

   ABSORB:

   (a) `https://www.ibm.com/think/topics/linear-regression`

   (b) `https://www.colorado.edu/amath/sites/default/files/attached-files/ch12_0.pdf`

   DO:

   (a) Jelaskan dengan kata-kata dan pendapat anda sendiri, apa itu garis regresi dan mengapa garis itu dianggap sebagai garis yang paling cocok?

   Jawaban:

   Garis regresi adalah garis lurus yang menggambarkan hubungan linear antara dua variabel, yaitu variabel independen (x) dan variabel dependen (y). Garis ini merupakan representasi matematis dari trend atau pola hubungan antara kedua variabel tersebut.

   Garis regresi dianggap sebagai garis yang paling cocok karena:

   - **Meminimalkan kesalahan**: Garis ini diperoleh melalui metode kuadrat terkecil (least squares method) yang meminimalkan jumlah kuadrat selisih antara nilai observasi dengan nilai prediksi.
   - **Mengurangi jarak total**: Jarak vertikal dari setiap titik data ke garis regresi diminimalkan secara keseluruhan.
   - **Memberikan prediksi terbaik**: Dengan kriteria statistik yang objektif, garis ini memberikan prediksi yang paling akurat untuk nilai y berdasarkan nilai x yang diberikan.
   - **Memiliki sifat matematis optimal**: Garis regresi memiliki sifat bahwa jumlah residual (selisih) sama dengan nol, dan varians residual minimum.

   (b) Tentukan persamaan regresi linear yang dapat dibentuk dari tabel berikut, dimana x menyatakan jumlah jam belajar dan y menyatakan nilai ujian.

x	1	2	3	4	5
y	60	70	80	90	100

   Jawaban:

   Menentukan persamaan regresi linear $y = a + bx$.

   i. Menghitung rata-rata dari kedua variabel x dan y terlebih dahulu

   $\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

   $\bar{y} = \frac{60+70+80+90+100}{5} = \frac{400}{5} = 80$

   ii. Menghitung koefisien b (slope)

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

x	y	x^2	xy
1	60	1	60
2	70	4	140
3	80	9	240
4	90	16	360
5	100	25	500
$\Sigma x = 15$	$\Sigma y = 400$	$\Sigma x^2 = 55$	$\Sigma xy = 1300$

   $b = \frac{5(1300) - (15)(400)}{5(55) - (15)^2}$

   $b = \frac{6500 - 6000}{275 - 225} = \frac{500}{50} = 10$

   iii. Menghitung koefisien a (intercept)
   $a = \bar{y} - b\bar{x}$
   $a = 80 - 10(3)$
   $a = 80 - 30 = 50$
   $a = 50$

   iv. Sehingga persamaan regresi linear adalah:
   $y = 50 + 10x$

Langkah 2: Hitung koefisien b (slope)

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	60	-2	-20	40	4
2	70	-1	-10	10	1
3	80	0	0	0	0
4	90	1	10	10	1
5	100	2	20	40	4
			Total	100	10

$$b = \frac{100}{10} = 10 \tag{1}$$

Langkah 3: Hitung koefisien a (intercept)

$$a = \bar{y} - b\bar{x} = 80 - 10(3) = 80 - 30 = 50 \tag{2}$$

Persamaan regresi linear: $\boxed{y = 50 + 10x}$

Interpretasi:

- Setiap penambahan 1 jam belajar akan meningkatkan nilai ujian sebesar 10 poin
- Jika tidak belajar sama sekali (x=0), nilai dasar yang diperkirakan adalah 50

2. Regresi Linear Berganda

ABSORB:

(a) `https://www.investopedia.com/terms/m/mlr.asp`

(b) `https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/`

DO:

Seorang analis properti ingin mengembangkan sebuah model untuk memprediksi harga rumah di kota. Tujuannya adalah untuk membantu agen properti dan calon pembeli dalam menentukan harga wajar sebuah rumah berdasarkan fitur-fitur utamanya.

Analis tersebut percaya bahwa harga rumah tidak hanya bergantung pada satu faktor, melainkan dari beberapa faktor sekaligus. Lantas, ia memutuskan untuk menggunakan analisis regresi linear berganda.

Variable yang digunakan: Analis mengumpulkan data dari 30 rumah yang baru saja terjual di kota.

- Variabel Dependen (Y) = harga rumah (dalam jutaan rupiah)
- Variabel Independen 1 (X1) = luas bangunan (dalam meter persegi)
- Variabel Independen 2 (X2) = jumlah kamar tidur (jumlah unit)
- Variabel Independen 3 (X3) = jarak ke pusat kota (dalam kilometer)

SeSetelah melakukan analisis data, analis tersebut menemukan model persamaan regresi sebagai berikut : $Y = 50 + 1,5X1 + 25X2 - 10X3$. Selain itu ditemukan nilai R-squared sebesar 0,85.

Pertanyaan: Bagaimana anda menginterpretasikan persamaan tersebut?

Jawaban:

Persamaan regresi $Y = 50 + 1,5X_1 + 25X_2 - 10X_3$ dapat diinterpretasikan sebagai berikut:

Interpretasi Koefisien:

- **Konstanta (50):** Nilai dasar harga rumah adalah 50 juta rupiah ketika semua variabel independen bernilai nol.
- **Koefisien X_1 (1,5):** Setiap penambahan 1 meter persegi luas bangunan akan meningkatkan harga rumah sebesar 1,5 juta rupiah, dengan asumsi variabel lain tetap (ceteris paribus).
- **Koefisien X_2 (25):** Setiap penambahan 1 kamar tidur akan meningkatkan harga rumah sebesar 25 juta rupiah, dengan asumsi variabel lain tetap.
- **Koefisien X_3 (-10):** Setiap penambahan 1 kilometer jarak ke pusat kota akan menurunkan harga rumah sebesar 10 juta rupiah, dengan asumsi variabel lain tetap. Tanda negatif menunjukkan hubungan terbalik.

Analisis Model:

- **R-squared = 0,85:** Model ini mampu menjelaskan 85% variasi harga rumah, yang menunjukkan model memiliki kekuatan prediksi yang sangat baik.

- **Faktor paling berpengaruh:** Jumlah kamar tidur memiliki dampak terbesar terhadap harga (koefisien 25), diikuti oleh jarak ke pusat kota (koefisien -10) dan luas bangunan (koefisien 1,5).
- **Validitas ekonomi:** Model ini masuk akal secara ekonomi karena rumah dengan luas lebih besar dan kamar lebih banyak umumnya lebih mahal, sedangkan rumah yang jauh dari pusat kota cenderung lebih murah.

Contoh Prediksi:

Rumah dengan luas 100 m^2, 3 kamar tidur, dan berjarak 5 km dari pusat kota:

$$Y = 50 + 1,5(100) + 25(3) - 10(5) \tag{3}$$
$$= 50 + 150 + 75 - 50 = 225 \text{ juta rupiah} \tag{4}$$

3. Polynomial Regression

ABSORB:

Cermati materi regresi polinomial berikut:

(a) `https://home.iitk.ac.in/~shalab/regression/Chapter12-Regression-PolynomialRegression.pdf`

(b) `https://www.geeksforgeeks.org/machine-learning/python-implementation-of-polynomial-regression/`

(c) `https://www.geeksforgeeks.org/machine-learning/linear-vs-polynomial-regression-understanding-the-di`

DO:

Seorang analis data mencoba memodelkan hubungan antara pengeluaran iklan bulanan dan jumlah pengunjung situs web. Setelah membuat plot data, ia melihat bahwa hubungan antara kedua variabel tersebut tidak lurus (non-linear), melainkan membentuk kurva.

Dalam situasi ini mengapa penggunaan regresi polinomial lebih tepat dibandingkan regresi linear sederhana, dan apa resiko utama yang harus diperhatikan saat memilih derajat (pangkat tertinggi) dari polinomial tersebut?

Jawaban:

Mengapa Regresi Polinomial Lebih Tepat:

(a) **Hubungan Non-Linear:** Data menunjukkan hubungan yang membentuk kurva, bukan garis lurus. Regresi linear sederhana hanya dapat menangkap hubungan linear dan akan menghasilkan prediksi yang buruk untuk data non-linear.

(b) **Fleksibilitas Model:** Regresi polinomial dapat menangkap pola yang lebih kompleks seperti:
- Titik maksimum/minimum (dengan polinomial derajat 2)
- Beberapa titik belok (dengan polinomial derajat lebih tinggi)
- Pola kurva yang tidak dapat direpresentasikan garis lurus

(c) **Akurasi Prediksi:** Model polinomial akan memberikan fit yang lebih baik terhadap data dan prediksi yang lebih akurat dalam rentang data yang ada.

(d) **Relevansi Praktis:** Dalam konteks iklan-pengunjung, hubungan non-linear masuk akal karena:
- Efektivitas iklan mungkin meningkat pesat pada awalnya
- Mencapai titik saturasi dimana penambahan budget tidak efektif
- Atau bahkan menurun karena over-exposure

Risiko Utama dalam Memilih Derajat Polinomial:

(a) **Overfitting:**
- Model dengan derajat terlalu tinggi akan "menghafal" noise dalam data training
- Performa sangat baik pada data training tetapi buruk pada data baru
- Model menjadi terlalu kompleks dan tidak dapat digeneralisasi

(b) **Underfitting:**
- Model dengan derajat terlalu rendah tidak dapat menangkap pola sebenarnya
- Bias tinggi karena model terlalu sederhana
- Prediksi tidak akurat bahkan pada data training

(c) **Instabilitas Numerik:**
- Polinomial derajat tinggi dapat menyebabkan masalah numerik
- Koefisien menjadi sangat besar atau sangat kecil
- Prediksi di luar rentang data menjadi tidak realistis

(d) **Interpretabilitas:**

- Model menjadi sulit diinterpretasikan ketika derajat terlalu tinggi
- Kehilangan makna bisnis dari koefisien-koefisien

Solusi untuk Mengatasi Risiko:

- Gunakan cross-validation untuk memilih derajat optimal
- Terapkan regularization (Ridge, Lasso) untuk mencegah overfitting
- Pertimbangkan trade-off antara kompleksitas model dan performa
- Validasi model pada data yang belum pernah dilihat (test set)
- Mulai dengan derajat rendah dan tingkatkan secara bertahap