TUGAS UJIAN AKHIR SEMESTER

Mata Kuliah: Biostatistika

"ANALISIS MODEL REGRESI LOGISTIK ORDINAL"

(Studi Kasus Penyakit Jantung di AS)

Evi Nor Laili Solikh Amin

22/502120/PPA/06415

PROGRAM STUDI S2 MATEMATIKA

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGTAHUAN ALAM

UNIVERSITAS GADJAH MADA

YOGYAKARTA

2024

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan suatu kondisi dimana jantung tidak dapat berungsi dengan baik, sehingga menyebabkan kerja jantung sebagai pompa darah dan oksigen dalam tubuh terganggu. Terganggunya sirkulasi oksigen dan darah dapat mengakibatkan bercampurnya darah bersih dan darah kotor akibat melemahnya jantung, celah antara atrium kiri dan kanan.

Menurut statistik dunia, ada 9,4 juta kematian setiap tahun yang disebabkan oleh penyakit kardiovaskuler dan 45% kematian tersebut disebabkan oleh penyakit jantung koroner. Diperkirakan angka tersebut akan meningkat hingga 23,3 juta pada tahun 2030 (Wong,2014).

Penyakit jantung koroner (PJK) tetap menjadi masalah kesehatan yang signifikan dengan dampak sosio-ekonomi yang besar, disebabkan oleh biaya obat-obatan yang tinggi, durasi perawatan yang panjang, dan kebutuhan akan pemeriksaan tambahan selama proses pengobatan. Oleh karena itu, pencegahan melalui deteksi dini faktor risiko dan upaya pengendalian sangat penting untuk dilakukan.

Identifikasi faktor risiko Penyakit jantung koroner (PJK) sangat bermanfaat untuk perencanaan intervensi pencegahan. Berbagai penelitian telah berhasil mengidentifikasi faktor- faktor risiko penyakit jantung koroner antara lain herediter, usia, jenis kelamin, sosioekonomi, letak geografi, makanan tinggi lemak dan kalori, kurang makan sayur buah, merokok, alkohol, aktifitas fisik kurang, hipertensi, obesitas, diabetes mellitus, aterosklerosis, penyakit arteri perifer, stroke dan dislipidemia (Mendis et al, 2011).

Berdasarkan Cardiovascular Disease Risk Factor yang menyebabkan penyakit jantung sangat beragam dan kompleks. Studi mengenai faktor penyebab penyakit jantung sangat penting untuk memahami dan mengembangkan strategi pencegahan yang efektif. Salah satu metode analisis yang dapat digunakan untuk mengidentifikasi faktor-faktor ini adalah regresi logistik ordinal.

The Behavioral Risk Factor Surveillance System (BRFSS) adalah survei telepon terkait kesehatan yang dikumpulkan setiap tahun oleh CDC. Setiap tahun, survei ini mengumpulkan tanggapan dari lebih dari 400.000 orang Amerika mengenai perilaku berisiko terkait kesehatan, kondisi kesehatan kronis, dan penggunaan layanan pencegahan. Survei ini telah dilakukan setiap tahun sejak tahun 1984. Untuk proyek ini, dataset csv yang tersedia di Kaggle untuk tahun 2020.

Berdasarkan permasalahan tersebut maka pada penelitian ini dilakukan aplikasi model regresi logistik ordinal untuk mengetahui faktor-faktor yang mempengaruhi status penyakit jantung dataset kumpulan tanggapan dari 319.795 reponden. Sehingga dengan diketahui faktor-faktor yang menyebabkan tingkatan diharapkan dapat lebih sadar akan Kesehatan diri dan sekitar.

1.2 Rumusan Masalah

Berdasarkan permasalahan tersebut, maka rumusan masalah dalam penelitian ini adalah

1. Bagaimana pemodelan faktor-faktor yang mempengaruhi penyakit jantung dengan regresi logistik ordinal?
2. Bagaimana mengetahui faktor-faktor resiko yang paling berpengaruh terhadap penyakit Jantung?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, maka tujuan dari penelitian ini adalah

1. Memodelkan faktor-faktor yang mempengaruhi penyakit Jantung dengan regresi logistik ordinal.
2. Mengetahui faktor-faktor resiko yang paling berpengaruh terhadap penyakit Jantung.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah

1. Penulis dapat lebih menguasai dan mengkaji ilmu terutama di bidang Biostatistika, untuk dapat diterapkan pada kehidupan nyata.

2. Pembaca dapat menambah ilmu serta wawasan mengenai metode regresi logistik ordinal dan dapat menjadi referensi untuk melakukan penelitian-penelitian selanjutnya mengenai metode regresi logistik ordinal terkhusus di bidanag Biostatistika.

1.5 Batasan Masalah

Batasan masalah yang digunakan dalam penelitian ini adalah :

1. Data yang digunakan adalah data sekunder yang diperoleh dari Kaggel dengan web: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
2. Penelitian ini hanya dibatasi 4 variabel prediktor untuk melihat faktor-faktor yang mempengaruhi penyakit Jantung
3. Metode yang digunakan adalah metode regresi logistik ordinal.

BAB II

METPEN DAN KAJIAN TEORI

Bab ini menjelaskan mengenai metode penelitian dan kajian pustaka yang akan digunakan pada pembahasan. Adapun tujuan dari penelitian ini adalah untuk mengetahui factor-faktor yang mempengaruhi dari penyakit jantung. Kemudian dilakukan analisis, dan metode statistika yang digunakan adalah regresi logistic ordinal dengan aplikasi R.

2.1 Tabel Kontingensi (*Cross Tabulation*)

Tabel Kontingensi *(cross tabulation/* tabulasi silang) adalah tabel yang berisi data jumlahatau frekuensi atau beberapa klasifikasi (kategori) (Agresti, 2002). Metode tabulasi silangdapat menjawab hubungan antara dua atau lebih variabel penelitian tetapi bukan hubungan sebab akibat. Secara umum jika memiliki dua variabel A dan B, dimana variabel A terdiri atas I sel, yaitu $A_1,\ldots A_i,\ldots A_I$ dan variabel B terdiri atas J sel, yaitu $B_1,\ldots B_j,\ldots B_J$ maka akan mempunyai tabel dengan baris sebanyak I dan kolom sebanyak J seperti pada tabel 3.1 berikut:

Tabel 3.1 Tabel Kontingensi $I \times J$

Variabel A	Variabel B				Total
	1	2	…	J	
1	n_{11}	n_{12}	…	n_{1J}	n_1
2	n_{21}	n_{22}	…	n_{2J}	n_2
⋮	⋮	⋮	⋱	⋮	⋮
I	n_{I1}	n_{I2}	…	n_{IJ}	n_I
Total	n_1	n_2	…	n_J	n

Keterangan:

n_{IJ} : pengamatan pada sel ke i, j dengan $i = 1,2, \ldots, I$ dan $j = 1,2, \ldots, J$.

n_I : jumlah pengamatan pada sel ke i dengan $i = 1,2, \ldots, I$

n_J : jumlah pengamatan pada sel ke j dengan $j = 1,2, \ldots, J$

n : jumlah keseluruhan pengamatan pada sel

2.2 Uji Independensi

Uji independensi digunakan untuk mengetahui hubungan antara dua variabel. Hubungan dua variabel yang dimaksud adalah antara variabel respon dengan variabel prediktor (Agresti, 2002). Setiap sel dari variabel-variabel tersebut harus memenuhi syarat sebagaiberikut.

2.2.1 Homogen

Homogen adalah dalam setiap sel tersebut harus merupakan obyek yang sama.

2.2.2 *Mutually Exclusive* dan *Mutually Exhaustive*

Mutually exclusive artinya unit sel suatu variabel harus saling asing sehingga setiap pengamatan akan termuat dalam satu sel. *Mutually exhaustive* artinya pengklasifikasian harus mencakup seluruh bagian variabel sehingga tidak akan terjadi pengamatan yang tidak termasuk dalam sel.

2.2.3 Skala Nominal dan Skala Ordinal

Skala nominal dan skala ordinal adalah skala yang bersifat kategorikal atau klasifikasi. Perbedaan kedua skala tersebut adalah skala nominal dapat berfungsi untuk membedakan saja tetapi tidak ada tingkatan sedangkan skala ordinal berfungsi membedakan dan ada tingkatan.

Pengujian yang dilakukan pada uji independensi adalah sebagai berikut.

Hipotesis:

H_0 : Tidak ada hubungan antara dua variabel yang diamati

H_1 : Ada hubungan antara dua variabel yang diamati

Statistik uji yang digunakan adalah statistik *Pearson Chi-Square* dengan daerahpenolakannya adalah H_0 ditolak jika $\chi^2 > \chi^2_{\alpha,df(I-1)(J-1)}$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \qquad (3.1)$$

dengan $e_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$

dimana:

n_{ij} : jumlah pengamatan pada baris ke i kolom ke j

e_{ij} : nilai ekspektasi pengamatan pada baris ke i kolom ke j

n_i : jumlah pengamatan pada baris ke i

n_j : jumlah pengamatan pada kolom ke j

2.3 Uji Multikolinieritas

Uji multikolinieritas bertujuan untuk menemukan suatu korelasi atau hubungan antar variabel independen pada suatu model regresi, sehingga uji multikolinieritas ini hanya digunakan dan diolah pada beberapa variabel independen saja. Untuk melihat terjadinya multikolinieritas antar variabel yaitu dengan melihat hasil pada nilai VIF (*Variance Inflation Factor*). Suatu model regresi menunjukkan adanya multikolinieritas jika:

1. Nilai VIF > 10.
2. Nilai *tolerance* < 0,10.
3. Tingkat korelasi antar variabel independennya > 95%.

Langkah-langkah pengujian multikolinieritas:

1. Hitung nilai korelasi antar variabel independen (r)
2. Kuadratkan nilai korelasi antar variabel independen (r^2)
3. Hitung nilai $VIF = \frac{1}{1-R^2}$
4. Hitung nilai *tolerance* $(TOL) = \frac{1}{VIF}$

Jika terjadi multikolinieritas antar variabel independen maka model yang dihasilkan tidak berasumsi untuk digunakan, maka cara mengatasi multikolinieritas yang dapat dilakukan yaitu

1. Menghilangkan atau menambahkan variabel independent.
2. Tranformasi variabel (memasukkan persamaan tambahan ke model regresi).
3. Penambahan data atau memperbesar ukuran sampel.

2.4 Regresi Logistik Ordinal

Regresi logistik ordinal merupakan salah satu metode statistika yang digunakan untuk menganalisis hubungan antara variabel respon berskala ordinal dengan tiga kategori atau lebih dan variabel prediktor yang dapat bersifat kategori atau kuantitatif maupun kontinu (Hosmer & Lemeshow, 2000). Model untuk regresi ordinal sederhana disebut *cumulative logit models*. Pada model logit ini sifat ordinal dari respon Y dituangkan dalam peluang kumulatif sehingga *cumulative logit model* merupakan model yang didapat dengan membandingkan peluang kumulatif yaitu peluang kurang dari atau sama dengan kategori respon ke-j pada p variabel prediktor yang dinyatakan dalam vektor x_i, $P(Y \leq j|x_i)$, dengan peluang lebih besar dari kategori respon ke-j, x_i, $P(Y \leq j|x_i)$. Nilai peluang kumulatif ke-j adalah:

$$\pi_k(x) = P(Y \leq j) = \frac{exp\left[g_j(x_k)\right]}{1 + exp\left[g_j(x_k)\right]} = \frac{exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]}{1 + exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]};$$

$k = 1,2, \dots, j, \dots, r$

$$\pi_k(x) = P(Y \leq j) = \pi_1 + \pi_2 + \cdots + \pi_r \qquad (2.2)$$

Apabila $P(Y \leq j)$ dibandingkan dengan peluang suatu respon pada kategori $(j + 1)$ sampai dengan kategori r, maka hasilnya adalah sebagai berikut:

$$\frac{P(Y \leq j)}{P(Y > j)} = \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \frac{\dfrac{exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]}{1 + exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]}}{\dfrac{1}{exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]}}$$

$$\frac{P(Y \leq j)}{P(Y > j)} = \frac{P(Y \leq j)}{1 - P(Y \leq j)} = exp\left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right]$$

$$\frac{P(Y \leq j)}{P(Y > j)} = \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \frac{\pi_1 + \pi_2 + \cdots + \pi_r}{\pi_{j+1} + \pi_{j+2} + \cdots + \pi_r} \qquad (2.3)$$

Pada rumusan 2.3 dilakukan transformasi logistik menjadi model regresi logistik (logit)ordinal atau logit kumulatif.

$$Logit\ [P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left(\frac{\pi_1 + \pi_2 + \cdots + \pi_r}{\pi_{j+1} + \pi_{j+2} + \cdots + \pi_r}\right)$$

$$Logit\,[P(Y \leq j)] = \left[\beta_{oj} + \sum_{k=1}^{r} \beta_k x_k\right] \qquad (2.4)$$

Dengan nilai β_k untuk $k = 1,2,\dots,r$ pada setiap model regresi logistik ordinal adalah sama.

Jika terdapat empat kategori respon dimana $j = 1,2,3,4$ maka nilai dari peluang kategori respon diperoleh sebagai berikut:

$$P(Y = 1) = \pi_1(x) = \frac{exp[\beta_{01} + x'\beta]}{1 + exp[\beta_{01} + x'\beta]} \qquad (2.5)$$

$$P(Y = 2) = \pi_2(x) = \frac{exp[\beta_{02} + x'\beta]}{1 + exp[\beta_{02} + x'\beta]} - \pi_2(x) \qquad (2.6)$$

$$P(Y = 3) = \pi_3(x) = \frac{exp[\beta_{03} + x'\beta]}{1 + exp[\beta_{03} + x'\beta]} - \pi_3(x) \qquad (2.7)$$

$$P(Y = 4) = \pi_4(x) = 1 - \pi_1(x) - \pi_2(x) - \pi_3(x) \qquad (2.8)$$

Nilai $\pi(x_j)$ pada persamaan (3.5), (3.6), (3.7), (3.8) akan dijadikan pedoman pengklasifikasian. Suatu pengamatan akan masuk dalam respon kategori-j berdasarkan nilai $\pi(x_j)$ yang terbesar (Hosmer., dkk, 2000).

2.4.1 Estimasi Parameter

Estimasi parameter dapat dipergunakan dalam metode maksimum *likelihood.* Metode ini memperoleh dugaan maksimum *likelihood* bagi β dengan langkah awal yaitu membentuk fungsi *likelihood*. Estimasi dari parameter regresi logistik ordinal didapatkan dengan menurunkan fungsi log *likelihood* terhadap parameter yang akan diestimasi dan disamakan dengan nol. Bentuk umum dari fungsi *likelihood* untuk sampel dengan n independen observasi $(x_i, y_i), i = 1,2,\dots,n$ adalah sebagai berikut.

$$L(\beta) = \prod_{i=1}^{n} [\pi_0(x_i)^{y_{0i}}\, \pi_1(x_i)^{y_{1i}}\, \pi_2(x_i)^{y_{2i}}] \qquad (2.9)$$

Sehingga didapatkan fungsi ln-*likelihood* dapat diperoleh dengan cara mendiferensialkan $L(\beta)$ terhadap β dan menyamakannya dengan nol (Agresti,

2002). Nilai β diestimasi dengan metode numerik karena persamaannya bersifat nonlinier. Metode untuk mengestimasi varians dan kovarians dari taksiran β dikembangkan menurut teori MLE *(Maximum Likelihood Estimator)* yang menyatakan bahwa estimasi varians dan kovarians diperoleh dari turunan kedua fungsi ln-*likelihood* (Agresti, 2002).

Nilai taksiran β diperoleh dari penyelesaian turunan pertama fungsi ln-*likelihood* yang non linier digunakan iterasi Newton- Raphson dengan rumus:

$$\beta^{(t+1)} = \beta^{(t)} - \left(H^{(t)}\right)^{-1} q^{(t)} \qquad (2.10)$$

Dimana, $q^T = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \frac{\partial L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial L(\beta)}{\partial \beta_p}\right)$ (2.11)

H matriks Hessian dengan elemen-elemen $h_{ab} = \frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}$

$$H = \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1p} \\ h_{21} & h_{22} & \ldots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & \ldots & h_{pp} \end{pmatrix}$$

Langkah-langkah berikut merupakan langkah-langkah metode iterasi *Newton Raphson*:

1. Menentukan nilai awal estimasi parameter yaitu $\beta^{(0)}$. Sehingga dengan mensubtitusikan ke dalam Persamaan (3.5), (3.6), (3.7), (3.8) diperoleh peluang masing-masing kategori respon $\pi_j(x_i)$
2. Mencari matriks Hessian $H^{(0)}$ dan vektor $q^{(0)}$
3. Iterasi berlanjut untuk $t > 0$
4. Langkah tersebut dilakukan terus-menerus hingga didapatkan estimasi parameter, $\hat{\beta}$, yang mencapai kondisi konvergen d untuk setiap k yaitu:

$$\left|{\beta_k}^{(t+1)} - \widehat{\beta_k}\right| \leq d\left|{\beta_k}^{(t)} - \widehat{\beta_k}\right|; d > 0 \qquad (2.12)$$

2.4.2 Pengujian Signifikansi Parameter

Setelah mendapatkan parameter, maka selanjutnya adalah menguji signifikansi dari parameter yang telah diestimasi tersebut. Pengujian parameter digunakan untuk menguji koefisien β dari model yang telah diperoleh. Dalam model regresi logistik terdapat dua jenis pengujian yaitu pengujian serentak (keseluruhan) dan pengujian parsial (individual).

1. Uji Serentak (Keseluruhan)

   Uji serentak dilakukan untuk memeriksa keberartian koefisien β secara keseluruhan atau serentak. Jika parameter yang diuji signifikan maka dapat dikatakan jika model yang dibentuk sesuai untuk memodelkan variabel respon.

   Hipotesis yang dilakukan pada uji serentak adalah sebagai berikut:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (Variabel independen tidak berhubungan secara serentak terhadap variable dependen pada model)

H_1 : minimal ada satu $\beta_k \neq 0$ dengan $k = 1,2, \ldots, p$. (Variabel independen berhubungan secara simultan terhadap variabel dependen pada model)

Statistik uji yang digunakan adalah statistik uji G^2 atau *Likelihood Rasio Tests*.

$$G^2 = -2 \log \left[\frac{\left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \left(\frac{n_3}{n}\right)^{n_3}}{\prod_{i=1}^{n} [\pi_0(x_i)^{y_{0i}} \, \pi_1(x_i)^{y_{1i}} \, \pi_2(x_i)^{y_{2i}} \, \pi_3(x_i)^{y_{3i}}]} \right] \quad (2.13)$$

Dimana,

$$n_0 = \sum_{i=1}^{n} y_{0i} \, , n_1 = \sum_{i=1}^{n} y_{1i} \, , n_2 = \sum_{i=1}^{n} y_{2i} \, , n_3 = \sum_{i=1}^{n} y_{3i}$$

$$n = n_0 + n_1 + n_2 + n_3$$

Di bawah H_0 statistik uji G^2 akan mengikuti distribusi *Chi-square* dengan derajat bebas k (Hosmer., dkk, 2000). Sehingga untuk memperoleh keputusan, nilai statistik uji G^2 dibandingkan dengan nilai $\chi^2_{(\alpha,p)}$. Kriteria penolakan H_0 adalah jika $G^2 > \chi^2_{(\alpha,p)}$.

2. Uji Parsial (Individual)

   Uji parsial digunakan untuk mengetahui signifikansi parameter terhadap variabel respon. Statistik uji yang digunakan adalah uji *Wald*. Berdasarkan hasil uji *Wald*

maka akan diketahui apakah suatu variabel prediktor layak atau tidak masuk dalam model.

Hipotesis yang digunakan adalah sebagai berikut:

$H_0: \beta_1 = 0$ (tidak ada hubungan antara X terhadap Y)

$H_1: \beta_k \neq 0$ dengan $k = 1,2, \dots, p$ (terdapat hubungan antara X terhadap Y).

Statistik uji yang digunakan dalam uji parsial ini adalah sebagai berikut.

$$W^2 = \left(\frac{\widehat{\beta_k}}{SE(\widehat{\beta_k})} \right)^2 \qquad (2.14)$$

Statistik uji W^2 mengikuti distribusi *Chi-Square*, sehingga pengujian ini dilakukan dengan membandingkan nilai dari *Wald test* dengan nilai ${\chi^2}_{(\alpha,db)}$ pada tabel. Kriteria penolakan H_0 yang berarti parameter signifikan bila W^2 lebih besar dari ${\chi^2}_{(\alpha,db)}$ atau $p - value \leq \alpha$ dengan $db = j$.

2.4.3 Uji Kesuaian Model

Uji kesesuaian model dilakukan untuk mengetahui apakah model dengan variabel dependen tersebut merupakan model yang sesuai. Statistik uji yang digunakan adalah uji pearson chi-square. Hipotesis yang digunakan adalah sebagai berikut.

H_0 : Model sesuai (tidak ada perbedaan yang nyata antara hasil observasidengan kemungkinan hasil prediksi model)

H_1 : Model tidak sesuai (ada perbedaan yang nyata antara hasil observasidengan kemungkinan hasil prediksi model)

Statistik uji:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - {n'}_k \bar{\pi}_k)^2}{{n'}_k \bar{\pi}_k (1 - \bar{\pi}_k)} \qquad (2.15)$$

Keterangan:

o_k : observasi pada grup ke-k $\left(\sum_{j=1}^{c_j} y_j \text{ dengan } c_k: \text{respon } (0,1) \right)$

$\bar{\pi}_k$: rata-rata taksiran peluang

g : jumlah grup (kombinasi kategori dalam model serentak)

n'_k : banyak observasi pada grup ke-k

Daerah penolakan H_0 adalah jika $\hat{C} \geq \chi^2_{(\alpha,db)}$atau $p_{value} \leq \alpha$ dengan derajat bebas pada uji ini adalah $db = \mathrm{P} - (k+1)$ dimana k adalah jumlah variabelprediktor. Semakin tinggi nilai χ^2 dan semakin rendah p_{value} mengindikasikan bahwa terdapat kemungkinan model tidak sesuai dengan data (Hosmer., dkk, 2000).

2.4.4 Koefisien Determinasi Model

Pengujian koefisien determinasi model dilakukan untuk melihat seberapa besar variabel-variabel independen yang mempengaruhi nilai variabel-variabel dependen. Besarnya nilai koefisien determinasi pada model regresi logistik ditunjukkan oleh nilai *Mc. Fadden, Cox dan Snell, dan Nagelkerke R-square*. Koefisien *Nagelkerke* didapat dari penyempurnaan nilai koefisien determinasi *Cox* dan *Snell*. Berikut rumus ketiga koefisien determinasi:

$$R^2_{MF} = 1 - \left[\frac{likelihood(ModelB)}{likelihood(ModelA)}\right] \qquad (2.16)$$

Pada persamaan (2.16), R^2_{MF} merupakan merupakan koefisien determinasi *McFadden*. Persamaan (2.17) dibawah ini merupakan rumus untuk mencari koefisien determinasi *Cox* dan *Snell*.

$$R^2_{MF} = 1 - \exp\left[-\frac{2}{n}[likelihood(modelB) - likelihood(modelA)]\right] \quad (2.17)$$

$$R^2_{MAX} = 1 - \exp\left[-\frac{2}{n}[likelihood(modelA)]\right] \qquad (2.18)$$

Dari persamaan (2.17) dan (2.18) diatas maka didapatlah rumus untuk koefisien determnasi *Nagelkerke* yang dapat dilihat pada persamaan (2.19) berikut:

$$R^2_N = \left[\frac{R^2_{CS}}{R^2_{MAX}}\right] \qquad (2.19)$$

2.4.5 Interpretasi Model

Interpretasi koefisien untuk model regresi logistik ordinal dapat dilakukan dengan menggunakan nilai *odds ratio*nya. *Odds ratio* pada kategori $Y \leq j$ merupakan perbandingan antara x_1 dan x_2 adalah:

$$L_j(x_1) - L_j(x_2) = \log\left(\frac{\frac{P(Y \leq j|x_1)}{P(Y > j|x_1)}}{\frac{P(Y \leq j|x_2)}{P(Y > j|x_2)}}\right)$$

$$L_j(x_1) - L_j(x_2) = \log\left(\frac{\frac{F_j(x_1)}{\left(1 - F_j(x_1)\right)}}{\frac{F_j(x_2)}{\left(1 - F_j(x_2)\right)}}\right)$$

$$L_j(x_1) - L_j(x_2) = \beta_i(x_1 - x_2) \qquad (2.20)$$

Keterangan:

i = 1,2, … , m

m =banyaknya peubah penjelas

Parameter β_i diartikan sebagai peubah nilai fungsi logit kumulatif yang disebabkan oleh peubah satu unit peubah penjelas ke-i yang disebut log odds (misalnya antara $x = x_1$ dan $x = x_2$) yang dinotasikan sebagai:

$$Ln\,[\psi(x_1, x_2)] = g(x = x_1) - g(x = x_2) = \beta_i(x_1 - x_2)$$

Sehingga didapatkan penduga untuk *odds ratio* $(\hat{\psi})$ sebagai berikut:

$$(\hat{\psi}) = \exp[\beta_i(x_1 - x_2)] \qquad (2.21)$$

2.5 Deskripsi dan Sumber data

Data yang digunakan adalah data sekunder yang dipublikasi oleh Kaggle https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease adalah Data Penyakit Jantung yang awalnya berasal dari CDC dan merupakan bagian utama dari Behavioral Risk Factor Surveillance System (BRFSS), yang melakukan survei telepon tahunan untuk mengumpulkan data mengenai status kesehatan penduduk AS dengan 319.795 responden.

2.6 Variabel Penelitian

Dalam penelitian ini variabel respon (Y) adalah penyakit jantung (iya atau tidak), sedangkan variabel predictor (X) adalah Merokok, Alkohol, Diabetes dan

Umur.Dalam pemodelan regresi logistik ordinal ini digunakan 4 variabel yang akan ditunjukkan pada Tabel 3.1

Variabel	Keterangan
HeartDisease	Status penyakit jantung (0=tidak dan 1=iya)
Smoking	Status merokok (0=tidak dan 1=iya)
AlcoholDrinking	Status minum alkohol (0=tidak dan 1=iya)
Diabetic	Status Diabetes (0=tidak dan 1=iya)
CategoryAge	Kategori umur (1>70, 2=55-69, 3<55)

2.7 Struktur Data Penyakit Jantung

Berikut ini disajikan data penyakit dalam tabel dengan 5 data teratas sebagai berikut:

Heart Disease	Smoking	Alcohol Drinking	Diabetic	CategoryAge
0	1	0	1	2
0	0	0	0	3
0	1	0	1	2
0	0	0	0	3
0	0	0	0	1

BAB III
HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengetahui mengetahui faktor-faktor yang secara signifikan mempengaruhi sesoraang mengidap penyakit jantung. Data yang digunakan dalam penelitian ini adalah data sekunder dari CDC dan merupakan bagian utama dari Behavioral Risk Factor Surveillance System (BRFSS), yang melakukan survei telepon tahunan untuk mengumpulkan data mengenai status kesehatan penduduk AS dengan 319.795 responden. Metode statistika yang digunakan pada penelitian ini adalah Regresi logistic ordinal menggunakan software SPSS.

3.1 Deskriptif Data

Deskriptif data bertujuan untuk menyajikan data dengan lebih efektif sehingga mudah dipahami. Untuk data-data numerik disajikan dalam bentuk ringkasan ukuran pusatnya sedangkan untuk data kategorik disajikan dalam bentuk sebagai berikut:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
jantung	319795	0	1	.09	.280
Merokok	319795	0	1	.41	.492
alkohol	319795	0	1	.07	.252
diabet	319795	0	1	.14	.342
umur	319795	1	3	1.78	.805
Valid N (listwise)	319795				

Hasil statistic deskriptif di atas menunjukan bahwa jumlah responden sebanyak 319.795. Karena semua variabel berisi variabel dummy menyebabkan informasi pada nilai minimum dan maksimum menujukan 0 dan 1 kecuali umur yakni minimum 1 dan maksimum 3.

3.2 Uji Pearson Chi Square

Uji Chi-Square, hipotesis yang dirumuskan adalah tentang hubungan atau asosiasi antara dua variabel kategoris. Hipotesis tersebut terdiri dari hipotesis nol (H0) dan hipotesis alternatif (H1). Hipotesis untuk uji Chi Square sebagai berikut:

H_0 : tidak ada hubungan antara variabel prediktor dengan Penyakit Jantung

H_1 : ada hubungan antara variabel prediktor dengan Penyakit Jantung

Taraf signifikansi: $\alpha = 0{,}05$

Hipotesis diatas dapat dianalisis menggunakan statistik uji chi square

1. Variabel Prediktor Merokok

Merokok * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
Merokok	tidak	176551	11336	187887
	ya	115871	16037	131908
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3713.816[a]	1	.000		
Continuity Correction[b]	3713.033	1	.000		
Likelihood Ratio	3645.329	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	3713.804	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 11290.73.

b. Computed only for a 2x2 table

Berdasarkan hasil Analisa di atas menunjukan bahwa nilai Pearson Chi-Square adalah 3713.816 dengan 1 derajat kebebasan (df) dan nilai p sebesar 0.000. Nilai p (Asymp. Sig.) lebih kecil dari 0.05, menunjukkan bahwa ada hubungan yang signifikan secara statistik antara status merokok dan penyakit jantung.Tabel Kontingen si Alkohol. Ini berarti bahwa kejadian penyakit jantung tidak terjadi secara acak sehubungan dengan

status merokok. Orang yang merokok memiliki prevalensi penyakit jantung yang berbeda dibandingkan dengan orang yang tidak merokok.

2. Variabel Prediktor Alkohol

alkohol * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
alkohol	tidak	271786	26232	298018
	ya	20636	1141	21777
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	329.104[a]	1	.000		
Continuity Correction[b]	328.649	1	.000		
Likelihood Ratio	374.120	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	329.103	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1864.01.

b. Computed only for a 2x2 table

Berdasarkan hasil Analisa di atas menunjukan bahwa nilai Pearson Chi-Square adalah 329.104 dengan 1 derajat kebebasan (df) dan nilai p sebesar 0.000. Nilai p (Asymp. Sig.) lebih kecil dari 0.05, menunjukkan bahwa ada hubungan yang signifikan secara statistik antara konsumsi alkohol dan penyakit jantung. Ini berarti bahwa kejadian penyakit jantung tidak terjadi secara acak sehubungan dengan konsumsi alkohol. Orang yang mengonsumsi alkohol memiliki prevalensi penyakit jantung yang berbeda dibandingkan dengan orang yang tidak mengonsumsi alcohol.

3. Variabel Prediktor Jantung

diabet * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
diabet	tidak	258126	18308	276434
	ya	34296	9065	43361
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9769.372[a]	1	.000		
Continuity Correction[b]	9767.548	1	.000		
Likelihood Ratio	7667.732	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	9769.342	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 3711.50.

b. Computed only for a 2x2 table

Berdasarkan hasil Analisa di atas menunjukan nilai Pearson Chi-Square adalah 9,769.372 dengan 1 derajat kebebasan (df) dan nilai p sebesar 0.000. Nilai p (Asymp. Sig.) kurang dari 0.05, menunjukkan bahwa ada hubungan yang signifikan secara statistik antara status diabetes dan penyakit jantung. Ini berarti bahwa kejadian penyakit jantung tidak terjadi secara acak sehubungan dengan status diabetes. Orang dengan diabetes lebih cenderung memiliki penyakit jantung dibandingkan dengan orang tanpa diabetes.

4. Variabel Prediktor Umur

umur * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
umur	>70	142103	3398	145501
	69-55	87964	9630	97594
	<55	62355	14345	76700
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17497.261[a]	2	.000
Likelihood Ratio	17854.253	2	.000
Linear-by-Linear Association	17461.570	1	.000
N of Valid Cases	319795		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6565.17.

Berdasarkan hasil uji Chi-Square menunjukan nilai Pearson Chi-Square adalah 17,497.261 dengan 2 derajat kebebasan (df) dan nilai p sebesar 0.000. Nilai p (Asymp. Sig.) lebih kecil dari 0.05, yang berarti ada hubungan yang signifikan secara statistik antara umur dan penyakit jantung., terdapat hubungan yang signifikan antara umur dan penyakit jantung (p value kurang dar 0.05/2). Ini berarti bahwa frekuensi kejadian penyakit jantung tidak terjadi secara acak sehubungan dengan umur. Ada pola yang menunjukkan bahwa kategori umur tertentu lebih mungkin atau kurang mungkin untuk memiliki penyakit jantung.

Selain itu Umur >70: Sebagian besar orang dalam kategori ini tidak memiliki penyakit jantung, namun ada sejumlah kecil yang memiliki penyakit jantung. Umur 69-55: Ada lebih banyak orang yang memiliki penyakit jantung dibandingkan dengan kategori umur >70. Umur <55: Kategori ini memiliki jumlah orang dengan penyakit jantung tertinggi. Hal ini menunjukkan bahwa ada perbedaan signifikan dalam prevalensi penyakit jantung di antara kelompok umur yang berbeda.

3.3 Uji Independensi Sample T Test

Uji t-independen (Independent Sample t-Test) digunakan untuk membandingkan rata-rata dua kelompok independen untuk melihat apakah perbedaannya signifikan secara statistik.

Hipotesis:

H0: Tidak ada perbedaan yang signifikan antara rata-rata dua kelompok.

H1:Ada perbedaan yang signifikan antara rata-rata dua kelompok.

Taraf Signifikasi :0.05

Hasil Uji Independensi dapat dilihat dengan 3 cara yakni membandingkan t hitung

dengan t table, p value kurang dari alfa dan memperhatikan nilai lower dan upper tidak melewati nol.Pembahasan kali ini akan dianalisis pada p value dan lower dan upper.

1. Variabel Merokok

Group Statistics

	Merokok	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	187887	.06	.238	.001
	ya	131908	.12	.327	.001

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
jantung	Equal variances assumed	15261.467	.000	-61.298	319793	.000	-.061	.001	-.063	-.059
	Equal variances not assumed			-58.093	226467.018	.000	-.061	.001	-.063	-.059

Pada hasil independensi sample T test berdasarkan merokok diperoleh nilai p-value kurang dari 0.05 dan nilai upper dan lower tidak melewati nol. Sehingga keputusan tolak Ho yang artinya Berdasarkan Merokok terdapat hubungan antara variable merokok dengan penyakit jantung.

2. Variabel diabetes

Group Statistics

	diabet	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	276434	.07	.249	.000
	ya	43361	.21	.407	.002

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
jantung	Equal variances assumed	1423.082	.000	18.151	319793	.000	.036	.002	.032	.039
	Equal variances not assumed			22.313	27197.406	.000	.036	.002	.032	.039

Pada hasil independensi sample T test berdasarkan merokok diperoleh nilai p-value (.000) kurang dari 0.05 dan nilai upper dan lower tidak melewati nol. Sehingga keputusan tolak Ho yang artinya Berdasarkan Merokok terdapat hubungan antara seseorang yang punya penyakit diabetes dengan penyakit jantung.

3. Uji independesi sample T test Alkohol

Group Statistics

	alkohol	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	298018	.09	.283	.001
	ya	21777	.05	.223	.002

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
jantung	Equal variances assumed	1423.082	.000	18.151	319793	**.000**	.036	.002	**.032**	**.039**
	Equal variances not assumed			22.313	27197.406	**.000**	.036	.002	**.032**	**.039**

Pada hasil independensi sample T test berdasarkan merokok diperoleh nilai p-value (.000) kurang dari 0.05 dan nilai upper dan lower tidak melewati nol. Sehingga keputusan tolak Ho yang artinya Berdasarkan Merokok terdapat hubungan antara seseorang yang memiliki riwayat meminum alkohol dengan penyakit jantung.

3.4 Regresi Logistik Ordinal

1. Uji Multikolinearitas

Uji multikolinieritas bertujuan untuk menemukan suatu korelasi atau hubungan antar variabel independen pada suatu model regresi, sehingga uji multikolinieritas ini hanya digunakan dan diolah pada beberapa variabel independen saja. Untuk melihat terjadinya multikolinieritas antar variabel yaitu dengan melihat hasil pada nilai VIF (*Variance Inflation Factor*).

Hipotesis untuk uji independensi adalah sebagai berikut:

H_0 : VIF < 10 artinya tidak terdapat korelasi atau hubungan antar variabel independen pada suatu model regresi.

H_1 : VIF > 10 artinya terdapat korelasi atau hubungan antar variabel independen pada suatu model regresi.

Hasil pengujian Multikolinearitas ditunjukkan pada tabel berikut

Tabel 3. Uji Multikolinearitas

Variabel	Tolerance	VIF	Keputusan
Merokok	.974	1.026	Terima H_0
Alkohol	.979	1.021	
Diabetes	.964	1.037	
Umur	.956	1.046	

Berdasarkan tabel 3. menunjukkan bahwa nilai VIF < 10 untuk semua variable prediktor. Sehingga dapat keputusannya adalah terima H_0 yang berarti tidak terdapat korelasi atau hubungan antar variabel independen pada suatu model regresi.

2. Uji Simultan

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	25256.584			
Final	728.051	24528.533	5	.000

Link function: Logit.

Berdasarkan hasil perhitungan uji simultan di atas menunjukan bahwa nilai chi square sebesar 24528.533 dan p-value sebesar .000< 0.05. Maka keputusan yang di ambil adalah tolak Ho, sehingga model dengan variabel predictor lebih baik dari pada model tanpa variabel respon.

3. Uji Kesesuaian Model

Uji kesesuaian model dilakukan untuk mengetahui kesesuaian model yang telah terbentuk dalam analisis regresi logistik ordinal atau tidak terdapat perbedaan antara hasil pengamatan dengan kemungkinan hasil prediksi pada model. Pengujian ini akan dianalisis dengan hipotesis sebagai berikut.

H_0 : Model telah sesuai (tidak terdapat perbedaan yang signifikan antarahasil pengamatan dengan kemungkinan hasil prediksi model)

H_1 : Model tidak sesuai (terdapat perbedaan yang signifikan antara hasil pengamatan dengan kemungkinan hasil prediksi model)

Taraf signifikan : $\alpha = 0{,}05$

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	569.437	18	.000
Deviance	549.325	18	.000

Link function: Logit.

Berdasarkna uji kesesuaian model diperoleh nilai Chi square 569.437 dan nilai deviance 549.325 sedangkan nilai p-*value* sebesar 0,000 kurang dari nilai α yaitu 0,05. Oleh karena itu, dapat diputuskan tolak H_0 sehingga dapat disimpulkan bahwa model tidak sesuai atau terdapat perbedaan yang signifikan antara hasil pengamatan dengan kemungkinan hasil prediksi model.

4. Koefisien determinasi.

Pengujian koefisien determinasi model dilakukan untuk melihat seberapa besar variabel-variabel independen yang mempengaruhi nilai variabel-variabel dependen. Besarnya nilai koefisien determinasi pada model regresi logistik

ditunjukkan oleh nilai *Mc. Fadden, Cox dan Snell, dan Nagelkerke R-square*. Berikut hasil analisisnya:

Tabel Nilai Koefisien Determinasi

Cox and Snell	0,074
Nagelkerke	0,167
McFadden	0,131

Berdasarkan table di atas menunjukkan bawah hasil nilai koefisien determinasi. Nilai tertinggi ada pada nilai Nagelkerke yaitu sebesar 0,167, artinya variable predictor mampu mempengaruhi variable respon sebesar 16.7%.

5. Uji Parsial

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[HeartDisease = 0]	.811	.035	547.394	1	.000	.743	.879
Location	[Smoking=0]	-.641	.013	2285.657	1	.000	-.667	-.615
	[Smoking=1]	0[a]	.	.	0	.	.	.
	[AlcoholDrinking= 0]	.377	.032	136.810	1	.000	.314	.440
	[AlcoholDrinking= 1]	0[a]	.	.	0	.	.	.
	[Diabetic=0]	-.970	.015	4341.239	1	.000	-.999	-.941
	[Diabetic=1]	0[a]	.	.	0	.	.	.
	[AgeCategory=1]	-2.043	.020	10444.127	1	.000	-2.082	-2.004
	[AgeCategory=2]	-.706	.014	2382.938	1	.000	-.734	-.678
	[AgeCategory=3]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Berdasarkan hasil di atas untuk semua variable prediktor (merokok,alcohol, diabetes dan umur) memiliki nilai p value .000<0.05 dan memiliki nilai lower dan upper yang tidak melewati nol sehingga Tolak Ho. Sehingga merokok, alcohol, diabetes dan umur mempengaruhi penyakit jantung.

6. Model

$$g(x) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

$$= .811 - 0.641x_1 + 0.377x_2(1) - 0.97x_3(1) - 2.043x_4(1) - 0.706x_4(2)$$

7. Odd Rasio (OR)

$$OR = EXP(\beta_i)$$

Dengan interval konfidensi 95% dengan perubahan Δ unit dari variable predictor adalah

$$\exp(\Delta\beta_i \pm 1.96\Delta(standart\ eror\ of\ \beta_i\ )$$

Berikut ini nilai Odd Rasio berdasarkan hasil table kontingensi di poin 3.2

OR merokok

Hitung odds ratio menggunakan rumus:

$$\text{Odds Ratio}(OR) = \frac{a \cdot d}{b \cdot c}$$

Substitusi nilai a, b, c, dan d ke dalam rumus:

$$\text{OR} = \frac{176551 \times 16037}{11336 \times 115871}$$

Hitung nilai OR:

$$\text{OR} = \frac{176551 \times 16037}{11336 \times 115871} = \frac{2830278987}{1313592856} \approx 2.154$$

Jadi, odds ratio (OR) dari data crosstabulation yang diberikan adalah sekitar 2.154.

Ini berarti bahwa orang yang merokok memiliki sekitar 2.154 kali lebih besar odds untuk mengidap penyakit jantung dibandingkan dengan orang yang tidak merokok.

Minum alcohol

$$\text{OR} = \frac{271786 \times 1141}{26232 \times 20636}$$

Hitung nilai OR:

$$\text{OR} = \frac{271786 \times 1141}{26232 \times 20636} = \frac{309671426}{541453152} \approx 0.572$$

Jadi, odds ratio (OR) dari data crosstabulation yang diberikan adalah sekitar 0.572.

Ini berarti bahwa orang yang tidak minum alkohol memiliki sekitar 0.572 kali odds untuk mengidap penyakit jantung dibandingkan dengan orang yang minum alkohol. Dengan kata lain, orang yang minum alkohol memiliki lebih tinggi odds untuk mengidap penyakit jantung dibandingkan dengan orang yang tidak minum alkohol.

Substitusi nilai a, b, c, dan d ke dalam rumus:

$$\text{OR} = \frac{258126 \times 9065}{18308 \times 34296}$$

Hitung nilai OR:

$$\text{OR} = \frac{258126 \times 9065}{18308 \times 34296} = \frac{2340691690}{627974368} \approx 3.727$$

Jadi, odds ratio (OR) dari data crosstabulation yang diberikan adalah sekitar 3.727.

Ini berarti bahwa orang yang memiliki diabetes memiliki sekitar 3.727 kali lebih besar odds untuk mengidap penyakit jantung dibandingkan dengan orang yang tidak memiliki diabetes.

$$\text{OR} = \frac{142103 \times 14345}{3398 \times 62355}$$

Hitung nilai OR:

$$\text{OR} = \frac{2037507335}{211846290} \approx 9.62$$

Jadi, odds ratio (OR) untuk kelompok umur >70 dibandingkan dengan <55 adalah sekitar 9.62. Ini berarti bahwa orang yang berumur >70 memiliki sekitar 9.62 kali lebih besar odds untuk mengidap penyakit jantung dibandingkan dengan orang yang berumur <55.

Umur 69-55 vs. <55

$$\text{OR} = \frac{87964 \times 14345}{9630 \times 62355}$$

Hitung nilai OR:

$$\text{OR} = \frac{1261499980}{600508650} \approx 2.10$$

Jadi, odds ratio (OR) untuk kelompok umur 69-55 dibandingkan dengan <55 adalah sekitar 2.10. Ini berarti bahwa orang yang berumur 69-55 memiliki sekitar 2.10 kali lebih besar odds untuk mengidap penyakit jantung dibandingkan dengan orang yang berumur <55.

Sehingga

- **Umur >70 vs. <55**: OR ≈ 9.62
- **Umur 69-55 vs. <55**: OR ≈ 2.10

Orang yang berumur >70 memiliki odds yang jauh lebih besar untuk mengidap penyakit jantung dibandingkan dengan orang yang berumur <55. Orang yang berumur 69-55 juga memiliki odds yang lebih besar untuk mengidap penyakit jantung dibandingkan dengan orang yang berumur <55, tetapi tidak sebesar kelompok umur >70.

3.5 Interpretasi Model

Berdasarkan hasil dari tabel parameter dapat dintrepretasikan sebagai berikut:

1. Dipoleh semua variabel predictor yakni merokok, minum alcohol, diabetes dan usia mempengaruhi penyakit jantung.
2. Selain itu tidak merokok, tidak minum alcohol, tidak memiliki diabetes, dan menjadi lebih muda (kategori usia 1 dan 2) semuanya mengurangi log odds terkena penyakit jantung. Atau jika diperjelas menjadi jika seseorang merokok, minum alcohol,memiliki diabetes dan usia nya semakin bertambah banyak memiliki peluang terkena penyakit jantung semakin tinngi. Dan yang memiliki nilai OR paling tinggi adalah variable umur, yakni ketika seseorang memiliki umur diatas 70 tahun memiliki resiko lebih besar 9.62 kali dibandingkan umur dibawahnya.

BAB IV
KESIMPULAN

Berdasarkan hasil analisis dan pembahasan ppada penelitian ini, diperoleh kesimpulan sebagai berikut:

1. Berdasarkan model fungsi logit dari regresi logistic ordinal sebagai berikut:

$$g(x) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$
$$= .811 - 0.641\ merokok + 0.377\ Alkohol(1)$$
$$- 0.97 Diabetes\ (1) - 2.043 Umur(1) - 0.706 Umur(2)$$

2. Berdasarkan data dengan 319.795 reponden (penderita penyakit jantung 27.373 dan tidak penyakit jantung 292.422) Faktor-faktor resiko yang paling berpengaruh terhadap penyakit Jantung adalah semua variabel predictor dari penelitian, yakni merokok, minum alcohol, diabetes dan umur.

DAFTAR PUSTAKA

A. Agresti. 2017. "*An Introduction to Categorical Data Analysis Second Edition.*". Second Edition, Canada.

Agresti, A. 2002. *Categorical Data Analysis*. New York: John Wiley and Sons, Inc.

Cardiovascular Disease Risk Factor: [cited 2024]. Available from: http://www. w o r l d - h e a r t - f e d e r a t i o n . o rg / f i l e a d m i n / user_upload/documents/Fact_sheets/2012/ PressBackgrounderApril2012RiskFactors pdf.

D.W. Hosmer and S. Lemeshow. 2000."Applied logistic regression (Wiley Series in probability and statistics).", Ed. Sixth, John and Wiley.

Hosmer, D. W. Lemeshow, S., & Sturdivant, R. X. 2000. *Applied Logistic Regression*. New York: John Wiley and Sons, Inc.

Kaggle Inc (2020*). Indicator of Heart Disease*. Diakses pada 25 Mei 2024. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Kleinbum, David G, dan Mitchel Klein. 2010. *Logistic Regression "A Self-Learning Text".*Atlanta: Springer.

Mendis, S., Puska, P., Norrving, B.E. and World Health Organization, 2011. Global atlas on cardiovascular disease prevention and control. World Health Organization.

Wong, N.D., 2014. Epidemiological studies of CHD and the evolution of preventive cardiology. Nature Reviews Cardiology, 11(5), pp.276-289.

LAMPIRAN

Data

https://github.com/evinorlailisa/uas_biostat/raw/main/data_biostat.xlsx

Hasil SPSS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
jantung	319795	0	1	.09	.280
Merokok	319795	0	1	.41	.492
alkohol	319795	0	1	.07	.252
diabet	319795	0	1	.14	.342
umur	319795	1	3	1.78	.805
Valid N (listwise)	319795				

Tabel Kontingensi

Merokok * jantung Crosstabulation

Count

		jantung		Total
		tidak	ya	
Merokok	tidak	176551	11336	187887
	ya	115871	16037	131908
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3713.816[a]	1	.000		
Continuity Correction[b]	3713.033	1	.000		
Likelihood Ratio	3645.329	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	3713.804	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 11290.73.

b. Computed only for a 2x2 table

alkohol * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
alkohol	tidak	271786	26232	298018
	ya	20636	1141	21777
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	329.104[a]	1	.000		
Continuity Correction[b]	328.649	1	.000		
Likelihood Ratio	374.120	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	329.103	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1864.01.

b. Computed only for a 2x2 table

diabet * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
diabet	tidak	258126	18308	276434
	ya	34296	9065	43361
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9769.372[a]	1	.000		
Continuity Correction[b]	9767.548	1	.000		
Likelihood Ratio	7667.732	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	9769.342	1	.000		
N of Valid Cases	319795				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 3711.50.

b. Computed only for a 2x2 table

umur * jantung Crosstabulation

Count

		jantung		
		tidak	ya	Total
umur	>70	142103	3398	145501
	69-55	87964	9630	97594
	<55	62355	14345	76700
Total		292422	27373	319795

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17497.261[a]	2	.000
Likelihood Ratio	17854.253	2	.000
Linear-by-Linear Association	17461.570	1	.000
N of Valid Cases	319795		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6565.17.

Group Statistics

	Merokok	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	187887	.06	.238	.001
	ya	131908	.12	.327	.001

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
jantung	Equal variances assumed	15261.467	.000	-61.298	319793	.000	-.061	.001	-.063	-.059
	Equal variances not assumed			-58.093	226467.018	.000	-.061	.001	-.063	-.059

Group Statistics

	diabet	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	276434	.07	.249	.000
	ya	43361	.21	.407	.002

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
jantung	Equal variances assumed	1423.082	.000	18.151	319793	.000	.036	.002	.032	.039
	Equal variances not assumed			22.313	27197.406	.000	.036	.002	.032	.039

Group Statistics

	alkohol	N	Mean	Std. Deviation	Std. Error Mean
jantung	tidak	298018	.09	.283	.001
	ya	21777	.05	.223	.002

Variabel	Tolerance	VIF	Keputusan
Merokok	.974	1.026	Terima H_0
Alkohol	.979	1.021	
Diabetes	.964	1.037	
Umur	.956	1.046	

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
jantung	Equal variances assumed	1423.082	.000	18.151	319793	**.000**	.036	.002	**.032**	**.039**
	Equal variances not assumed			22.313	27197.406	**.000**	.036	.002	**.032**	**.039**

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	25256.584			
Final	728.051	24528.533	5	.000

Link function: Logit.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	569.437	18	.000
Deviance	549.325	18	.000

Link function: Logit.

Tabel Nilai Koefisien Determinasi

Cox and Snell	0,074
Nagelkerke	0,167
McFadden	0,131

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[HeartDisease = 0]	.811	.035	547.394	1	.000	.743	.879
Location	[Smoking=0]	-.641	.013	2285.657	1	.000	-.667	-.615
	[Smoking=1]	0[a]	.	.	0	.	.	.
	[AlcoholDrinking= 0]	.377	.032	136.810	1	.000	.314	.440
	[AlcoholDrinking= 1]	0[a]	.	.	0	.	.	.
	[Diabetic=0]	-.970	.015	4341.239	1	.000	-.999	-.941
	[Diabetic=1]	0[a]	.	.	0	.	.	.
	[AgeCategory=1]	-2.043	.020	10444.127	1	.000	-2.082	-2.004
	[AgeCategory=2]	-.706	.014	2382.938	1	.000	-.734	-.678
	[AgeCategory=3]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.