Ujian Akhir Semester Mata Kuliah Analisis Data Kategorik

ANALISIS PREDIKTIF DIAGNOSIS MONKEYPOX BERDASARKAN GEJALA KLINIS MENGGUNAKAN REGRESI LOGISTIK BINER

disusun untuk memenuhi
Ujian Akhir Semester Mata Kuliah Analisis Data Kategorik

Oleh:

KELOMPOK 1

MUHAMMAD ZULKARNAINI	**2208108010015**
RAIZA SABILA PUTRI	**2208108010007**
ISTI KAMILA NANDA ZAHRA	**2208108010068**
MUHAMMAD ANNAZARI ALWAFI	**2208108010011**

DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH

2025

1. Topik Penelitian

Analisis Prediktif Status Infeksi *Monkeypox* Berdasarkan Gejala Klinis Menggunakan Metode Regresi Logistik Biner pada Data Kategorik. *Monkeypox* adalah penyakit zoonosis yang disebabkan oleh virus *Monkeypox*, anggota genus *Orthopoxvirus*, dengan gejala klinis menyerupai cacar namun tingkat keparahannya cenderung lebih ringan. Penyakit ini dapat ditularkan dari hewan ke manusia dan antar manusia melalui kontak langsung dengan lesi kulit, cairan tubuh, atau droplet pernapasan. Gejala umum yang sering muncul antara lain demam, pembengkakan kelenjar getah bening, nyeri otot, lesi kulit, nyeri rektum, dan pembengkakan amandel, yang dapat bervariasi tergantung pada kondisi imun pasien dan tingkat keparahan infeksi. Untuk memahami karakteristik klinis yang berkaitan dengan infeksi ini, dilakukan analisis prediktif terhadap status infeksi *Monkeypox* berdasarkan data gejala klinis menggunakan metode regresi logistik biner. Metode ini lazim digunakan dalam epidemiologi untuk memodelkan hubungan antara satu atau lebih variabel prediktor dengan probabilitas terjadinya kejadian biner, seperti status infeksi. Dalam analisis ini, gejala pasien dikodekan secara biner untuk memudahkan pemodelan dan interpretasi, sehingga memungkinkan identifikasi gejala yang paling berkontribusi terhadap status infeksi serta memberikan gambaran mengenai tingkat risiko infeksi berdasarkan kombinasi gejala yang diamati.

2. Tujuan Penelitian

Penelitian ini bertujuan untuk menganalisis hubungan antara gejala klinis yang dialami pasien dengan kemungkinan diagnosis *Monkeypox* melalui pendekatan analisis data kategorik. Fokus utama studi adalah mengidentifikasi gejala-gejala spesifik yang paling berpengaruh terhadap hasil diagnosis, serta mengukur tingkat kontribusi masing-masing gejala tersebut dalam memprediksi status infeksi. Dengan menggunakan metode statistik yang tepat untuk data kategorik, penelitian ini berupaya mengembangkan model prediktif yang dapat membantu proses skrining awal pasien yang diduga terinfeksi *Monkeypox*.

3. Metodologi Penelitian

Pada penelitian ini dilakukan metode analisis Regresi Logistik Biner. Regresi Logistik Biner merupakan salah satu model regresi non-linear yang memiliki variabel terikat berupa variabel biner, yaitu variabel yang memiliki nilai nol dan satu, atau variabel yang memiliki sebaran binomial. Pada variabel bebas berupa variabel numerik atau kategorik. Regresi Logistik Biner bertujuan guna memprediksi peluang kejadian pada suatu persitiwa atau kasus berdasarkan variabel faktor risiko. Oleh karena itu, metode Regresi logistik biner ini dipilih karena variabel dependen yang digunakan bersifat dikotomis (biner), yaitu:

- 1 = Pasien terdiagnosis *Monkeypox*
- 0 = Pasien tidak terdiagnosis *Monkeypox*

Metode ini efektif untuk memodelkan hubungan antara satu atau lebih variabel independen kategorikal dengan probabilitas terjadinya suatu peristiwa biner. Regresi logistik biner dapat mengukur hubungan antara variabel independen (baik numerik maupun kategorikal) dengan probabilitas suatu kejadian.

4. Data yang Digunakan

a. Sumber Data

Data yang digunakan merupakan *Monkeypox Patients Dataset* yang tersedia secara publik di Kaggle dengan jumlah 11 variabel dan data ini terdiri dari 18.784 dataset. Dimana variabel target yang menunjukkan apakah pasien tersebut terinfeksi *Monkeypox* atau tidak. Data lengkap dapat dilihat pada: https://www.kaggle.com/datasets/muhammad4hmed/*Monkeypox*-patients-dataset

b. Jumlah Sampel

Pada penelitian ini, untuk menentukan jumlah sampel yang tepat, dilakukan dengan menggunakan metode slovin agar data yang diteliti tetap representatif dan efisien. Pemilihan metode pengambilan jumlah sampel ini bertujuan untuk memastikan bahwa sampel yang di ambil dapat menggambarkan populasi secara keseluruhan tanpa melakukan survei kembali. Berdasarkan perhitungan menggunakan rumus tersebut, jumlah sampel yang diperoleh sebanyak 392 pasien.

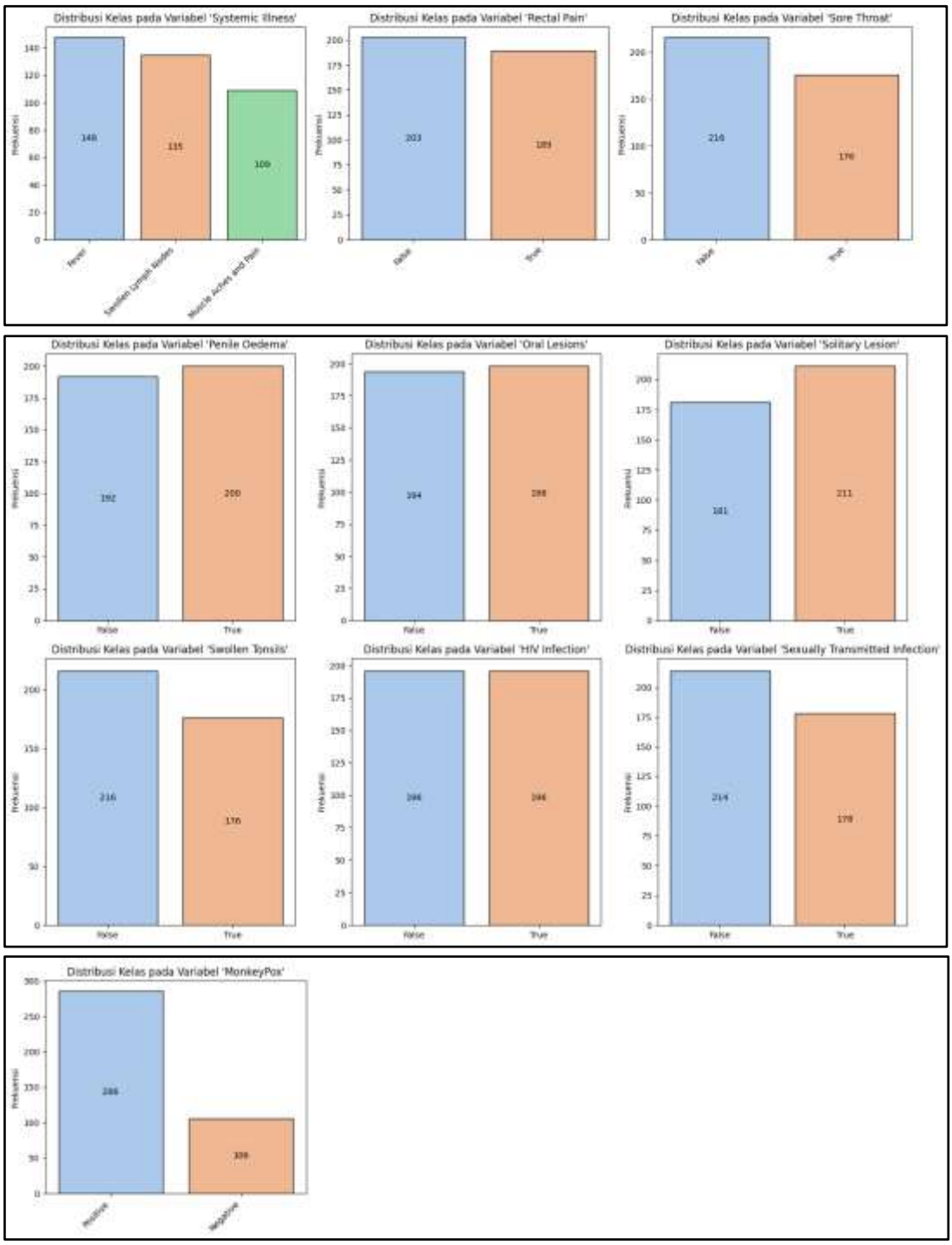
Berikut adalah daftar variabel yang digunakan dalam analisis:

Tabel 1. Deskripsi Data

Nama Variabel	**Tipe Data**	**Deskripsi**
Patient_ID	Nominal	ID unik untuk masing-masing pasien
Systemic Illness	Kategorikal	Kategori gejala utama yang dialami (contoh: Fever, Swollen Lymph Nodes, None, Other)
Rectal Pain	Biner	Apakah pasien mengalami nyeri rektum (True/False)
Sore Throat	Biner	Apakah pasien mengalami sakit tenggorokan (True/False)
Penile Oedema	Biner	Apakah pasien mengalami pembengkakan penis (True/False)
Oral Lesions	Biner	Apakah pasien memiliki lesi di mulut (True/False)
Solitary Lesion	Biner	Apakah pasien memiliki satu lesi tunggal (True/False)
Swollen Tonsils	Biner	Apakah pasien mengalami pembengkakan amandel (True/False)
HIV Infection	Biner	Apakah pasien terinfeksi HIV (True/False)
Sexually Transmitted Infection	Biner	Apakah pasien memiliki infeksi menular seksual (True/False)
Monkeypox	Biner (Target)	Status diagnosis pasien terhadap *Monkeypox*: **Positive** atau **Negative**

5. Analisis Data

a. Analisis Deskriptif



Gambar 1. Barchart Setiap Variabel

Berdasarkan hasil visualisasi distribusi data Monkeypox, terdapat 286 kasus positif *Monkeypox* dibandingkan 106 kasus negatif *Monkeypox*. Gejala sistemik yang paling sering ditemukan adalah demam (*Fever*), diikuti oleh pembengkakan kelenjar getah bening (*Swollen Lymph Nodes*) serta nyeri otot dan sendi (*Musle Aches and*

Pain). Sementara itu, pada gejala lokal seperti lesi tunggal, pembengkakan penis (*penile oedema*), dan lesi pada area mulut (*oral lesions*) juga muncul dengan frekuensi tinggi, Gejala lain seperti nyeri rektal, radang tenggorokan, dan pembengkakan amandel tercatat dalam jumlah yang lebih rendah namun tetap relevan. Selain itu, infeksi HIV dan infeksi menular seksual (STI) cukup signifikan, di mana penderita HIV tercatat hampir seimbang antara yang terinfeksi dan tidak.

b. Analisis Inferensia

1. Model Awal Regresi Logistik Biner

   Model Regresi Logistik Biner ini dikatakan model awal karena perlu dilakukan pengujian terlebih dahulu baik secara simultan maupun secara parsial

```
> model_awal

Call:  glm(formula = MonkeyPox ~ Systemic.Illness + Rectal.Pain + Sore.Throat +
    Penile.Oedema + Oral.Lesions + Solitary.Lesion + Swollen.Tonsils +
    HIV.Infection + Sexually.Transmitted.Infection, family = binomial(link = "logit"),
    data = df)

Coefficients:
                                (Intercept)  Systemic.IllnessMuscle Aches and Pain
                                    0.15958                                -1.17364
  Systemic.IllnessSwollen Lymph Nodes                            Rectal.PainTrue
                                    0.25183                                 0.31061
                        Sore.ThroatTrue                         Penile.OedemaTrue
                                    0.58835                                -0.07127
                      Oral.LesionsTrue                       Solitary.LesionTrue
                                    0.71686                                 0.09476
                   Swollen.TonsilsTrue                          HIV.InfectionTrue
                                    0.10497                                 0.29508
  Sexually.Transmitted.InfectionTrue
                                    0.47825
```

Gambar 2. Model Umum Regresi Logistik Biner

Pada *output* di atas diperoleh model awal regresi logistik biner ialah

$$Y = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = 0{,}160 - 1{,}17364_{\text{ Systemic Illness(Muscle Achesand Pain)}} + 0{,}25183_{\text{ Systemic Illness(Swollen Lymph Nodes)}} + 0{,}31061_{\text{ Rectal Pain(True)}} + 0{,}58835_{\text{Sore Throat(True)}} - 0{,}07127_{\text{Penile Oedema(True)}} + 0{,}71686_{\text{Oral Lesions(True)}} + 0{,}09476_{\text{Solitary Lesion(True)}} + 0{,}10497_{\text{ Swollen Tonsils(True)}} - 0{,}29508_{\text{ HIV Infection(True)}} + 0{,}47825_{\text{ Sexually Transmitted Infection(True)}}$$

2. Uji Simultan (Likelihood Rasio)

   Pengujian ini dilakukan untuk mengetahui signifikansi dari estimasi parameter yang diperoleh secara keseluruhan.

   - Hipotesis

     $H_0 : \beta_j = 0$ (variabel prediktor tidak berpengaruh secara signifikan terhadap variabel respon)

     $H_1 : \beta_j \neq 0$ (paling tidak ada satu variabel prediktor berpengaruh secara signifikan terhadap variabel respon)

   - Taraf Nyata

     $\alpha = 0{,}05$

- Daerah Penolakan

  $P_{value} < \alpha$ (0,05)

- Statistik Uji

```
> cat("Nilai G (Likelihood Ratio):", G_statistic, "\n")
Nilai G (Likelihood Ratio): 44.76623
> cat("Nilai chi-square:", chi_square_critical, "\n")
Nilai chi-square: 18.30704
```

Gambar 3. Uji Simultan (Likelihood Rasio) Model Awal

Berdasarkan *output* di atas dapat dilihat bahwa nilai Likelihood ratio (G) sebesar 44,76623 dengan nilai Chi-Square sebesar 18,30704.

- Keputusan dan Kesimpulan

  Dari Statistik uji di atas dapat dilihat bahwa nilai $G > X^2$ yaitu 44,76623 > 18,30704 artinya Tolak H_0. Maka dapat disimpulkan bahwa paling tidak ada satu variabel prediktor berpengaruh secara signifikan terhadap variabel respon.

3. Uji Parsial (Uji Wald) Model Awal

   Pengujian parsial dilakukan untuk mengetahui signifikansi β secara individu.

   - Hipotesis

     $H_0 : \beta_j = 0$ (parameter β tidak berpengaruh secara signifikan terhadap variabel respon secara individu)

     $H_1 : \beta_j \neq 0$ (parameter β berpengaruh secara signifikan terhadap variabel respon secara individu)

   - Taraf Nyata

     $\alpha = 0,05$

   - Daerah Penolakan

     $P_{value} < \alpha$ (0,05)

   - Statistik Uji

```
> summary(model_awal)

Call:
glm(formula = MonkeyPox ~ Systemic.Illness + Rectal.Pain + Sore.Throat +
    Penile.Oedema + Oral.Lesions + Solitary.Lesion + Swollen.Tonsils +
    HIV.Infection + Sexually.Transmitted.Infection, family = binomial(link = "logit"),
    data = df)

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                          0.15958    0.37213   0.429  0.66805
Systemic.IllnessMuscle Aches and Pain -1.17364    0.28980  -4.050 5.13e-05 ***
Systemic.IllnessSwollen Lymph Nodes   0.25183    0.30779   0.818  0.41325
Rectal.PainTrue                       0.31061    0.24454   1.270  0.20402
Sore.ThroatTrue                       0.58835    0.25015   2.352  0.01867 *
Penile.OedemaTrue                    -0.07127    0.24830  -0.289  0.77252
Oral.LesionsTrue                      0.71686    0.24708   2.901  0.00372 **
Solitary.LesionTrue                   0.09476    0.25213   0.376  0.70702
Swollen.TonsilsTrue                   0.10497    0.24605   0.427  0.66967
HIV.InfectionTrue                     0.29508    0.24567   1.202  0.22952
Sexually.Transmitted.InfectionTrue    0.47825    0.24855   1.918  0.05511 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gambar 4. Uji Parsial (Uji Wald) Model Awal

- Keputusan dan kesimpulan
  Dengan tingkat signifikansi 5% variabel Systematic Ilness Muscle aches and pain, Sore Throat, dan Oral Lesions berpengaruh secara parsial terhadap MonkeyPox, sedangkan variabel Systematic Ilness Swollen Lymph Nodes, Rectal Pain, Penile Oedema, Solitary Lesion, Swollen Tonsil, HIV Infection, dan Sexually Transmitted Infection tidak berpengaruh secara signifikan pada tingkat signifikansi pada level 5%

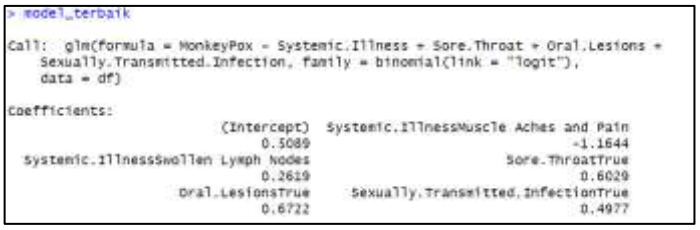
4. Uji Godness Of Fit (Hosmer- Lemeshow) Model Awal
   Pengujian yang dilakukan untuk mengetahui apakah model yang digunakan telah sesuai dengan data (FIT). berdasarkan nilai-nilai prediksi peluang.
   - Hipotesis
     H_0 : Model yang digunakan sesuai dengan data (tidak ada perbedaan antara hasil observasi dengan hasil prediksi)
     H_1 : Model yang digunakan tidak sesuai dengan data (ada perbedaan antara hasil observasi dengan hasil prediksi)
   - Taraf Nyata
     $\alpha = 0{,}05$
   - Daerah Penolakan
     $P_{value} < \alpha$ (0,05)
   - Statistik Uji

```
> #Uji kesesuaian model Awal (Goodnest OF FIT)
> hoslem_test = hoslem.test(model_awal$y, fitted(model_awal))
> hoslem_test

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_awal$y, fitted(model_awal)
X-squared = 11.091, df = 8, p-value = 0.1966
```

Gambar 5. Uji Godness Of Fit (Hosmer- Lemeshow) Model Awal

   - Keputusan dan kesimpulan
     Berdasarkan *output* di atas dapat dilihat bahwa nilai $P_{value} > \alpha$ yaitu 0,1966 > 0,05 yang artinya tidak dapat menolak H_0. Maka, dapat di simpulkan bahwa model yang digunakan sesuai dengan data.

5. Model Regresi Logistik Biner Terbaik



```
> model_terbaik

Call:  glm(formula = MonkeyPox ~ Systemic.Illness + Sore.Throat + Oral.Lesions +
    Sexually.Transmitted.Infection, family = binomial(link = "logit"),
    data = df)

Coefficients:
                          (Intercept)  Systemic.IllnessMuscle Aches and Pain
                               0.5089                                -1.1644
  Systemic.IllnessSwollen Lymph Nodes                          Sore.ThroatTrue
                               0.2619                                 0.6029
                      Oral.LesionsTrue     Sexually.Transmitted.InfectionTrue
                               0.6722                                 0.4977
```

Gambar 6. Model Regresi Logistik Biner Terbaik

$$Y = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = 0{,}5089 - 1{,}1644_{\text{Systemic Illness(Muscle Aches and Pain)}} + 0{,}2619_{\text{Systemic Illness(Swollen Lymph Nodes)}} + 0{,}6029_{\text{Sore Throat(True)}} + 0{,}6722_{\text{Oral Lesions(True)}} + 0{,}4977_{\text{Sexually Transmitted Infection (True)}}$$

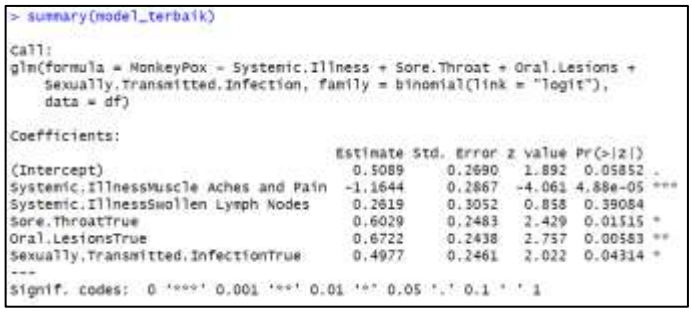
6. Uji Parsial (Uji Wald) Model Terbaik
   - Hipotesis

     $H_0 : \beta_j = 0$ (parameter β tidak berpengaruh secara signifikan terhadap variabel respon secara individu)

     $H_1 : \beta_j \neq 0$ (parameter β berpengaruh secara signifikan terhadap variabel respon secara individu)
   - Taraf Nyata

     $\alpha = 0{,}05$
   - Daerah Penolakan

     $P_{value} < \alpha$ (0,05)
   - Statistik Uji



```
> summary(model_terbaik)

Call:
glm(formula = MonkeyPox ~ Systemic.Illness + Sore.Throat + Oral.Lesions +
    Sexually.Transmitted.Infection, family = binomial(link = "logit"),
    data = df)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.5089     0.2690   1.892  0.05852 .
Systemic.IllnessMuscle Aches and Pain  -1.1644     0.2867  -4.061 4.88e-05 ***
Systemic.IllnessSwollen Lymph Nodes     0.2619     0.3052   0.858  0.39084
Sore.ThroatTrue                    0.6029     0.2483   2.429  0.01515 *
Oral.LesionsTrue                   0.6722     0.2438   2.757  0.00583 **
Sexually.Transmitted.InfectionTrue     0.4977     0.2461   2.022  0.04314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gambar 7. Uji Parsial (Uji Wald) Model Terbaik

   - Keputusan dan kesimpulan

     Dengan tingkat signifikansi 5%, semua variable dependen (Systemic_illness, Sore_Throat, Oral_Lesions dan Sexually_Transmitted_Infection) pada model terbaik berpengaruh signifikan terhadap variable respon (MonkeyPox), kecuali variabel Systemic_illness dengan kategori Swollen Lymph Nodes tidak berpengaruh secara signifikan pada tingkat signifikansi pada level 5%

7. Uji Simultan (Likelihood Rasio) Model Terbaik

- Hipotesis

  $H_0 : \beta_j = 0$ (variabel prediktor tidak berpengaruh secara signifikan terhadap variabel respon)

  $H_1 : \beta_j \neq 0$ (paling tidak ada satu variabel prediktor berpengaruh secara signifikan terhadap variabel respon)

- Taraf Nyata

  $\alpha = 0{,}05$

- Daerah Penolakan

  $P_{value} < \alpha$ (0,05)

- Statistik Uji

```
> cat("Nilai G (Likelihood Ratio):", G_statistic, "\n\n")
Nilai G (Likelihood Ratio): 41.26991

> cat("Nilai chi-square:", chi_square_critical, "\n")
Nilai chi-square: 11.0705
```

Gambar 8. Uji Simultan (Likelihood Rasio) Model Terbaik

Berdasarkan *output* di atas dapat dilihat bahwa nilai Likelihood ratio (G) sebesar 41,26991 dengan nilai Chi-Square sebesar 11,0705.

- Keputusan dan Kesimpulan

  Dari Statistik uji di atas dapat dilihat bahwa nilai $G > X^2$ yaitu 41,26991 > 11,0705 artinya Tolak H_0. Maka dapat disimpulkan bahwa paling tidak ada satu variabel prediktor berpengaruh secara signifikan terhadap variabel respon.

8. Uji Godness Of Fit (Hosmer- Lemeshow) Model Terbaik

- Hipotesis

  H_0 : Model yang digunakan sesuai dengan data (tidak ada perbedaan antara hasil observasi dengan hasil prediksi)

  H_1 : Model yang digunakan tidak sesuai dengan data (ada perbedaan antara hasil observasi dengan hasil prediksi)

- Taraf Nyata

  $\alpha = 0{,}05$

- Daerah Penolakan

$P_{value} < \alpha$ (0,05)

- Statistik Uji

```
> hoslem_test = hoslem.test(model_terbaik$y, fitted(model_terbaik))
> hoslem_test

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_terbaik$y, fitted(model_terbaik)
X-squared = 6.6574, df = 8, p-value = 0.574
```

Gambar 9. Uji Godness Of Fit (Hosmer- Lemeshow) Model Terbaik

- Keputusan dan kesimpulan

  Berdasarkan *output* di atas dapat dilihat bahwa nilai $P_{value} > \alpha$ yaitu 0,574 > 0,05 yang artinya tidak dapat menolak H_0. Maka,dapat di simpulkan bahwa model yang digunakan sesuai dengan data.

9. Odds Ratio

```
> exp(coef(model_terbaik))
                  (Intercept) Systemic.IllnessMuscle Aches and Pain   Systemic.IllnessSwollen Lymph Nodes
                    1.6635423                            0.3120962                             1.2993369
        Sore.ThroatTrue                  Oral.LesionsTrue   Sexually.Transmitted.InfectionTrue
                    1.8274920                            1.9586234                             1.6449377
```

Gambar 10. Odds Ratio

Interpretasi:

- Nilai odds ratio untuk variabel Systemic Illness Muscle Aches and Pain (Gejala Sistemik) didapatkan sebesar 0,3121. Hal ini menunjukkan bahwa pasien yang mengalami penyakit sistemik dengan gejala nyeri otot justru memiliki peluang lebih kecil (sekitar 0.3121 kali) untuk terkena *MonkeyPox* dibandingkan pasien yang tidak mengalaminya gejala sistemik nyeri otot.
- Nilai odds ratio untuk variabel Systemic Illness Swollen Lymph Nodes (Pembengkakan Kelenjar Getah Bening) didapatkan sebesar 1,2993. Ini menunjukkan bahwa pasien yang mengalami Penyakit Sistemik dengan gejala pembengkakan kelenjar getah bening memiliki peluang 1,2993 kali lebih besar untuk terkena *monkeypox* dibandingkan pasien tanpa gejala tersebut .
- Nilai odds ratio untuk variabel Sore Throat (Sakit Tenggorokan) didapatkan sebesar 1,8275. Hal ini menunjukkan bahwa pasien yang mengalami sakit tenggorokan memiliki peluang 1,8275 kali lebih besar untuk terkena *monkeypox* dibandingkan pasien yang tidak mengalami sakit tenggorokan.
- Nilai odds ratio untuk variabel Oral Lesions (Lesi Mulut) didapatkan sebesar 1,9586. Hal ini menunjukkan bahwa pasien dengan lesi di area mulut

memiliki peluang 1,9586 kali lebih besar untuk terkena *monkeypox* dibandingkan pasien yang tidak memiliki lesi di mulut.

- Nilai odds ratio untuk variabel Sexually Transmitted Infection (Infeksi Menular Seksual) didapatkan sebesar 1,6449. Ini berarti bahwa pasien yang memiliki infeksi menular seksual berpeluang 1,6449 kali lebih besar untuk terkena *monkeypox* dibandingkan pasien yang tidak memiliki infeksi tersebut.

6. Kesimpulan

Penelitian ini menunjukkan bahwa beberapa gejala klinis memiliki pengaruh signifikan terhadap kemungkinan diagnosis Monkeypox pada pasien. Berdasarkan hasil analisis regresi logistik biner, gejala seperti nyeri otot, sakit tenggorokan, lesi mulut, dan infeksi menular seksual terbukti memiliki hubungan yang signifikan dengan status infeksi Monkeypox. Pasien yang mengalami gejala-gejala ini, terutama lesi di mulut dan sakit tenggorokan, memiliki peluang lebih tinggi untuk terinfeksi Monkeypox. Sebaliknya, gejala seperti nyeri otot dan pembengkakan kelenjar getah bening menunjukkan peluang lebih kecil untuk terinfeksi. Uji goodness of fit menunjukkan bahwa model yang digunakan cocok dengan data, sehingga model regresi logistik biner ini dapat digunakan sebagai alat prediktif yang efektif dalam proses skrining awal pasien yang terduga terinfeksi Monkeypox. Dengan demikian, hasil penelitian ini dapat membantu dalam identifikasi gejala yang paling berisiko dan memberikan panduan untuk pengelolaan pasien yang lebih baik.