PEMODELAN REGRESI LOGISTIK BINER BAYESIAN DENGAN METODE *ADAPTIVE SYNTHETIC SAMPLING* PADA DATA PENYAKIT JANTUNG ISKEMIK

IZZUL HAQ
H051201069

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

PEMODELAN REGRESI LOGISTIK BINER BAYESIAN DENGAN METODE *ADAPTIVE SYNTHETIC SAMPLING* PADA DATA PENYAKIT JANTUNG ISKEMIK

IZZUL HAQ
H051201069

Skripsi

sebagai salah satu syarat untuk mencapai gelar sarjana

Program Studi Statistika

pada

PROGRAM STUDI STATISTIKA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM UNIVERSITAS HASANUDDIN
MAKASSAR
2024

SKRIPSI

PEMODELAN REGRESI LOGISTIK BINER BAYESIAN DENGAN METODE *ADAPTIVE SYNTHETIC SAMPLING* PADA DATA PENYAKIT JANTUNG ISKEMIK

IZZUL HAQ
H051201069

Skripsi,

telah dipertahankan di depan Panitia Ujian Sarjana Statistika pada tanggal 15 Agustus 2024 dan dinyatakan telah memenuhi syarat kelulusan pada

Program Studi Statistika
Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin
Makassar

Mengesahkan:
Pembimbing Tugas Akhir,

Sitti Sahriman, S.Si., M.Si.
NIP. 19881018 201504 2 002

Mengetahui:
Ketua Program Studi,

Dr. Anna Islamiyati, S.Si., M.Si.
NIP. 19770808 200501 2 002

vii

PERNYATAAN KEASLIAN SKRIPSI

DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, skripsi berjudul "Pemodelan Regresi Logistik Biner Bayesian Dengan Metode *Adaptive Synthetic Sampling* Pada Data Penyakit Jantung Iskemik" adalah benar karya saya dengan arahan dari pembimbing Sitti Sahriman, S.Si., M.Si. sebagai Pembimbing Utama. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbirkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka skripsi ini. Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini adalah karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut berdasarkan aturan yang berlaku.

Dengan ini saya melimpahkan hak cipta (hak ekonomis) dari karya tulis saya berupa skripsi ini kepada Universitas Hasanuddin.

Makassar, 15 Agustus 2024

Izzul Haq
NIM H051201069

ix

UCAPAN TERIMA KASIH

Segala puji bagi Allah *Subhanahu Wa ta'ala* atas segala limpahan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul **"Pemodelan Regresi Logistik Biner Bayesian Dengan Metode *Adaptive Synthetic Sampling* Pada Data Penyakit Jantung Iskemik"**. Shalawat dan salam senantiasa tercurahkan kepada baginda Rasulullah *Shallallahu 'Alaihi Wa sallam* beserta keluarga dan para sahabatnya.

Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada **Ibu Sitti Sahriman, S.Si., M.Si.** selaku Pembimbing Utama yang dengan sabar dan tulus meluangkan waktu di tengah berbagai kesibukannya untuk membimbing, memberi masukan, dan motivasi dalam penulisan skripsi ini. Terima kasih kepada **Bapak Drs Raupong, M.Si.** dan **Bapak Siswanto, S.Si., M.Si.** selaku Tim Penguji yang senantiasa memberikan saran dan kritikan yang membangun dalam penyempurnaan penulisan skripsi ini. Terima kasih kepada **Pimpinan Universitas Hasanuddin**, **Departemen Statistika**, **Jajaran Dosen**, dan **Staf Departemen Statistika** yang telah membekali ilmu dan kemudahan selama penulis menempuh studi.

Dengan rasa hormat, penulis mengucapkan terima kasih kepada kedua orang tua, Ibunda **Nurbiah** dan Ayahanda **Muh. Ishak Jumarang** atas inspirasi, pendidikan, serta cinta dan kasih sayang yang telah mengiringi setiap langkah penulis dengan doa dan restunya. Terima kasih juga kepada **Saudara** saya. Terima kasih kepada teman-teman di **STATISTIKA 2020** atas kebersamaan, semangat, dan motivasi yang telah diberikan selama menjalani pendidikan di Departemen Statistika. Juga, terima kasih kepada keluarga besar **Himastat FMIPA Unhas**, khususnya **POIS20N** atas ilmu dan pengalaman yang berharga. Penulis bangga menjadi bagian dari keluarga ini dan mengajak untuk tetap **BERSAMA DAN KUAT DI DALAM PERBEDAAN**. Kepada teman-teman **Pengurus BEM Periode 2023/2024**, atas cerita dan kerjasamanya. Kepada teman-teman **KKN GEL.111**, khususnya **Zydane, Rady dan Ochang**, atas Kerjasama dan teman bercerita bagi penulis. Kepada teman-teman **GAU22IAN dan AG23GASI**, terima kasih karena telah menjadi tempat bertukar cerita bagi penulis. Tetaplah **SELALU ADA**. Kepada sahabat-sahabat penulis, **Ngkal**, **Kur**, **Ryan, Razy, Fadlan**, **Azhar**, **Ara**, **Azalia**, **Uci**, **Liza**, **Maryana, Ruslinda**, **Alisha, Nahda**, **Irma**, **Naje**, **Sabila**, **Divia, Heri, Hakam, Reza**, **Mukhlis**, **Fahmi**, **Ayuni, Febi, Nahdia**, **Much**, **Theo, Russy, Rifky.** Kepada kanda-kanda terkhusus **Kanda Snuf, Kanda juni** terima kasih atas bimbingan, masukan dan arahannya selama ini dan yang tak sempat saya sebutkan namanya, terima kasih atas segala momen kebersamaan, pembelajaran, dan diskusi yang terus membangun penulis.

Penulis menyadari bahwa masih banyak terdapat kekurangan dalam penyusunan skripsi ini. Oleh karena itu, dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini dapat memberikan manfaat untuk berbagai pihak.

Penulis,

Izzul Haq

ABSTRAK

IZZUL HAQ. **Pemodelan Regresi Logistik Biner Bayesian Dengan Metode *Adaptive Synthetic Sampling* Pada Data Penyakit Jantung Iskemik**
(dibawah bimbingan ibu Sitti Sahriman, S.Si., M.Si).

Latar Belakang. Analisis regresi logistik digunakan untuk mengetahui hubungan antara satu variabel respon dikotomi dengan beberapa variabel prediktor. Namun, regresi logistik biner memiliki kekurangan, yaitu adanya asumsi bahwa hubungan antara variabel prediktor dan logit dari variabel respon bersifat linier. Asumsi ini tidak selalu akurat, sehingga hasil analisis dapat menjadi kurang tepat jika hubungan sebenarnya bersifat non-linier. Oleh karena itu, regresi logistik biner bayesian digunakan untuk menangani hubungan non-linear dengan lebih baik. Selain itu, teknik *oversampling* seperti *adaptive synthetic sampling* (ADASYN) diterapkan untuk menangani data yang tidak seimbang, sehingga meningkatkan kinerja klasifikasi dan mengurangi risiko *overfitting*. **Tujuan**. Penelitian ini bertujuan untuk mengetahui faktor-faktor yang mempengaruhi penyakit jantung iskemik berdasarkan model regresi logistik biner bayesian dengan ADASYN. **Metode**. Analisis data dilakukan dengan menyeimbangkan data untuk menyamakan skala variabel prediktor, kemudian membentuk model regresi logistik biner bayesian dengan ADASYN yang diterapkan pada data penyakit jantung iskemik. **Hasil**. Analisis menunjukkan setelah dilakukan standarisasi pada variabel prediktor, penggunaan metode ADASYN pada model regresi logistik biner bayesian menghasilkan kinerja yang lebih baik dibandingkan dengan model tanpa ADASYN. Akurasi model dengan ADASYN mencapai 82.11% sedangkan tanpa ADASYN sebesar 51.30%. Uji kebaikan model menunjukkan nilai sebesar 0.141 mengindikasikan peningkatan kinerja model. **Kesimpulan**. Model yang terbentuk dari regresi logistik biner bayesian dengan ADASYN menunjukkan bahwa penyakit jantung iskemik dipengaruhi oleh faktor usia (X_1), berat badan (X_2), indeks masa tumbuh (X_3), hemoglobin (X_4), hematokrit (X_5), leokosit (X_6), trombosit (X_7) dan hipertensi (X_8). Sedangkan faktor yang tidak dipengaruhi oleh penyakit jantung iskemik adalah kolestrol total (X_9) dan jenis kelamin (X_{10}).

Kata Kunci: Penyakit Jantung Iskemik, Regresi Logistik Biner Bayesian, A*daptive Synthetic Sampling*

ABSTRACT

IZZUL HAQ. **Bayesian Binary Logistic Regression Modelling with Adaptive Synthetic Sampling Method on Ischaemic Heart Disease Dataset**
(under the guidance of Mrs Sitti Sahriman, S.Si., M.Si).

Introduction. Logistic regression is used to analyze the relationship between a dichotomous response variable and multiple predictors. However, its assumption of linearity between predictors and the logit can be inaccurate, especially with nonlinear relationships. To address this, Bayesian binary logistic regression offers a better approach. Additionally, Adaptive Synthetic Sampling (ADASYN) is employed to correct data imbalance, enhancing classification and reducing overfitting risks. **Purpose**. This study aims to identify factors influencing ischemic heart disease using a Bayesian binary logistic regression model with ADASYN. **Method**. Analysis was performed by balancing the data to standardize the predictor variable scales and then constructing a bayesian binary logistic regression model with ADASYN applied to ischemic heart disease. **Results**. The analysis revealed that after standardizing the predictor variables, the use of ADASYN in the Bayesian binary logistic regression model resulted in better performance compared to the model without ADASYN. The accuracy of the model with ADASYN reached 82.11%, while the model without ADASYN achieved 51.30%. The goodness-of-fit test showed a value of 0.141, indicating an improvement in model performance. **Conclusion**. The model derived from the Bayesian binary logistic regression with ADASYN indicated that ischemic heart disease was influenced by factors such as age (X_1), weight (X_2), body mass index (X_3), hemoglobin (X_4), hematocrit (X_5), leukocytes (X_6), platelets (X_7), and hypertension (X_8). On the other hand, factors that were not influenced by ischemic heart disease included total cholesterol (X_9) and gender (X_{10}).

Keywords: Ischaemic Heart Disease, Bayesian Binary Logistic Regression, Adaptive Synthetic Sampling

DAFTAR ISTILAH

Istilah	Arti dan Penjelasan
Regresi	Analisis statistik yang digunakan untuk mengidentifikasi hubungan antara variabel respon berdasarkan satu atau lebih variabel prediktor
Biner	Variabel respon yang berupa data kualitatif dikotomi, yaitu bernilai 0 untuk gagal dan bernilai 1 untuk sukses
Bayesian	Informasi dari data sampel dan distribusi awal yang disebut sebagai distribusi *prior* dikombinasikan untuk menghasilkan distribusi *posterior* parameter populasi berdasarkan data sampel
Adaptive Synthetic Sampling	Metode yang digunakan untuk menyeimbangkan data pada data yang tidak seimbang
Estimasi	Proses memperkirakan nilai dari parameter populasi berdasarkan data sampel
Parameter	Nilai yang menggambarkan sifat atau karakteristik suatu populasi
Distribusi	Pola penyebaran data atau probabilitas kejadian dalam ruang sampel
Prior	Distribusi peluang yang merepresentasikan pengetahuan atau keyakinan awal tentang parameter sebelum data baru diperoleh
Posterior	Distribusi peluang yang diperoleh setelah memperbarui atau merevisi distribusi prior dengan informasi atau data yang baru
Markov Chain Monte Carlo	Teknik metode simulasi yang membangkitkan sejumlah sampel dari distribusi data tertentu untuk mendapatkan distribusi *posterior*
Gibbs Sampling	Teknik simulasi untuk membangkitkan variabel acak dari distribusi tertentu secara langsung tanpa harus menghitung densitasnya
Mean squared error	Metode alternatif untuk mengevaluasi teknik peramalan masing-masing kesalahan
Standarisasi Data	Proses mengubah skala variabel sehingga memiliki nilai rata-rata nol dan standar deviasi satu

DAFTAR LAMBANG DAN SINGKATAN

Lambang/singkatan	Arti dan Penjelasan
y_i	Variabel respon pada pengamatan ke-i
ε_i	Galat acak pada pengamatan ke-i
x_{ij}	Variabel prediktor ke-j pada pengamatan ke-i
β_j	Koefisien regresi dari variabel prediktor ke-j
β_0	Koefisien konstanta
$g(x_i)$	Nilai estimasi logit pada pengamatan ke-i
$\pi(x_i)$	Peluang kejadian sukses pada pengamatan ke-i
$\boldsymbol{x}$	Vektor kolom berukuran $n \times 1$ dari variabel prediktor
$\boldsymbol{\beta}$	Vektor kolom berukuran parameter regresi
s_j	Standar deviasi
$\bar{x}_j$	Rata-rata pada variabel ke-j
Z_{ij}	Standarisasi data ke-i variabel ke-j
$\hat{y}_i$	Nilai prediksi dari model untuk data ke-i.
π	Konstanta pi
μ	Rata-rata dari distribusi
σ^2	Variance dari distribusi
d	Derajat keseimbangan
r	Rasio
s_i	Sampel data sintetis
$\hat{r}_i$	Distribusi kerapatan
MCMC	*Markov Chain Monte Carlo*
MSE	*Mean Squared Error*
ADASYN	*Adaptive Synthetic Sampling*

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR LAMPIRAN

Lampiran **Halaman**

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung iskemik adalah salah satu jenis penyakit kardiovaskular yang memiliki jumlah penderita yang besar dan merupakan penyebab utama kematian di dunia, selain stroke. Penyakit ini terjadi ketika aliran darah ke otot jantung berkurang akibat penyempitan atau penyumbatan pada arteri koroner, disebabkan oleh penumpukan plak *aterosklerotik*. Kondisi ini dapat memicu gejala seperti nyeri dada dan meningkatkan risiko serangan jantung (Sechtem dkk., 2020). Pada tahun 2019, jumlah kasus penyakit jantung iskemik di dunia mencapai 197 juta kasus, dengan jumlah kematian sebanyak 9,74 juta jiwa. Di Indonesia, menurut data dari *Institute for Health Metrics and Evaluation* (IHME), terdapat 245.343 jiwa kematian per tahun akibat penyakit jantung iskemik (Fajriati & Prasetiyo, 2023). Rumah Sakit Umum Pusat Dr. Wahidin Sudirohusodo Makassar sebagai salah satu pusat jantung terkemuka di Indonesia Timur yang menangani banyak kasus penyakit jantung iskemik. Perencanaan dalam pencegahan penyakit jantung iskemik dapat dilakukan dengan mengidentifikasi faktor risiko penyakit jantung iskemik sejak dini sehingga dapat dilakukan tindakan lebih lanjut (Ghani dkk., 2016). Berdasarkan permasalahan tersebut, dibutuhkan analisis untuk mengetahui faktor-faktor yang berpengaruh terhadap penyakit jantung iskemik, salah satunya adalah dengan menggunakan analisis regresi.

Analisis regresi adalah metode statistik yang digunakan untuk menggambarkan hubungan antara variabel respon dan satu atau lebih variabel prediktor, salah satu jenis analisis regresi adalah regresi logistik, dimana regresi logistik biner merupakan bentuk yang paling sederhana dari analisis regresi logistik (Hosmer Jr dkk., 2013). Regresi logistik biner digunakan untuk menganalisis hubungan antara satu variabel respon dan beberapa variabel prediktor, dengan variabel respon yang berupa data kualitatif dikotomi, yaitu bernilai 0 untuk gagal dan bernilai 1 untuk sukses (Saragih dkk., 2020). Namun, regresi logistik biner memiliki beberapa kekurangan, termasuk asumsi bahwa *logit* dari variabel respon adalah fungsi linier dari variabel prediktor, yang tidak selalu benar. Jika hubungan ini tidak linier, hasil analisis bisa menjadi tidak akurat (Hosmer Jr dkk., 2013). Oleh karena itu, digunakan analisis regresi logistik biner bayesian untuk mengatasi permasalahan yang terdapat di regresi logistik biner.

Analisis regresi logistik biner bayesian merupakan analisis regresi logistik biner dengan pendekatan bayesian. Pendekatan ini menyediakan distribusi *posterior* yang memperhitungkan ketidakpastian dalam estimasi parameter dan dapat menangani hubungan non-linier dengan lebih baik, memberikan hasil analisis yang lebih akurat. Metode bayesian adalah metode estimasi parameter dimana parameter diasumsikan sebagai variabel acak yang mempunyai distribusi tertentu. Metode bayesian menggabungkan fungsi *likelihood* dan distribusi *prior* dari parameter untuk mendapatkan distribusi *posterior* yang selanjutnya menjadi dasar dalam estimasi parameter (Shobri dkk., 2021). Adapun distribusi *posterior* dalam penyelesaiannya

sering dijumpai tidak dapat diselesaikan secara analitis sehingga dibutuhkan simulasi yaitu *markov chain monte carlo* yang memungkinkan pengambilan sampel numerik dari distribusi *posterior* yang mendasarinya (Sumae dkk., 2022).

Markov chain monte carlo digunakan untuk memperbarui parameter. *Markov chain monte carlo* telah menjadi alat komputasi yang sangat penting dalam statistik bayesian karena dapat menerapkan teknik integrasi kompleks yang tidak dapat dihitung secara analitik. Metode bayesian penarikan kesimpulan tidak hanya berdasarkan informasi dari data sampel *(likelihood)* akan tetapi ada penambahan informasi subjektif mengenai peluang dari parameter yang tidak diketahui (Farahdiba, 2020). Meskipun metode analisis regresi logistik biner bayesian memiliki banyak keunggulan, tetapi terdapat kekurangan dalam menangani data yang tidak seimbang. Salah satu pendekatan yang untuk mengatasi ketidakseimbangan data adalah dengan menggunakan teknik *oversampling* seperti *adaptive synthetic* (ADASYN). Keunggulan dari *adaptive synthetic* adalah menghasilkan lebih banyak sampel sintetis untuk contoh kelas minoritas yang lebih sulit diklasifikasikan, meningkatkan kinerja klasifikasi dan mengurangi risiko *overfitting* yang sering terjadi dengan teknik *oversampling* lainnya (Ramadhan, 2021).

Penelitian terdahulu dilakukan oleh Shobri dkk. (2021) menggunakan regresi logistik biner bayesian pada kasus risiko kematian pasien covid-19 dengan menunjukkan bahwa jumlah komorbid memiliki pengaruh yang signifikan terhadap risiko kematian pasien covid-19. Penelitian lain dilakukan oleh Meliza dkk. (2020), menggunakan regresi logistik biner bayesian dalam menganalisis faktor-faktor yang berpengaruh terhadap kejadian kanker payudara menggunakan kuantil 2.5% dan kuantil 97.5%. Dhitama & Bachtiar (2020) juga telah melakukan penelitian terkait *oversampling* data menggunakan *adaptive synthetic* (ADASYN) dimana dengan menggunakan *adaptive synthetic* memiliki nilai akurasi yang lebih tinggi. Namun, penelitian tersebut belum menggabungkan pendekatan regresi logistik biner bayesian dengan penerapan *adaptive synthetic sampling* untuk menangani data tidak seimbang.

Berdasarkan uraian tersebut, penulis tertarik untuk melakukan penelitian yang menggabungkan metode regresi logistik biner bayesian dengan metode *adaptive synthetic sampling*. Penggunaan regresi logistik biner bayesian memberikan keunggulan dalam memberikan estimasi parameter yang lebih baik. Sementara itu, penerapan *adaptive synthetic sampling* (ADASYN) efektif dalam menangani ketidakseimbangan data, yang sering terjadi dalam data medis. Oleh karena itu, diusulkan penelitian berjudul "Pemodelan Regresi Logistik Biner Bayesian Dengan Metode *Adaptive Synthetic Sampling* Pada Data Penyakit Jantung Iskemik" untuk mengatasi keterbatasan tersebut dan mengembangkan model prediktif yang lebih baik dalam menganalisis faktor-faktor yang mempengaruhi penyakit jantung iskemik.

1.2 Batasan Masalah

Batasan masalah penilitian ini adalah:

1. Data yang digunakan merupakan data penyakit jantung iskemik pada bulan juli 2020 sampai bulan juli 2021 yang bersumber dari Pusat Jantung Terpadu Rumah Sakit Umum Pusat Dr. Wahidin Sudirohusodo Makassar.
2. Distribusi *prior* yang digunakan adalah *non-informatif* dengan probabilitas berdistribusi normal $\beta_j \sim N(0, 10^4)$.

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Mendapatkan dugaan model regresi logistik biner bayesian dengan metode *adaptive synthetic sampling* dalam menganalisis data penyakit jantung iskemik.
2. Mengetahui faktor-faktor yang menyebabkan terkenanya penyakit jantung iskemik menggunakan model regresi logistik biner bayesian dengan metode *adaptive synthetic sampling.*

1.4 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Bagi Peneliti, penelitian ini mampu menambah wawasan tentang analisis dengan menggunakan model regresi logistik biner bayesian dengan metode *adaptive synthetic sampling*
2. Bagi Tenaga Kesehatan, membantu meningkatkan kualitas pelayan medis dalam mendeteksi penyakit jantung iskemik.
3. Bagi Masyarakat, memberikan informasi tentang faktor-faktor yang mempengaruhi penyakit jantung iskemik, sehingga penderita dapat melakukan tindakan pencegahan sejak awal.

1.5 Teori

1.5.1. Regresi Logistik Biner

Analisis regresi merupakan salah satu analisis yang bertujuan untuk mengetahui pengaruh variable respon terhadap variabel prediktor. Secara umum, model regresi linear dapat dituliskan sebagai berikut (Tampil dkk., 2017).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

dengan $i = 1,2, \ldots, n$ dan $j = 1,2, \ldots, p$

Regresi logistik merupakan suatu metode regresi yang menggambarkan hubungan antara suatu variabel respon dan satu atau lebih variabel prediktor. Perbedaan antara model regresi logistik dengan model regresi linear adalah variabel respon dari regresi logistik bersifat *dikotomus* yang memiliki dua kategori. Dalam model ini, $\pi(x_i)$ merupakan peluang nilai sukses dari variabel prediktor x dan peluang ini merupakan parameter dari distribusi *binomial* (Yudissanta & Ratna, 2012). Apabila variabel respon terdiri dari dua kategori, metode regresi logistik yang dapat digunakan adalah regresi logistik biner (Hosmer Jr dkk., 2013).

Model dalam regresi logistik biner termasuk dalam distribusi keluarga eksponensial, distribusi eksponensial yang dimaksud adalah distribusi b*ernoulli*, yaitu distribusi peubah acak yang hanya mempunyai dua fungsi kategorik yaitu bernilai 0

dan 1 (Islamiyati, 2015). Variabel respon pada regresi logistik biner terdiri dari 2 kategori, misalkan y = 1 menyatakan hasil yang diperoleh 'sukses' dan y = 0 menyatakan hasil yang diperoleh 'gagal'. Variabel respon y mengikuti distribusi *bernoulli* karena variabel respon y hanya memiliki dua kategori (Hosmer Jr dkk., 2013) dengan fungsi kepadatan peluang distribusi *bernoulli* pada Persamaan (2).

$$f(y_i) = \pi(x_i)^{y_i}\big(1-\pi(x_i)\big)^{(1-y_i)} \; ; \; y_i = 0{,}1 \tag{2}$$

dimana jika $y_i = 0$ maka $f\,(y_i) = (1-\pi(x_i))$ dan jika $y_i = 1$ maka $f(y_i) = \pi(x_i)$, maka $\pi(x_i)$ terdapat pada Persamaan (3).

$$\pi(x_i) = \frac{\exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)}{1 + \exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)} \tag{3}$$

dengan $i = 1{,}2, \ldots, n$ dan $j = 1{,}2, \ldots, p$

Fungsi $\pi(x_i)$ dalam Persamaan (3) adalah fungsi non-linear, sehingga diperlukan transformasi *logit* untuk memperoleh fungsi yang linear dan memungkinkan analisis hubungan antara variabel respon dan variabel prediktor. Oleh karena itu, Persamaan (3) digunakan untuk mendapatkan transformasi *logit*. Nilai transformasi *logit* diperoleh dengan menyederhanakan Persamaan (3).

$$\pi(x_i) + \pi(x_i)\,\exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big) = \exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)$$

$$\pi(x_i) = \exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big) - \pi(x_i)\,\exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)$$

$$\pi(x_i) = [1 - \pi(x_i)]\,\exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)$$

$$\frac{\pi(x_i)}{[1-\pi(x_i)]} = \exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)$$

$$ln\,\frac{\pi(x_i)}{[1-\pi(x_i)]} = \ln\,[\exp\big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)]$$

$$ln\,\frac{\pi(x_i)}{[1-\pi(x_i)]} = \big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big)$$

Sehingga didapatkan nilai transformasi *logit* pada Persamaan (4).

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \big(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}\big) \tag{4}$$

dengan $i = 1{,}2, \ldots, n$ dan $j = 1{,}2, \ldots, p$

1.5.2. Estimasi Parameter

Regresi logistik biner melibatkan pengamatan yang mengikuti distribusi *bernoulli*, sehingga estimasi parameternya dapat ditentukan menggunakan fungsi *likelihood* (Fitrah dkk., 2021). Terdapat dua metode dalam mengestimasi parameter, yaitu metode klasik dan metode bayesian. Metode klasik estimasi parameter dilakukan menggunakan data sampel sebagai objek observasi tanpa mempertimbangkan distribusi awal yang disebut *prior*. Sebaliknya, metode bayesian estimasi parameter dilakukan dengan menggunakan kombinasi data sampel dan distribusi *prior*. Pendekatan bayesian menghasilkan estimasi parameter berdasarkan *mean* dan *median posterior*. (Bolstad & Curran, 2007). Berdasarkan Persamaan (3), parameter yang akan ditaksir adalah β dimana diketahui y_i berdistribusi binomial, Oleh karena

itu, fungsi kepadatan peluang dapat dilihat pada Persamaan (2) dan fungsi *likelihood* diperoleh pada Persamaan (5).

$$L(\boldsymbol{\beta}; y) = \prod_{i=1}^{n} f(y_i)$$
$$L(\boldsymbol{\beta}; y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{(1-y_i)} \quad (5)$$

1.5.3. Metode Bayesian

Estimasi dalam statistika inferensial didasarkan pada data sampel yang diambil dari populasi. Berbeda dengan pendekatan bayesian, dimana estimasi menggabungkan informasi dari data sampel dan distribusi awal disebut distribusi *prior* yang menghasilkan distribusi *posterior*, sehingga diperoleh estimasi bayesian yang merupakan *mean* dan modus dari distribusi *posterior*. Hasil yang dinyatakan dalam bentuk distribusi *posterior* yang kemudian menjadi dasar dalam metode bayesian (Ntzoufras, 2011). Analisis regresi logistik biner bayesian adalah regresi logistik biner yang diterapkan dengan metode bayesian. Metode bayesian mengestimasi parameter dengan menganggap parameter sebagai variabel acak yang mengikuti distribusi tertentu. Pendekatan ini memprediksi peluang di masa depan berdasarkan pengalaman di masa lalu (Xhemali dkk., 2009).

Penelitian yang dilakukan oleh Azhar (2012) dan Khairiyah & Diana (2018) menyatakan bahwa estimasi parameter dengan pendekatan bayesian adalah esmtimasi parameter yang lebih baik dibandingkan metode klasik seperti metode kuadrat terkecil dan *maksimum likelihood*. Metode bayesian tidak hanya menarik kesimpulan berdasarkan informasi dari *likelihood*, tetapi juga menambahkan informasi subjektif mengenai peluang parameter yang tidak diketahui, yang disebut distribusi *prior* (Shobri dkk., 2021).

A. Distribusi *prior*

Analisis bayesian pada suatu populasi mengikuti distribusi tertentu dengan suatu parameter didalamnya, parameter tersebut dapat mengikuti distribusi peluang tertentu yang disebut distribusi *prior* (Bain & Engelhardt, 1992). Distribusi *prior* merupakan distribusi awal yang memberikan informasi mengenai parameter. Distribusi *prior* mencerminkan kepercayaan subjektif parameter sebelum sampel diambil (Wijaya & Wulandari, 2016). Distribusi *posterior* akan bergantung pada pemilihan distribusi *prior*. Dengan demikian, permasalahan utama adalah memilih distribusi *prior* untuk parameter yang tidak diketahui namun relevan dengan permasalahan data penelitian. Pemilihan distribusi *prior* dapat didasarkan pada ruang parameternya (Shobri dkk., 2021).

Pada dasarnya, distribusi *prior* adalah representasi subjektif peneliti terhadap suatu nilai parameter yang diduga. Distribusi *prior* yang berkaitan dengan bentuk distribusi hasil identifikasi pola datanya dikelompokkan menjadi:

a) Distribusi konjugat

   Distribusi *prior* konjugat mengacu pada acuan analisis model terutama dalam pembentukan fungsi *likelihood* sehingga dalam penentuan *prior* konjugat selalu dipikirkan mengenai pembentukan pola distribusi *prior* yang mempunyai fungsi kepekatan peluang pembangun *likelihood*.

b) Distribusi tidak konjugat

Apabila pemberian distribusi *prior* pada suatu model tidak mempertimbangkan pola pembentuk fungsi *likelihood*.

Berdasarkan penentu parameter pada pola distribusinya, distribusi *prior* dikelompokkan menjadi:

a) Distribusi *informatif*

Distribusi *prior informatif* mengacu pada pemberian parameter dari distribusi *prior* yang telah dipilih, baik distribusi *prior* konjugat atau tidak konjugat. Pemberian nilai parameter pada distribusi *prior* ini akan sangat mempengaruhi bentuk distribusi *posterior* yang akan didapat pada informasi data yang diperoleh.

b) Distribusi *non - informatif*

Distribusi *prior non-informatif*, informasi mengenai parameter tidak tersedia, pemilihannya tidak didasarkan pada data yang ada atau distribusi *prior*nya tidak mengandung informasi tentang parameter (Shobri dkk., 2021).

Penelitian ini menggunakan distribusi *non-informatif* dengan distribusi normal. Persamaan distribusi normal yang digunakan dapat dilihat pada Persamaan (6).

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)\right) \quad (6)$$

Distribusi *non-informatif* dari distribusi normal tersebut dijelaskan lebih lanjut dalam Persamaan (7).

$$f(\beta_j) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{(\beta_j-\mu)^2}{\sigma^2}\right)\right); j = 1,2,\dots,p \quad (7)$$

B. Distribusi *Posterior*

Distribusi *posterior* merupakan distribusi yang menyatukan informasi data sampel dengan informasi distribusi *prior* dengan teorema bayesian. Distribusi dinyatakan sebagai fungsi kepadatan peluang bersyarat dari β dengan pengamatan $\boldsymbol{Y}$**,** dengan $\boldsymbol{Y} = y_1, y_2, \dots, y_n$ (Shobri dkk., 2021). Dalam estimasi bayesian, setelah informasi tentang sampel diperoleh dan *prior* dapat ditentukan maka distribusi *posterior* dapat dicari dengan cara mengalikan *prior* dengan informasi sampel yang diperoleh dari *likelihood* (Bolstad & Curran, 2007).

Box dan Tiao (1973) mendefinisikan bahwa distribusi *posterior* $f(\beta_j|y)$ adalah fungsi kepadatan bersyarat parameter β jika diketahui nilai observasi y, dapat ditulis pada Persamaan (8).

$$f(\beta_j|y) = \frac{f(y,\beta_j)}{f(y)} \quad (8)$$

dengan $f(y,\beta_j)$ adalah fungsi kepadatan bersama dari y dan β sedangkan $f(y)$ merupakan distribusi marginal y. Fungsi kepadatan bersama $f(y,\beta_j)$ merupakan perkalian dua fungsi kepadatan yaitu distribusi *prior* $(f(\beta_j))$ dan distribusi data $(f(y|\beta_j))$, yang ditulis pada Persamaan (9).

$$f(y,\beta_j) = f(y|\beta_j)\, f(\beta_j) \tag{9}$$

Sedangkan distribusi marginal y dapat dihitung dengan

$$f(y) = \begin{cases} \int f(y,\beta_j)d\beta = \int f\left(y\middle|\beta_j\right) f(\beta_j)\, d\beta, & \beta\ kontinu \\ \sum_\beta f(y,\beta_j)d\beta = \sum_\beta f\left(y\middle|\beta_j\right) f(\beta_j), & \beta\ diskrit \end{cases}$$

Distribusi parameter β yaitu $f(\beta_j)$ disebut sebagai *prior* dan $f(y|\beta_j)$ sebagai *likelihood* yang merupakan fungsi parameter dari β. Karena $f(y)$ tidak bergantung pada β, maka $\frac{1}{f(y)}$ dapat dianggap konstan, misalkan C. dengan kata lain, $f(y)$ adalah konstanta yang disebut *normalized constant*. Sehingga *posterior* dapat ditulis pada Persamaan (10).

$$f(\beta_j|y) = L(y|\beta_j) f(\beta_j) \tag{10}$$

Persamaan (10) menunjukkan bahwa *posterior* adalah proporsional terhadap *likelihood* dikalikan dengan *prior* dari parameter model. Nilai tengah dari distribusi *posterior* yang akan digunakan untuk menentukan estimator dari parameter yang tidak diketahui. Fungsi kepadatan peluang dari β jika diketahui contoh pengamatan $\boldsymbol{Y} = y_1, y_2, \dots, y_n$ yang terdapat pada Persamaan (11).

$$f(\beta_j|\boldsymbol{Y},\boldsymbol{X}) = \frac{L(\boldsymbol{Y},\boldsymbol{X};\beta) \times f(\beta_j)}{\int_{-\infty}^{\infty} L(\boldsymbol{Y},\boldsymbol{X};\beta) \times f(\beta_j) d\beta_j} \tag{11}$$

Adapun distribusi *posterior* dalam penyelesaiannya tidak dapat diselesaikan secara analitis sehingga dibutuhkan simulasi yaitu *markov chain monte carlo* yang memungkinkan pengambilan sampel numerik dari distribusi *posterior* yang mendasarinya dan digunakan untuk memperbarui parameter (Sumae dkk., 2022).

1.5.4. *Markov Chain Monte Carlo*

Markov chain monte carlo (MCMC) adalah suatu teknik metode simulasi yang membangkitkan sejumlah sampel dari distribusi data tertentu untuk mendapatkan distribusi *posterior* (Ntzoufras, 2011). Analisis bayesian dapat dipermudah dengan penggunaan MCMC, sehingga keputusan dari hasil analisis dapat diambil dengan cepat dan akurat. Terdapat dua keuntungan utama yang diperoleh dari penggunaan metode MCMC pada analisis bayesian. Pertama, metode MCMC dapat menyederhanakan integral yang kompleks dengan dimensi besar menjadi integral yang lebih sederhana dengan satu dimensi. Kedua, metode MCMC memungkinkan estimasi densitas data dengan membangkitkan rantai markov yang cukup besar, sebanyak N iterasi (Iriawan, 2001).

MCMC pada dasarnya adalah metode integrasi monte carlo yang menggunakan rantai markov. Metode ini adalah teknik simulasi untuk menarik sampel dari distribusi *posterior*. MCMC menarik sampel secara berturut-turut dari distribusi target, membangkitkan data sampel parameter β yang mengikuti distribusi tertentu melalui algoritma tertentu. Proses ini mengandalkan nilai pada setiap langkah yang bergantung pada nilai langkah sebelumnya, membentuk rantai markov (Gilks & Roberts, 1996). Terdapat dua metode MCMC yang paling populer yaitu algoritma *metropolis-hastings* dan *gibbs sampling*. Metode *gibbs sampling* merupakan teknik yang sering dipakai oleh pengguna metode bayesian (Castanheira dkk., 2004). *Gibbs sampling* adalah teknik simulasi yang digunakan untuk

membangkitkan variabel acak dari distribusi tertentu secara langsung tanpa perlu menghitung densitasnya. Teknik ini memungkinkan menghindari perhitungan yang sulit (Casella & George, 1992).

Gibbs sampling adalah algoritma MCMC yang melibatkan sampling iteratif dari distribusi bersyarat dimana parameter β dipartisi menjadi beberapa bagian $\beta = \beta_1, \beta_2, \dots, \beta_p$. Distribusi *posterior full* counditional adalah $p = (\beta_p | x, \beta_1, \beta_2 \dots, \beta_{p-1})$.*Gibbs sampling* bekerja dengan Langkah-langkah sebagai berikut.

1. Menentukan nilai awal dari $\beta^{(0)} = \beta_1^{(0)}, \beta_2^{(0)}, \dots, \beta_p^{(0)}$
2. Bangkitkan nilai $\beta_1^{(1)}$ berasal dari $f(\beta_1 | \beta_2^{(0)}, \beta_3^{(0)}, \dots, \beta_p^{(0)}$,
   nilai $\beta_2^{(1)}$ berasal dari $f(\beta_2 | \beta_1^{(0)}, \beta_3^{(0)}, \dots, \beta_p^{(0)}$
   $\vdots$
   nilai $\beta_p^{(1)}$ berasal dari $f(\beta_p | \beta_1^{(0)}, \beta_2^{(0)}, \dots, \beta_{p-1}^{(0)}$ sehingga diperoleh parameter baru $\beta_1^{(1)}, \beta_2^{(1)}, \dots, \beta_p^{(1)}$
3. Menggunakan $\beta_1^{(1)}, \beta_2^{(1)}, \dots, \beta_p^{(1)}$ untuk peroses selanjutnya sampai k iterasi yang mencapai konvergen.
4. Memeriksa konvergen dengan melihat *trace plot, density posterior, dynamic trace* dan *autocorrelation plot* dikatakan konvergen ketika semua menunjukkan indikasi bahwa rantai markov telah melewati seluruh ruang parameter dengan baik dan bahwa sampel yang dihasilkan benar-benar berasal dari distribusi posterior yang diinginkan.
5. Mendapatkan *mean*, *median*, *standar deviasi* dan distribusi *posterior*.

1.5.5. Uji Signifikansi Parameter

Uji signifikansi parameter dilakukan untuk mengetahui kelayakan model dengan menguji signifikan tiap variabel respon yang digunakan dalam model. Tidak adanya pengaruh perlakuan dinyatakan dengan hipotesis null (H_0) yang nilai parameternya menyebabkan perlakuan bernilai nol. Hipotesis lawannya yang menyatakan sebuah perlakuan tidak bernilai nol disebut hipotesis alternatif (H_1). Uji hipotesis untuk bayesian menggunakan yang bentuk sederhananya dapat ditujukan oleh kuantil 2.5% dan 97.5%. Apabila nilai mean β dari proses simulasi berada di dalam kuantil tersebut, maka dapat disimpulkan tidak ada pengaruh perlakuan terhadap variabel respon (H_0) ditolak. Sebaliknya, apabila nilai mean berada di luar kuantil, maka disimpulkan ada pengaruh perlakuan terhadap variabel respon (H_0) diterima (Robert & Ntzoufras, 2012).

1.5.6. Uji Kesesuain Model

Uji kesesuaian model yang dipakai yaitu, *Mean squared error* (MSE). MSE adalah metode alternatif untuk mengevaluasi teknik peramalan dimana masing-masing kesalahan (selisih data aktual terhadap data peramalan) dikuadratkan, dijumlahkan dan dibagi dengan jumlah data (Pramana dkk., 2022). Perhitungan *mean squared error* dilakukan menggunakan Persamaan (12).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

1.5.7. *Odds Ratio*

Odds ratio adalah perbandingan risiko terjadinya suatu dari kejadian suatu kategori yang satu terhadap kategori yang lain. *Odds ratio* memudahkan interpretasi model regresi logistik yang diperoleh. Interpretasi parameter bertujuan untuk memahami makna nilai taksiran parameter pada variabel prediktor (Hosmer Jr dkk., 2013). *Odds ratio* berhubungan dengan transformasi *logit*, seperti yang diketahui agar menjadi bentuk yang linear fungsi logistik perlu ditransformasi sedemikian rupa. Transformasi *logit* dapat dituliskan.

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Sehingga model *odds ratio* dituliskan pada persamaan (13).

$$OR = \frac{\pi(x_i)}{1-\pi(x_i)} = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \quad (13)$$

1.5.8. *Confusion Matrix*

Confusion matrix adalah alat yang berfungsi untuk menganalisis seberapa model klasifikasi mengenali *tuple* dari data yang berbeda (Probo & Irawan, 2016).

Tabel 1. Confusion matrix

	Positive	Negatif	
Positive	TP	FN	TP+FN
Negatif	FP	TN	FP+TN
	TP+FP	FN+TN	

Akurasi adalah proporsi dari total prediksi *true* dari semua data. Rumus akurasi terdapat pada Persamaan (14) (Nasution & Hayaty, 2019):

$$Akurasi\ (Accuration) = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (14)$$

Keterangan :

- TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
- TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
- FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
- FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

1.5.9. Standarisasi Data

Standarisasi data dilakukan untuk mengurangi variasi data antar variabel karena setiap variabel memiliki satuan yang berbeda-beda. Perbedaan satuan dapat mempengaruhi hasil analisis karena rentang nilai yang besar antar variabel (Bau dkk., 2023). Sebagai contoh, jika variabel penghasilan mempunyai satuan juta (0000.0000) sedangkan usia seseorang hanya mempunyai satuan puluhan (00), perbedaan mencolok ini dapat membuat perhitungan jarak menjadi tidak valid.

Masalah ini dapat diatasi dengan teknik standarisasi yang melibatkan penentuan nilai rata-rata dan varians, seperti yang dijelaskan dalam Persamaan (15) (Athifaturrofifah dkk., 2019).

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{15}$$

1.5.10. *Adaptive Synthetic Sampling*

Pendekatan yang digunakan untuk mengatasi ketidakseimbangan kelas adalah *oversampling*, yaitu dengan memperbanyak sampel dari kelas minoritas hingga jumlahnya sebanding dengan kelas mayoritas (Permana dkk., 2024). Metode *adaptive synthetic sampling* adalah salah satu teknik yang digunakan untuk menyeimbangkan data pada dataset yang tidak seimbang. *Adaptive synthetic sampling* menggunakan bobot distribusi pada data kelas minoritas berdasarkan tingkat kesulitan yang dihadapi model dalam mengklasifikasikan data. Data sintesis dihasilkan dari kelas minoritas yang sulit untuk diklasifikasikan dibandingkan dengan data minoritas yang lebih mudah untuk diklasifikasikan (Dhitama & Bachtiar, 2020). *Adaptive synthetic sampling* meningkatkan kinerja model dengan dua cara. Pertama, mengurangi bias yang diakibatkan oleh ketidakseimbangan kelas dan yang kedua secara adaptif menggeser batas keputusan klasifikasi terhadap kesulitan data (Rahayu dkk., 2017). Langkah pembangkitan data sintetis dengan *adaptive synthetic sampling* adalah sebagai berikut:

1. Menentukan nilai parameter dari *adaptive synthetic sampling*, yaitu nilai dth (nilai dari maksimal toleran data *imbalance*) dan α (nilai level keseimbangan).
2. Menghitung derajat keseimbangan yang terdapat pada Persamaan (16).
$$d = \frac{m_{minoritas}}{m_{mayoritas}} \tag{16}$$
3. Menghitung banyaknya *instance* data sintetis yang akan dibuat untuk kelas minoritas yang terdapat pada Persamaan (17).
$$G = \left(m_{mayoritas} - m_{minoritas}\right) \times \alpha \tag{17}$$
4. Menghitung rasio berdasarkan K-*Nearest Neighbor* menggunakan *euclidean distance* yang terdapat pada Persamaan (18).
$$r_i = \frac{\Delta_i}{k} \tag{18}$$
dengan Δ_i adalah banyaknya *instance* pada *nearest neighbors* yang termasuk kedalam kelas, dan $i = 1, 2, 3, \ldots, m_{minoritas}$
5. Normalisasi r_i sehingga $\hat{r}_i$ adalah distribusi kerapatan yang terdapat pada Persamaan (19).
$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_{minoritas}} r_i} \tag{19}$$
6. Menghitung banyaknya *instance* data sintetis yang perlu dibangkitkan untuk setiap *instance* minoritas yang terdapat pada Persamaan (20).
$$g_i = \hat{r}_i \times G\,;\; i = 1, 2, 3, \ldots, m_{minoritas} \tag{20}$$

7. Pembangkitan sampel data sintetis dilakukan dengan menggunakan Persamaan (21).

$$s_i = x_i + (x_{ui} - x_i) \times \lambda \quad (21)$$

dengan x_i merupakan data pengamatan ke-i dari kelas minor, x_{ui} adalah data ke-i dari data latih yang dipilih secara acak, dan lambda (λ) adalah bilangan random antara 0 dan 1.

1.5.11. Penyakit Jantung Iskemik

Penyakit jantung adalah salah satu dari penyakit yang dapat menyebabkan kematian yang tinggi selain stroke, kanker paruparu, kanker payudara, dan AIDS (Khairunnisa dkk., 2023). Penyakit jantung dan pembuluh darah merupakan penyebab utama kematian di dunia. Insiden kematian mendadak akibat gangguan ini sangat tinggi. Prevalensi penyakit jantung di Indonesia sebesar 1,5%. Penyakit Jantung Iskemik (PJI) dikenal sebagai penyakit arteri koroner (PAK), didefinisikan sebagai kekurangan oksigen dan penurunan atau tidak adanya aliran darah ke miokardium yang disebabkan oleh penyempitan arteri koroner. Hipertensi merupakan faktor risiko penting terhadap PJI. Tekanan darah tinggi yang terus menerus dapat menyebabkan kerusakan pada sistem pembuluh darah arteri (Rukminingsih & Dewi, 2020).

Penyakit jantung iskemik terjadi ketika aliran darah ke bagian jantung berkurang atau terhenti akibat penyempitan arteri koroner, yang biasanya disebabkan oleh penumpukan plak aterosklerotik. Plak ini terdiri dari kolesterol, lemak, dan zat lainnya yang menyempitkan arteri dan mengurangi aliran darah ke jantung. Hal ini dapat mengakibatkan nyeri dada (angina) atau serangan jantung (infark miokard). Faktor risiko lain termasuk merokok, diabetes, obesitas, dan gaya hidup tidak aktif termasuk kedalam faktor yang mempengaruhi terkenanya penyakit jantung iskemik (American Heart Association, 2021). Pada tahun 2019, jumlah kasus penyakit jantung iskemik di dunia mencapai 197 juta kasus dengan jumlah kematian sebanyak 9,74 juta kematian. Di Indonesia, berdasarkan data *institute for health metrics and evaluation* (IHME) ada 245.343 kematian per tahun akibat penyakit jantung iskemik (Fajriati & Prasetiyo, 2023).

BAB II
METODOLOGI PENELITIAN

2.1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder rekam medis pasien pada bulan Juli 2021 sampai Juli 2022 yang bersumber dari Pusat Jantung Terpadu Rumah Sakit Umum Pusat Dr. Wahidin Sudirohusodo Makassar. Jumlah data yang digunakan pada penelitian ini sebanyak 230 data dan terdapat 11 variabel. Data dapat dilihat pada Lampiran 1.

2.2. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri atas satu variabel respon dan sepuluh variabel prediktor. Adapun rincian variabel yang digunakan adalah sebagai berikut.

Tabel 2. Variabel respon dan prediktor

Variabel	Indikator	Keterangan	Skala
y	Jantung	1: Pasien yang terkena penyakit jantung 0: Pasien yang tidak terkena penyakit jantung	Nominal
X_1	Usia	Umur dari pasien penyakit jantung dalam tahun	Rasio
X_2	Berat Badan	Berat Badan pasien dalam satuan (kg)	Rasio
X_3	Indeks Masa Tubuh	Indeks Masa Tubuh dalam satuan (kg/m^2)	Rasio
X_4	Hemoglobin	Jumlah hemoglobin dalam satuan (gr/dL)	Rasio
X_5	Hematokrit	Jumlah hematokrit dalam satuan (%)	Rasio
X_6	Leokosit	Jumlah leokosit dalam satuan $(sel/\mu L)$	Rasio
X_7	Trombosit	Jumlah trombosit dalam tubuh dalam satuan $(ribu/\mu L)$	Rasio
X_8	Hipertensi	1: Pasien yang memiliki riwayat hipertensi 0: Pasien yang tidak memiliki riwayat hipertensi	Nominal
X_9	Kolestrol Total	Kolestrol total dalam satuan (mg/dl)	Rasio
X_{10}	Jenis Kelamin	1: Pasien berjenis kelamin laki-laki 0: Pasien berjenis kelamin perempuan	Nominal

2.3. Metode Analisis

Pendekatan analisis yang digunakan dalam penelitian ini menggunakan regresi logistik biner bayesian dengan *adaptive synthetic sampling*. Adapun tahapan analisis data yang digunakan dalam mencapai tujuan penelitian ini adalah sebagai berikut:

1. Mendeskripsikan data yang telah didapatkan dengan mencari nilai mean, median, standard deviasi, nilai minimum dan maksimum pada data penyakit jantung iskemik di Rumah Sakit Umum Pusat Dr. Wahidin Sudirohusodo Makassar.
2. Melakukan standarisasi data pada variabel prediktor dengan menentukan nilai rata-rata dan varian pada Persamaan (15) untuk membantu memastikan analisis yang lebih akurat dan valid.
$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{15}$$
3. Melakukan *adaptive synthetic sampling* untuk menyeimbangkan data pada variabel respon sehingga tidak ada ketidakseimbangan pada data. Adapaun cara melakukan *adaptive synthetic sampling* dapat dilihat pada Persamaan (16), Persamaan (17), Persamaan (18), Persamaan (19), Persamaan (20) dan Persamaan (21).
4. Mengestimasi parameter model regresi logisti biner bayesian.
   a. Menentukan formula dari fungsi *likelihood* seperti pada Persamaan (6).
$$L(\boldsymbol{\beta}; y) = \prod\prod_{i=1}^{n} \pi(x_i)^{y_i}\big(1 - \pi(x)\big)^{(1-y_i)} \tag{6}$$
   b. Menetapkan distribusi *prior* dan distribusi *posterior* seperti pada Persamaan (7) dan Persamaan (11).
$$f(\beta_j) = \frac{1}{\sqrt{2\pi\sigma^2\beta_j}} \exp\left(-\frac{1}{2}\left(\frac{(\beta_j - \mu\beta_j)^{\,2}}{\sigma^2\beta_j}\right)\right); j = 1,2,\dots,p \tag{7}$$
$$f(\beta_j|\boldsymbol{Y},\boldsymbol{X}) = \frac{L(\boldsymbol{Y},\boldsymbol{X};\beta) \times f(\beta_J)}{\int_{-\infty}^{\infty} L(\boldsymbol{Y},\boldsymbol{X};\beta) \times f(\beta_J) d\beta_j} \tag{11}$$
   c. Melakukan simulasi *markov chain monte carlo* menggunakan algoritma *gibbs sampling*, yaitu pendekatan numerik yang membantu menemukan distribusi dari parameter yang diduga memiliki bentuk yang rumit.
5. Menggunakan uji signifikan parameter dengan pendekatan interval konvidensi 95% dari masing-masing parameter yang dihitung dengan batas bawah yaitu kuantil 2.5% dan batas atasnya adalah 97.5% pada setiap variabel prediktor, untuk variabel yang signifikan maka model regresi logistik biner bayesian terbentuk.
6. Menghitung nilai akurasi pada model yang terbentuk menggunakan Persamaan (14) untuk melihat seberapa layak model bisa digunakan dalam menganalisis.
$$Akurasi\ (Accuration) = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{14}$$
7. Menghitung nilai *mean square error* pada model yang terbentuk menggunakan Persamaan (12) untuk mengukur kinerja dari regresi logistik biner bayesian.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (12)$$

8. Menghitung dan menginterpertasi nilai *odds ratio* pada Persamaan (13) untuk melihat adanya peningkatan atau penurunan peluang pada kategori status penyakit jantung iskemik dari masing-masing variabel prediktor yang signifikan.

$$OR = \frac{\pi(x_i)}{1-\pi(x_i)} = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \quad (13)$$