MODEL REGRESI LOGISTIK *PRINCIPAL COMPONENT ANALYSIS* PADA PREDIKTOR KATEGORIK

SKRIPSI

DWI AULIYAH

H051171310

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

MODEL REGRESI LOGISTIK *PRINCIPAL COMPONENT ANALYSIS* PADA PREDIKTOR KATEGORIK

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains
pada Program Studi Statistika Departemen Statistika Fakultas
Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

DWI AULIYAH

H051171310

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2021

LEMBAR PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	: Dwi Auliyah
NIM	: H051171310
Program Studi	: Statistika
Jenjang	: Sarjana (S1)

Menyatakan dengan ini bahwa karya tulis saya yang berjudul

MODEL REGRESI LOGISTIK *PRINCIPAL COMPONENT ANALYSIS* PADA PREDIKTOR KATEGORIK

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 20 Agustus 2021

DWI AULIYAH

NIM. H051171310

LEMBAR PENGESAHAN

MODEL REGRESI LOGISTIK *PRINCIPAL COMPONENT ANALYSIS* PADA PREDIKTOR KATAGORIK

Disusun dan diajukan oleh

DWI AULIYAH
H051171310

Telah dipertahankan dihadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin
pada tanggal 20 Agustus 2021
dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pertama,
Dr. Anna Islamiyati, S.Si., M.Si.	Sitti Sahriman, S.Si., M.Si.
NIP. 19770808 200501 2 002	NIP. 19881018 201504 2 002

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.
NIP. 19720117 199703 2002

KATA PENGANTAR

Assalamu'alaikum Warohmatullahi Wabarokatuh.

Alhamdulillahi robbil'alamin, Puji syukur kepada Allah *Subhanahu Wa Ta'ala* atas segala limpahan rahmat, nikmat, dan hidayah yang diberikan kepada penulis sehingga dapat menyelesaikan penulisan skripsi dengan judul "Model Regresi Logistik *Principal Component Analysis* Pada Prediktor Kategorik" sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam.

Salam dan sholawat *InsyaAllah* senantiasa tercurah kepada Nabi Muhammad *Shallallahu'alaihi Wasallam*, sang kekasih tercinta yang telah memberikan petunjuk cinta dan kebenaran dalam kehidupan.

Dalam penyelesaian skripsi ini, penulis telah melewati perjuangan panjang dan pengorbanan yang tidak sedikit. Namun berkat rahmat dan izin-Nya serta dukungan dari berbagai pihak yang turut membantu baik moril maupun material sehingga akhirnya tugas akhir ini dapat terselesaikan. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang setinggi-tingginya dan penghargaan yang tak terhingga kepada Ayahanda *Demi Yazis* dan Ibunda tercinta *Yetti Silviyandri* yang telah membesarkan dan mendidik penulis dengan penuh kesabaran dan dengan limpahan cinta, kasih sayang, dan doa kepada penulis yang tak pernah habis, serta saudara-saudara penulis *Hayyu Wulandari* dan *Yuki Hilmi Yazis* yang selalu membantu jika ada kendala selama penulisan dan menjadi penyemangat untuk segera menyelesaikan masa studi penulis.

Ucapan terima kasih dengan penuh keikhlasan juga penulis ucapkan kepada:

1. Ibu Prof. Dr. Dwia Aries Tina Pulubuhu, MA, selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya.
2. Bapak Dr. Eng. Amiruddin, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh jajarannya.
3. Ibu Dr. Nurtiti Sunusi, S.Si., M.Si., selaku Ketua Departemen Statistika yang telah seperti orang tua sendiri. Segenap dosen pengajar dan staf Departemen Statistika yang telah membekali ilmu dan kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika.

4. Ibu Dr. Anna Islamiyati, S.Si. M.Si. selaku Pembimbing Utama dan Dosen Pembimbing Akademik yang telah seperti orang tua sendiri yang sangat ikhlas meluangkan waktu dan pemikirannya untuk memberikan arahan, pengetahuan, motivasi dan bimbingan ditengah kesibukan beliau serta menjadi tempat berkeluh kesah untuk penulis.
5. Ibu Sitti Sahriman, S.Si., M.Si. selaku Pembimbing Pertama sekaligus penasehat akademik penulis yang telah meluangkan waktunya ditengah kesibukan untuk memberikan arahan bagi penulis.
6. Ibu Anisa, S.Si., M.Si dan Bapak Siswanto, S.Si., M.Si selaku tim penguji yang telah memberikan saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini.
7. Sahabat tercinta yang menemani, mendengar, serta memberikan solusi atas keluh kesah penulis selama perkuliahan Cici Pulcerima.
8. Spesial untuk sahabat tercinta penulis selama perkuliahan, Syafira, Fadillah, Haura, Nurlia, Hana, yang telah menjadi sahabat terbaik sejak awal perkuliahan dan senantiasa mendengarkan curhatan, memberikan dorongan, semangat, dan motivasi dalam setiap keadaan sehingga penulis bisa mendapatkan lebih banyak pelajaran hidup.
9. Teman-teman Statistika 2017, terima kasih atas kebersamaan, suka, dan duka selama menjalani pendidikan di Departemen Statistika.
10. Kepada seluruh pihak yang tidak dapat penulis sebutkan satu persatu, terima kasih setinggi-tingginya untuk segala dukungan dan partisipasi yang diberikan kepada penulis semoga bernilai ibadah di sisi Allah *Subhanahu Wa Ta'ala.*

Penulis berharap skripsi ini dapat memberikan tambahan pengetahuan baru bagi para pembelajar statistika. Penulis menyadari bahwa dalam penulisan tugas akhir ini masih banyak terdapat kekurangan. Oleh karena itu, dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga dapat bermanfaat bagi pihak-pihak yang berkepentingan. *Aamiin Yaa Rabbal Alamin.*

Makassar, 20 Agustus 2021

Dwi Auliyah

ABSTRAK

Regresi logistik biner adalah analisis yang digunakan untuk memprediksi peubah respon yang mengandung dua kategorik atau memiliki dua kemungkinan hasil berdasarkan peubah prediktor yang berskala kategorik atau numerik. Ketika terjadi multikolinearitas pada variabel prediktor yang berskala kategorik, pendekatan statistika yang dapat digunakan untuk mengatasi masalah tersebut adalah metode *principal component analysis* pada prediktor kategorik. Estimasi parameter regresi logistik *principal component analysis* pada prediktor kategorik dilakukan melalui metode *maximum likelihood estimation*. Estimasi parameter yang dihasilkan merupakan bentuk persamaan implisit yang sulit diselesaikan dengan cara analitik sehingga digunakan iterasi *Newton Raphson.* Adapun pengujian parameter yang digunakan dalam model regresi logistik *principal component analysis* pada prediktor kategorik adalah uji *Likelihod Ratio*. Selanjutnya, model regresi logistik *principal component analysis* pada prediktor kategorik diterapkan pada data kadar gula darah masyarakat di Kabupaten Muna Barat. Dalam penelitian ini diperoleh nilai probabilitas responden yang memiliki kadar gula darah normal dan kadar gula darah terdampak penyakit diabetes dengan ketepatan klasifikasi 84,6547%.

Kata Kunci : Kadar Gula Darah, Multikolinieritas, *Principal Component Analysis* Pada Prediktor Kategorik, Regresi Logistik Biner, *Maximum Likelihood Estimation*.

ABSTRACT

Binary logistic regression is an analysis used to predict the response variable that contains two categorical or only having two possible outcomes based on predictor variables on a categorical or numerical scale. When multicollinearity occurs in predictor variables on a categorical scale, a statistical approach that can be used to overcome this problem is the principal component analysis method on categorical predictor. The estimation of principal component analysis logistic regression parameter on categorical predictors is done through the maximum likelihood estimation method. The resulting parameter estimate is a form of implicit equation that is difficult to solve analytically, so Newton Raphson iteration is used. The parameter testing used in the principal component analysis logistic regression model on categorical predictor is the Likelihood Ratio test. Furthermore, the principal component analysis logistic regression model on categorical predictor is still applied to the data on blood sugar levels in the people of West Muna Regency. In this study, the probability value of respondents who had normal blood sugar levels and blood sugar levels affected by diabetes was obtained with a classification accuracy of 84.6547%.

Keywords : *Blood Sugar Level, Multicollinearity, Principal Component Analysis On Categorical Predictor, Binary Logistic Regression, Maximum Likelihood Estimation.*

DAFTAR ISI

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Regresi logistik merupakan analisis regresi yang bertujuan untuk memprediksi peubah respon kategorik dengan beberapa peubah prediktor kategorik atau kontinu (Agresti, 2002). Analisis regresi logistik biner merupakan salah satu analisis regresi logistik yang digunakan untuk menganalisis hubungan antara peubah prediktor dengan peubah respon berskala dikotomi. Skala dikotomi adalah skala data nominal dengan dua kategori. Wahyuni dkk (2018) telah menggunakan model regresi logistik biner dengan peubah prediktor berskala campuran dalam memodelkan keputusan penerimaan beasiswa PPA. Safitri dkk (2019) telah menggunakan model regresi logistik biner dengan peubah prediktor kategorik dalam melihat faktor–faktor yang mempengaruhi tingkat pengangguran terbuka di Provinsi Sulawesi Barat. Akan tetapi, penelitian tersebut belum mempertimbangkan masalah multikolinearitas yang dapat saja terjadi pada data. Salah satu asumsi dalam analisis regresi logistik adalah tidak terjadi multikolinieritas pada peubah prediktor (Ohyver, 2013).

Multikolinieritas terjadi apabila terdapat hubungan atau korelasi diantara beberapa atau seluruh peubah prediktor (Soemartini, 2008). Analisis regresi memiliki kelemahan dalam prediksi pada saat terjadi multikolinieritas pada peubah prediktor. Jika ada multikolinieritas diantara peubah prediktor, maka koefisien regresi yang dihasilkan dalam suatu analisis menjadi sangat lemah sehingga tidak dapat memberikan hasil yang mewakili sifat atau pengaruh dari peubah prediktor (Montgomery dan Hines, 1990). Hal yang sama juga terjadi pada regresi logistik. Aguilera (2005) menjelaskan bahwa estimasi parameter model regresi logistik tidak akurat dan interpretasi *odds* rasio dapat keliru ketika ada multikolinieritas diantara peubah prediktor. Metode yang dapat digunakan untuk mengatasi multikolinieritas tanpa harus mengeluarkan peubah prediktor yang berkorelasi yaitu metode *principal component analysis* (Soemartini, 2008).

Metode *principal component analysis* bertujuan mereduksi dimensi untuk mengatasi masalah multikolinearitas pada data. Pada dasarnya *principal component*

analysis mentransformasi secara linier variabel prediktor yang umumnya saling berkorelasi menjadi sejumlah variabel yang lebih sedikit dan tidak saling berkorelasi yang disebut komponen utama. Setelah terbentuk beberapa komponen hasil *principal component analysis* yang bebas multikolinearitas, komponen-komponen tersebut menjadi variabel prediktor baru yang selanjutnya diregresikan pengaruhnya terhadap variabel respon (Ifadah, 2011).

Aguilera dkk (2005) telah menggunakan *principal component analysis* untuk mengatasi multikolinieritas pada regresi logistik. Islamiyati (2015) telah mengembangkan komponen utama untuk data campuran pada regresi logistik biner. Penelitian-penelitian tersebut melibatkan variabel prediktor yang berskala kuantitatif dan campuran. Namun pada data riil, sering ditemukan data dengan prediktor yang berjenis kategorik, misalnya data kuesioner.

Berdasarkan uraian diatas, penulis menyusunnya dalam sebuah penelitian dengan judul **"Model Regresi Logistik *Principal Component Analysis* Pada Prediktor Kategorik"**. Model akan digunakan pada data kadar gula darah masyarakat di Kabupaten Muna Barat. Kadar gula darah merupakan gula yang terdapat di dalam darah yang dapat menentukan terdampak atau tidaknya seseorang pada penyakit diabetes. Diabetes adalah penyakit metabolik akibat pankreas tidak dapat menggunakan hormon yang mengatur keseimbangan kadar gula darah secara efektif. Diabetes merupakan penyebab kematian terbesar nomor 3 di Indonesia dengan persentase sebesar 6,7% setelah penyakit jantung koroner (12,9%) dan stroke (21,1%). Faktor resiko yang dapat menyebabkan diabetes melitus diantaranya, tekanan darah sistolik, tekanan darah diastolik, lama waktu tidur, gaya belajar/bekerja, tingkat pengetahuan tentang diabetes, kebiasaan berolahraga, faktor keturunan diabet, indeks massa tubuh, usia, dan ukuran lingkar perut (Kemenkes, 2020). Faktor resiko diabetes ini dapat digunakan sebagai peubah prediktor pada penerapan model regresi logistik *principal component analysis* dengan skala satuan kategorik berdasarkan peubah respon dalam penelitian ini yaitu kadar gula darah yang terdiri dari dua kategori yaitu 1 (mewakili kadar gula darah tidak normal) dan 0 (mewakili kadar gula darah normal).

1.2 Rumusan Masalah

Rumusan masalah dalam penelitan ini berdasarkan latar belakang yang telah diuraikan adalah:

1. Bagaimana estimasi model regresi logistik *principal component analysis* pada prediktor kategorik yang bersesuaian dengan data kadar gula darah masyarakat di Kabupaten Muna Barat?
2. Bagaimana model hubungan antara kadar gula darah masyarakat di Kabupaten Muna Barat dengan faktor resiko diabetes berdasarkan regresi logistik *principal component analysis* pada prediktor kategorik?

1.3 Batasan Masalah

Data yang digunakan pada penelitian ini adalah data kadar gula darah, tekanan darah sistolik, tekanan darah diastolik, gaya bekerja, lama waktu tidur, riwayat keturunan penyakit diabetes, tingkat pengetahuan tentang penyakit diabetes, usia, indeks massa tubuh, dan ukuran lingkar perut masyarakat di Kabupaten Muna Barat. Selain itu, proporsi kumulatif keragaman total yang mampu dijelaskan oleh komponen-komponen utama yang dipilih dari hasil reduksi dan transformasi minimal 80%.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Memperoleh estimasi model regresi logistik *principal component analysis* pada prediktor kategorik yang bersesuaian dengan data kadar gula darah masyarakat di Kabupaten Muna Barat.
2. Memperoleh model hubungan antara kadar gula darah masyarakat di Kabupaten Muna Barat dengan faktor resiko diabetes berdasarkan regresi logistik *principal component analysis* pada prediktor kategorik.

1.5 Manfaat Penelitian

Hasil yang diharapkan dari penelitian ini adalah untuk menambah wawasan keilmuwan mengenai pemodelan regresi logistik *principal component analysis* pada prediktor kategorik dan pengaplikasiannya pada data kadar gula darah masyarakat di Kabupaten Muna Barat.

BAB 2
TINJAUAN PUSTAKA

2.1 Multikolinieritas

Multikolinieritas merupakan suatu kondisi yang terjadi ketika terdapat korelasi diantara variabel prediktor atau dapat dikatakan antar variabel prediktor tidak bersifat saling bebas (Yan & Su, 2009). Mendeteksi adanya kasus multikolinieritas di dalam model regresi dapat dilakukan dengan cara menganalisis matriks korelasi.

Mendeteksi adanya multikolinearitas pada variabel prediktor dilakukan dengan menghitung nilai koefisien korelasi sederhana (*simple correlation*) antar variabel prediktor berdasarkan matriks korelasi $\boldsymbol{R}$ sebagai berikut:

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$$

dengan $\boldsymbol{X}$ adalah variabel prediktor yang telah distandarisasi dan n adalah jumlah data. Jika terdapat koefisien korelasi sederhana yang mencapai atau melebihi 0,8 dan mencapai atau kurang dari -0,8 maka hal tersebut menunjukkan terjadinya masalah multikolinieritas dalam regresi (Gujarati, 1978).

2.2 Vektor Eigen dan Nilai Eigen

Jika $\boldsymbol{R}$ adalah sebuah matriks berukuran $n\ x\ n$, terdapat suatu skalar λ vektor tak nol $\boldsymbol{a}$ sehingga memenuhi persamaan sebagai berikut (Anton, 1987):

$$\boldsymbol{R}\boldsymbol{a} = \lambda\boldsymbol{a}$$

$$\boldsymbol{R}\boldsymbol{a} - \lambda\boldsymbol{a} = 0$$

Skalar λ disebut nilai eigen dari $\boldsymbol{R}$ dan $\boldsymbol{a}$ disebut sebagai vektor eigen dari $\boldsymbol{R}$ yang bersesuaian dengan λ. Untuk memperoleh nilai eigen matriks $\boldsymbol{R}$ yang berukuran $n\ x\ n$, maka $\boldsymbol{R}\boldsymbol{a} = \lambda\boldsymbol{a}$ dapat ditulis sebagai $\boldsymbol{R}\boldsymbol{a} = \lambda\mathbf{I}\boldsymbol{a}$ atau $(\boldsymbol{R} - \lambda\mathbf{I})\boldsymbol{a} = 0$. Agar λ menjadi nilai eigen, maka harus ada pemecahan tak nol dari persamaan $(\boldsymbol{R} - \lambda\mathbf{I})\boldsymbol{a} = 0$. Akan tetapi karena $\det(\boldsymbol{R}) \neq 0$, maka persamaan $(\boldsymbol{R} - \lambda\mathbf{I})\boldsymbol{a} = 0$ akan mempunyai persamaan tak nol jika dan hanya jika (Anton, 1987):

$$det(\boldsymbol{R} - \lambda\mathbf{I}) = 0 \text{ atau } |\boldsymbol{R} - \lambda\mathbf{I}| = 0$$

2.3 *Principal Component Analysis*

Principal component analysis adalah suatu teknik statistik yang secara linear mengubah bentuk sekumpulan variabel asli menjadi kumpulan variabel yang lebih

kecil dan tidak saling berkorelasi yang dapat mewakili informasi dari kumpulan variabel asli (Dunteman, 1989). Selanjutnya, menurut Johnson dan Wichern (2007), *principal component analysis* adalah teknik analisis statistik untuk mentransformasi peubah–peubah asli yang masih saling berkorelasi satu dengan yang lain menjadi satu set peubah baru yang tidak berkorelasi lagi. Peubah–peubah baru tersebut disebut sebagai Komponen Utama (*Principal Component*).

Principal component analysis menjelaskan bagian dari variasi dalam kumpulan variabel yang diamati atas dasar beberapa dimensi. Tujuan khusus *principal component analysis* yaitu untuk menghilangkan multikolinieritas antar variabel prediktor dan mereduksi sejumlah besar variabel menjadi sejumlah kecil faktor. Reduksi data pengamatan menggunakan *principal component analysis* dapat dilakukan tanpa mengurangi informasi dari semua data. Oleh karena itu, p*rincipal component analysis* dipandang sebagai transformasi dari $\boldsymbol{X_1, X_2, \dots, X_p}$ (Soemartini, 2008).

Misal $\boldsymbol{R}$ merupakan matriks korelasi dari variabel–variabel $\boldsymbol{X} = [\boldsymbol{X_1, X_2, \dots, X_p}]$ dengan pasangan nilai eigen dan vektor eigen yaitu $(\lambda_1, \boldsymbol{a_1})$, $(\lambda_2, \boldsymbol{a_2})$, …,$(\lambda_p, \boldsymbol{a_p})$ dimana $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Komponen utama yang dibentuk sebagai kombinasi linier dapat didefinisikan pada persamaan berikut (Johnson dan Wichern, 2007):

$$\boldsymbol{Z_1} = \boldsymbol{a_1'X} = a_{11}\boldsymbol{X_1} + a_{12}\boldsymbol{X_2} + \cdots + a_{1p}\boldsymbol{X_p}$$
$$\boldsymbol{Z_2} = \boldsymbol{a_2'X} = a_{21}\boldsymbol{X_1} + a_{22}\boldsymbol{X_2} + \cdots + a_{2p}\boldsymbol{X_p}$$
$$\vdots$$
$$\boldsymbol{Z_r} = \boldsymbol{a_r'X} = a_{r1}\boldsymbol{X_1} + a_{r2}\boldsymbol{X_2} + \cdots + a_{rp}\boldsymbol{X_p}$$

atau

$$\boldsymbol{a_r'X} = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{p1} \\ a_{12} & a_{22} & \dots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \dots & a_{rp} \end{bmatrix} \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_p} \end{bmatrix}$$

dengan r adalah jumlah komponen utama yang terbentuk.

Kriteria pemilihan komponen utama yaitu (Johnson dan Wichern, 2007) :

1. Proporsi kumulatif keragaman total yang mampu dijelaskan oleh komponen-komponen utama yang dipilih minimal 80% .

2. Komponen utama yang dipilih adalah komponen utama yang mempunyai nilai eigen lebih besar satu atau $\lambda_p > 1$.

2.4 *Categorical Principal Component Analysis*

Categorical Principal Component Analysis (CATPCA) merupakan salah satu metode yang dilakukan untuk mengatasi multikolinieritas pada data berskala kategorik dengan menggunakan penskalaan optimal yang mengubah label kategorik ke nilai-nilai numerik dengan memaksimalkan keragaman antar peubah (Linting & Kooij 2012).

Terdapat n individu dengan p peubah diberikan dengan $n\ x\ p$ pengamatan skor matriks $\boldsymbol{X}$ dimana masing-masing peubah didefinisikan oleh $\boldsymbol{X_j}$ dengan $j = 1,2, \dots, p$. Jika peubah $\boldsymbol{X_j}$ merupakan skala pengukuran nominal atau ordinal, maka transformasi linier skala optimal diamati pada masing-masing skor dengan mengubahnya menjadi kuantifikasi kategori. Misalkan $\boldsymbol{C_j}$ adalah matriks berukuran $v_j\ x\ 1$ dengan v_j adalah jumlah kategori untuk setiap variabel $\boldsymbol{X_j}$ dan nilai $\boldsymbol{C_j}$ adalah bilangan bulat berurutan. Untuk menemukan solusi masalah $\boldsymbol{X_j}$ dan nilai $\boldsymbol{C_j}$ dapat dirumuskan dengan meminimalkan fungsi kerugian sebagai berikut (Leeuw & Mair 2009):

$$\sigma(\overline{\boldsymbol{X}}; \boldsymbol{C_j}) = \frac{1}{n}\sum_{j=1}^{p}\frac{1}{d}\left\{tr\left(\overline{\boldsymbol{X}} - \boldsymbol{G_j C_j}\right)^T\left(\overline{\boldsymbol{X}} - \boldsymbol{G_j C_j}\right)\right\} \qquad (2.1)$$

dengan $d = j = 1,2, \dots p$ merupakan jumlah dimensi yang ditentukan dan $\overline{\boldsymbol{X}}$ adalah rata-rata $\boldsymbol{X_j}$.

Sebagai variabel numerik kontinu, variabel hasil kuantifikasi juga memiliki varians seperti pada umumnya variabel kontinu. Varians CATPCA dihitung dari memaksimalkan varians variabel kuantitatif hasil kuantifikasi (Linting & Kooij, 2012). Kuantifikasi optimal mengubah variabel kategorik menjadi variabel numerik karena varians hanya dimiliki variabel numerik. Pada CATPCA, korelasi dihitung di antara variabel hasil kuantifikasi. Kuantifikasi optimal bertujuan mengoptimalkan matriks korelasi dari variabel terkuantifikasi dan untuk memaksimalkan varians pada variabel terkuantifikasi (Linting dkk, 2007).

Minimalisasi fungsi kerugian diberikan oleh algoritma *Alternating Least Square* (ALS) yaitu algoritma komputasi untuk meminimalkan fungsi kerugian kuadrat terkecil. Algoritma ALS menemukan perkiraan kuadrat terkecil dari setiap

parameter dengan memperbarui setiap matriks parameter secara bergantian (Kuroda dkk, 2013). Untuk meminimalkan fungsi kerugian $\sigma(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3})$ parameter matriks $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$, dan $\boldsymbol{\theta_3}$, dengan $\boldsymbol{\theta^{(t)}}$ yaitu t estimasi dari $\boldsymbol{\theta}$ maka algoritma ALS memperbarui perkiraan $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$, dan $\boldsymbol{\theta_3}$ dengan memecahkan masalah kuadrat terkecil untuk setiap parameter (Kuroda dkk, 2013):

$$\begin{aligned} \boldsymbol{\theta_1^{(t+1)}} &= arg \min_{\boldsymbol{\theta_1}} \sigma\left(\boldsymbol{\theta_1}, \boldsymbol{\theta_2^{(t)}}, \boldsymbol{\theta_3^{(t)}}\right), \\ \boldsymbol{\theta_2^{(t+1)}} &= arg \min_{\boldsymbol{\theta_2}} \sigma\left(\boldsymbol{\theta_1^{(t+1)}}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3^{(t)}}\right), \\ \boldsymbol{\theta_3^{(t+1)}} &= arg \min_{\boldsymbol{\theta_3}} \sigma\left(\boldsymbol{\theta_1^{(t+1)}}, \boldsymbol{\theta_2^{(t+1)}}, \boldsymbol{\theta_3}\right). \end{aligned} \tag{2.2}$$

2.5 Regresi Logistik Biner

Regresi logistik biner merupakan suatu metode statistika yang digunakan untuk menggambarkan hubungan antara variabel respon (Y) yang bersifat biner dengan variabel prediktor (X) yang bersifat kualitatif, kuantitatif ataupun kombinasi keduanya. Variabel respon Y terdiri dari 2 kategori yaitu "sukses" dan "gagal" yang dinotasikan dengan Y = 1 (sukses) dan Y = 0 (gagal). Dalam keadaan demikian, variabel Y mengikuti distribusi Bernoulli untuk setiap observasi dengan distribusi peluang sebagai berikut (Hosmer & Lemeshow, 2000):

$$f(Y_i) = \pi(X_i)^{Y_i}[1 - \pi(X_i)]^{1-Y_i}, Y_i = 0,1$$

Model dari regresi logistik biner adalah sebagai berikut (Hosmer & Lemeshow, 2000):

$$\pi(X) = \frac{e^{(\beta_0 + \beta_1 \boldsymbol{X_1} + \cdots + \beta_p \boldsymbol{X_p})}}{1 + e^{(\beta_0 + \beta_1 \boldsymbol{X_1} + \cdots + \beta_p \boldsymbol{X_p})}}$$

dengan $\pi(X)$ adalah probabilitas sukses dan $\beta_0, \beta_1, \ldots, \beta_p$ adalah parameter regresi. Pada regresi logistik, $\pi(X)$ adalah fungsi yang nonlinier sehingga untuk mempermudah dalam pendugaan parameter, $\pi(X)$ ditransformasi dengan menggunakan transformasi logit sebagai berikut (Hosmer & Lemeshow, 2000):

$$\text{Logit } \pi(X) = g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Metode estimasi parameter regresi logistik dilakukan dengan metode *maximum likelihood estimation* (Hosmer & Lemeshow, 2000). Metode tersebut mengestimasi parameter $\boldsymbol{\beta}$ untuk memaksimumkan fungsi *likelihood*. Fungsi *likelihoodnya* dapat dilihat pada persamaan sebagai berikut:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \pi(X_i)^{Y_i}[1-\pi(X_i)]^{1-Y_i}$$

untuk mempermudah perhitungan, fungsi *likelihood* dimaksimumkan dalam bentuk $\ln L(\boldsymbol{\beta})$ sebagai berikut (Agresti, 2002):

$$\begin{aligned}
\ln L(\boldsymbol{\beta}) &= \ln(\textstyle\prod_{i=1}^{n} \pi(X_i)^{Y_i}[1-\pi(X_i)]^{1-Y_i}) \\
&= \textstyle\sum_{i=1}^{n} [Y_i \ln \pi(X_i) + (1-Y_i)\ln(1-\pi(X_i))] \\
&= \textstyle\sum_{i=1}^{n} \left[Y_i \ln\left(\frac{e^{g(X)}}{1+e^{g(X)}}\right) + (1-Y_i)\ln\left(1-\frac{e^{g(X)}}{1+e^{g(X)}}\right)\right] \\
&= \textstyle\sum_{i=1}^{n} \left[Y_i g(X) - \ln\left(1+e^{g(X)}\right)\right] \\
&= \textstyle\sum_{i=1}^{n} \left[Y_i(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}) - \ln\left(1+e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}\right)\right] \\
&= \textstyle\sum_{i=1}^{n} \left[Y_i(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}) - \ln\left(1+e^{\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}}\right)\right]
\end{aligned}$$

Turunan fungsi ln *likelihood* terhadap parameter–parameternya:

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n}\left[Y_i - \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}}}\right]$$

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}\left[Y_i X_{ij} - X_{ij}\frac{\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)}{1+\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)}\right]$$

Nilai turunan pertama dari fungsi ln *likelihood* tidak memberikan penyelesaian, sehingga digunakan iterasi *Newton-Raphson* untuk mendapatkan nilai taksirannya. Dalam metode *Newton-Raphson* dibutuhkan turunan pertama dan kedua dari fungsi ln *likelihoodnya* dengan bentuk persamaan sebagai berikut:

$$\begin{bmatrix}\hat{\beta}_{0(t+1)}\\ \hat{\beta}_{1(t+1)}\\ \hat{\beta}_{2(t+1)}\\ \vdots \\ \hat{\beta}_{p(t+1)}\end{bmatrix} = \begin{bmatrix}\hat{\beta}_{0(t)}\\ \hat{\beta}_{1(t)}\\ \hat{\beta}_{2(t)}\\ \vdots \\ \hat{\beta}_{p(t)}\end{bmatrix} - \boldsymbol{H^{-1}d} \qquad (2.3)$$

dengan :

$\boldsymbol{d} = \begin{bmatrix}\frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_0}\\ \frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_1}\\ \vdots \\ \frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_p}\end{bmatrix}$ adalah matriks turunan pertama fungsi ln *likelihood* terhadap parameternya.

H adalah matriks turunan kedua fungsi ln *likelihood* terhadap parameternya.

Turunan kedua fungsi ln *likelihood* terhadap parameter–parameternya adalah sebagai berikut:

$$\frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial^2 \beta_0} = -\sum_{i=1}^{n}\left[\frac{\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)}{\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)\right)^2}\right]$$

$$\frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_0 \partial \beta_j} = -\sum_{i=1}^{n}\left[X_{il}\frac{\exp(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij})}{\left(1+\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}\right)\right)^2}\right]$$

$$\frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_j \partial \beta_m} = -\sum_{i=1}^{n}\left[X_{il}X_{im}\frac{\exp(\beta_0+\sum_{i=1}^{n}\beta_i X_{ij})}{\left(1+\exp\left(\beta_0+\sum_{i=1}^{n}\beta_i X_{ij}\right)\right)^2}\right]$$

dengan $j, m = 1,2,\ldots,p$.

Persamaan (2.3) secara umum dapat dibentuk parameter taksiran dengan iterasi *Newton-Raphson* sebagai berikut:

$$\begin{bmatrix}\hat{\beta}_{0(t+1)}\\ \hat{\beta}_{1(t+1)}\\ \hat{\beta}_{2(t+1)}\\ \vdots\\ \hat{\beta}_{p(t+1)}\end{bmatrix} = \begin{bmatrix}\hat{\beta}_{0(t)}\\ \hat{\beta}_{1(t)}\\ \hat{\beta}_{2(t)}\\ \vdots\\ \hat{\beta}_{p(t)}\end{bmatrix} - \begin{bmatrix}\frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial^2 \beta_0} & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_0 \partial \beta_p}\\ \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial^2 \beta_1} & \cdots & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_1 \partial \beta_p}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial^2 \beta_p}\end{bmatrix}^{-1}$$

$$\begin{bmatrix}\frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_0}\\ \frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_1}\\ \frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_2}\\ \vdots\\ \frac{\partial \ln L\left((\beta_0,\beta_1,\ldots,\beta_p);Y_i\right)}{\partial \beta_p}\end{bmatrix} \quad (2.4)$$

2.6 *Principal Component Logistic Regression*

Principal component logistic regression merupakan teknik multivariat yang dapat dipakai untuk mereduksi dimensi dari variabel–variabel prediktor. *Principal component logistic regression* bertujuan meningkatkan estimasi parameter model logistik yang memiliki multikolinearitas dengan menggunakan komponen utama dari variabel prediktor. Secara umum bentuk persamaan dari model *principal component logistic regression* yaitu (Aguilera dkk, 2005):

$$\pi(Z) = \frac{\exp\{\beta_0+\sum_{k=1}^{r}\sum_{j=1}^{p}Z_k a_{jk}\beta_j\}}{1+\exp\{\beta_0+\sum_{k=1}^{r}\sum_{j=1}^{p}Z_k a_{jk}\beta_j\}} = \frac{\exp\{\beta_0+\sum_{k=1}^{r}Z_k\gamma_k\}}{1+\exp\{\beta_0+\sum_{k=1}^{r}Z_k\gamma_k\}}$$

dengan $\pi(Z)$ adalah probabilitas sukses dan $\beta_0, \gamma_1, \dots, \gamma_r$ adalah parameter regresi. Model *principal component logistic regression* tersebut dapat diformulasikan dalam bentuk matriks dengan transformasi fungsi logit $g = \ln(\frac{\pi(Z)}{1-\pi(Z)})$ sebagai berikut (Aguilera dkk, 2005):

$$g = \beta_0 + \sum_{k=1}^{r} Z_k \gamma_k + \varepsilon \tag{2.5}$$

dengan :

g = Probabilitas kejadian sukses pada $Y = 1$

β_0 = Konstanta

γ_k = Koefisien regresi logistik berdasarkan komponen utama yang terbentuk

Z_k = Komponen utama yang terbentuk

k = Banyaknya komponen utama yang terbentuk dari 1 hingga r

ε = Eror

Pendugaan parameter $(\beta_0, \gamma_1, \dots, \gamma_r)$ dapat diperoleh dengan metode *maksimum likelihood estimation* dimana metode ini ditaksir dengan memaksimumkan fungsi *likelihood*. Fungsi *likelihoodnya* dapat dilihat pada persamaan sebagai berikut (Aguilera, 2005):

$$\begin{aligned} f\big((\beta_0, \gamma_1, \gamma_2, \dots, \gamma_r); Y_i\big) &= \prod_{i=1}^{n} f(Y_i) \\ &= \prod_{i=1}^{n} (\pi(Z_i))^{Y_i} (1 - \pi(Z_i))^{1-Y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\pi(Z_i)}{1-\pi(Z_i)}\right)^{Y_i} (1 - \pi(Z_i)) \end{aligned}$$

dengan demikian fungsi ln *likelihoodnya* adalah:

$$\begin{aligned} ln\, L\big((\beta_0, \gamma_1, \gamma_2, \dots, \gamma_r); Y_i\big) &= ln \prod_{i=1}^{n} \left(\frac{\pi(Z_i)}{1-\pi(Z_i)}\right)^{Y_i} (1 - \pi(Z_i)) \\ &= \sum_{i=1}^{n} \Big[Y_i (\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k) - ln\big(1 + exp(\beta_0 + \\ &\quad \sum_{k=1}^{r} Z_k \gamma_k)\big) \Big] \end{aligned} \tag{2.6}$$

Selanjutnya Persamaan (2.6) diturunkan terhadap parameter-parameternya:

$$\frac{\partial ln\big((\beta_0,\gamma_1,\gamma_2,\dots,\gamma_r);Y_i\big)}{\partial \beta_0} = \sum_{i=1}^{n} \left[Y_i - \frac{exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)}{1 + exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)} \right]$$

$$\frac{\partial ln\big((\beta_0,\gamma_1,\gamma_2,\dots,\gamma_r);Y_i\big)}{\partial \gamma_1} = \sum_{i=1}^{n} \left[Y_i Z_{i1} - Z_{i1} \frac{exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)}{1 + exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)} \right]$$

$$\frac{\partial ln\big((\beta_0,\gamma_1,\gamma_2,\dots,\gamma_r);Y_i\big)}{\partial \gamma_r} = \sum_{i=1}^{n} \left[Y_i Z_{ir} - Z_{ir} \frac{exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)}{1 + exp\big(\beta_0 + \sum_{k=1}^{r} Z_k \gamma_k\big)} \right]$$

Nilai turunan pertama dari fungsi ln *likelihood* tidak memberikan penyelesaian, sehingga digunakan iterasi *Newton-Raphson* pada penaksiran parameter–parameter ini untuk mendapatkan nilai taksirannya. Dalam metode *Newton-Raphson* dibutuhkan turunan pertama dan kedua dari fungsi ln *likelihoodnya* dengan bentuk persamaan sebagai berikut:

$$\begin{bmatrix} \hat{\beta}_{0(t+1)} \\ \hat{\gamma}_{1(t+1)} \\ \hat{\gamma}_{2(t+1)} \\ \vdots \\ \hat{\gamma}_{r(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{0(t)} \\ \hat{\gamma}_{1(t)} \\ \hat{\gamma}_{2(t)} \\ \vdots \\ \hat{\gamma}_{r(t)} \end{bmatrix} - \boldsymbol{H^{-1}d} \tag{2.7}$$

dengan :

$(\hat{\beta}_{0(t)}, \hat{\gamma}_{1(t)}, \hat{\gamma}_{2(t)}, \dots, \hat{\gamma}_{r(t)})^T$ adalah parameter regresi.

$\boldsymbol{d} = \begin{bmatrix} \frac{\partial \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \beta_0} \\ \frac{\partial \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \gamma_1} \\ \frac{\partial \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \gamma_2} \\ \vdots \\ \frac{\partial \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \gamma_r} \end{bmatrix}$ adalah matriks turunan pertama fungsi *ln likelihood* terhadap parameternya.

$\boldsymbol{H}$ adalah matriks turunan kedua fungsi *ln likelihood* terhadap parameternya.

Turunan kedua fungsi *log likelihood* terhadap parameter–parameternya adalah sebagai berikut:

$$\frac{\partial^2 \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial^2 \gamma_0} = -\sum_{i=1}^{n}\left[\frac{exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)}{1+exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)}\right]$$

$$\frac{\partial^2 \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \beta_0 \partial \gamma_l} = -\sum_{i=1}^{n}\left[Z_{il}\frac{exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)}{\left(1+exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)\right)^2}\right]$$

$$\frac{\partial^2 \ln L((\beta_0,\gamma_1,\dots,\gamma_r);Y_i)}{\partial \gamma_l \partial \gamma_k} = -\sum_{i=1}^{n}\left[Z_{il}Z_{ik}\frac{exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)}{\left(1+exp(\beta_0+\sum_{k=1}^{r} Z_k\gamma_k)\right)^2}\right]$$

dengan $l, k = 1,2, \dots, r$.

Persamaan (2.7) secara umum dapat dibentuk parameter taksiran dengan iterasi *Newton-Raphson* sebagai berikut:

$$\begin{bmatrix}\hat{\beta}_{0(t+1)}\\ \hat{\gamma}_{1(t+1)}\\ \gamma_{2(t+1)}\\ \vdots\\ \hat{\gamma}_{r(t+1)}\end{bmatrix}=\begin{bmatrix}\hat{\beta}_{0(t)}\\ \hat{\gamma}_{1(t)}\\ \hat{\gamma}_{2(t)}\\ \vdots\\ \hat{\gamma}_{r(t)}\end{bmatrix}-\begin{bmatrix}\frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial^2\beta_0} & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\beta_0\partial\gamma_1} & \cdots & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\beta_0\partial\gamma_r}\\ \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_1\partial\beta_0} & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial^2\gamma_1} & \cdots & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_1\partial\gamma_r}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_r\partial\beta_0} & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_r\partial\gamma_1} & \cdots & \frac{\partial^2 \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial^2\gamma_r}\end{bmatrix}^{-1}$$

$$\begin{bmatrix}\frac{\partial \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\beta_0}\\ \frac{\partial \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\beta_1}\\ \frac{\partial \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_2}\\ \vdots\\ \frac{\partial \ln L((\beta_0,\gamma_1,\ldots,\gamma_r);Y_i)}{\partial\gamma_r}\end{bmatrix} \tag{2.8}$$

2.7 Pengujian Parameter

Model yang telah diperoleh perlu diuji signifikansi pada koefisien $\beta_0, \gamma_1, \ldots, \gamma_r$ terhadap variabel respon, yaitu dengan uji serentak. Pengujian ini dilakukan untuk memeriksa kemaknaan koefisien $\beta_0, \gamma_1, \ldots, \gamma_r$ terhadap variabel respon secara bersama-sama dengan menggunakan statistik uji.

Hipotesis:

H_0 : $\gamma_1 = \gamma_2 = \cdots = \gamma_r = 0$ (semua komponen utama dalam model regresi logistik tidak mempengaruhi variabel respon).

H_1 : paling sedikit ada satu $\gamma_k \neq 0; k = 1, 2, \ldots, r$ (paling sedikit ada satu komponen utama dalam model regresi logistik yang berpengaruh terhadap variabel respon).

Statistik uji yang digunakan adalah statistik uji G atau *likelihood ratio test*.

$$G^2 = -2 \ln \frac{l_0}{l_1} \tag{2.9}$$

dengan:

l_0 = nilai *likelihood* untuk model yang tidak mengandung variabel prediktor

l_1 = nilai *likelihood* untuk model yang mengandung variabel prediktor

Statistik uji G mengikuti distribusi chi kuadrat (χ^2) dengan daerah penolakan H_0 adalah jika $G > \chi^2_{(\alpha:r)}$ atau *p-value* $< \alpha, \alpha = 0{,}05$ dan $r =$ jumlah komponen utama (Hosmer & Lemeshow, 2000).

2.8 Ketepatan Klasifikasi

Salah satu ukuran untuk pemilihan model terbaik yang dapat digunakan pada pemodelan statistik yang melibatkan variabel respon kualitatif adalah ketepatan

klasifikasi (Ratnasari, 2012). Ketepatan klasifikasi dapat digunakan dalam suatu evaluasi model. Menurut Johnson & Wichern (2007), evaluasi ketepatan klasifikasi adalah suatu evaluasi yang melihat probabilitas kesalahan klasifikasi yang dilakukan oleh suatu fungsi klasifikasi. Nilai ketepatan klasifikasi tersebut dapat diperoleh dengan membandingkan nilai prediksi yang benar dari model dengan nilai observasi yang sebenarnya. Adapun tabel ketepatan klasifikasi yang biasa digunakan pada model regresi dengan variabel respon yang bersifat kategori disajikan pada Tabel 1 sebagai berikut:

Tabel 1. Ketepatan Klasifikasi

Hasil Observasi	Prediksi	
	Y_1	Y_2
Y_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
Y_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

Kemudian dirumuskan dalam persamaan berikut :

$$\text{Akurasi} = \frac{n_{1M}+n_{2M}}{n_1+n_2} x100\% \qquad (2.10)$$

Y_i = Variabel respon, $i = 1,2, \ldots$

n_{1C} = Nilai dari objek Y_1 yang benar diklasifikasikan sebagai objek Y_1

n_{1M} = Nilai dari objek Y_1 yang salah diklasifikasikan sebagai objek Y_2

n_{2C} = Nilai dari objek Y_2 yang benar diklasifikasikan sebagai objek Y_2

n_{2M} = Nilai dari objek Y_2 yang salah diklasifikasikan sebagai objek Y_1

2.9 Penyakit Diabetes

Diabetes atau kencing manis adalah suatu gangguan kesehatan berupa kumpulan gejala yang timbul pada seseorang yang disebabkan oleh peningkatan kadar gula dalam darah akibat kekurangan insulin ataupun resistensi insulin dan gangguan metabolik pada umumnya. Penyakit diabetes dapat menyerang semua lapisan umur dan sosial ekonomi (Alamsyah, dkk 2017).

Peningkatan prevalensi diabetes di dunia lebih menonjol perkembangannya di Negara berkembang dibandingkan dengan negara maju. WHO memprediksi akan ada kenaikan prevalensi diabetes di Indonesia dari 8,4 juta diabetisi pada tahun 2000, 14 juta diabetisi pada tahun 2006, dan akan meningkat menjadi sekitar 21,3 juta diabetisi pada tahun 2030. Faktor resiko diabetes melitus bisa dikelompokkan

menjadi faktor risiko yang tidak dapat dimodifikasi dan yang dapat dimodifikasi. Faktor risiko yang tidak dapat dimodifikasi adalah ras dan etnik, umur, jenis kelamin, riwayat keluarga dengan diabetes melitus, riwayat melahirkan bayi dengan berat badan lebih dari 4000 gram, dan riwayat lahir dengan berat badan lahir rendah (kurang dari 2500 gram). Sedangkan faktor risiko yang dapat dimodifikasi erat kaitannya dengan perilaku hidup yang kurang sehat, yaitu berat badan lebih, obesitas abdominal/sentral, kurangnya aktivitas fisik, hipertensi, dan diet tidak sehat/tidak seimbang (Riskesdas, 2018).