TESIS – SS14 2501

PERBANDINGAN METODE RARE EVENT WEIGHTED LOGISTIC REGRESSION DAN TRUNCATED REGULARIZED PRIOR CORRECTION UNTUK KLASIFIKASI PADA IMBALANCED DATA

(Studi Kasus: Klasifikasi Tingkat Kesejahteraan di Provinsi Bali)

SONY PUJI TRIASMORO

NRP. 06211650017002

DOSEN PEMBIMBING :

Dr. Vita Ratnasari, S.Si., M.Si.

Dr. Dra. Agnes Tuti Rumiati, M.Sc.

PROGRAM PASCASARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

TESIS – SS14 2501

PERBANDINGAN METODE RARE EVENT WEIGHTED LOGISTIC REGRESSION DAN TRUNCATED REGULARIZED PRIOR CORRECTION UNTUK KLASIFIKASI PADA IMBALANCED DATA

(Studi Kasus: Klasifikasi Tingkat Kesejahteraan di Provinsi Bali)

SONY PUJI TRIASMORO

NRP. 06211650017002

DOSEN PEMBIMBING :

Dr. Vita Ratnasari, S.Si., M.Si.

Dr. Dra. Agnes Tuti Rumiati, M.Sc.

PROGRAM PASCASARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

THESIS – SS14 2501

COMPARISON PERFORMANCE OF RARE EVENT WEIGHTED LOGISTIC REGRESSION AND TRUNCATED REGULARIZED PRIOR CORRECTION METHODS FOR CLASSIFICATION ON IMBALANCED DATA

(Case Study: Welfare Classification in Bali)

SONY PUJI TRIASMORO

NRP. 06211650017002

SUPERVISORS :

Dr. Vita Ratnasari, S.Si., M.Si.

Dr. Dra. Agnes Tuti Rumiati, M.Sc.

PROGRAM OF MAGISTER
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, DAN DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

PERBANDINGAN METODE RARE EVENT WEIGHTED LOGISTIC REGRESSION DAN TRUNCATED REGULARIZED PRIOR CORRECTION UNTUK KLASIFIKASI PADA IMBALANCED DATA
(Studi Kasus: Klasifikasi Tingkat Kesejahteraan di Provinsi Bali)

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Sains (M.Si)
di
Institut Teknologi Sepuluh Nopember
Oleh:

SONY PUJI TRIASMORO
NRP. 06211650017002

Tanggal Ujian : 23 Januari 2018
Periode Wisuda : Maret 2018

Disetujui Oleh:

1. Dr. Vita Ratnasari, S.Si., M.Si. (Pembimbing I)
   NIP. 19700910 199702 2 001
2. Dr. Dra. Agnes Tuti Rumiati, M.Sc. (Pembimbing II)
   NIP. 19570724 198503 2 002
3. Prof. Dr. Drs. I Nyoman Budiantara, M.Si. (Penguji)
   NIP. 19650603 198903 1 003
4. Dr. Siti Muchlisoh, M.Si. (Penguji)
   NIP. 19700219 199211 2 001

Dekan
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Prof. Dr. Basuki Widodo, M.Sc.
NIP. 19650605 198903 1 002

PERBANDINGAN METODE *RARE EVENT WEIGHTED LOGISTIC REGRESSION* DAN *TRUNCATED REGULARIZED PRIOR CORRECTION* UNTUK KLASIFIKASI PADA *IMBALANCED DATA*
(Studi Kasus: Klasifikasi Tingkat Kesejahteraan di Provinsi Bali)

Nama Mahasiswa : Sony Puji Triasmoro
NRP : 06211650017002
Dosen Pembimbing : Dr. Vita Ratnasari, S.Si., M.Si.
Dr. Dra. Agnes Tuti Rumiati, M.Sc.

ABSTRAK

Kemiskinan menjadi masalah yang tetap harus diselesaikan dan membutuhkan program-program penanggulangan kemiskinan yang dilakukan secara tepat. Kesalahan klasifikasi dalam pelaksanaan program penanggulangan kemiskinan seperti *inclusion error* (rumah tangga yang tidak berhak menerima manfaat, tetapi menerima manfaat) dan *exclusion error* (rumah tangga yang berhak menerima manfaat, tetapi tidak masuk sebagai penerima manfaat) harus diminimalisir sekecil mungkin. Untuk itu dibutuhkan metodologi dalam menentukan klasifikasi tingkat kesejahteraan rumah tangga yang tepat. Regresi logistik adalah metode analisis yang dapat digunakan untuk melakukan klasifikasi data. Akan tetapi ketika digunakan dalam kasus tingkat kesejahteraan di Indonesia yang datanya bersifat *imbalanced*, maka metode regresi logistik biasa tidak tepat digunakan. Hal ini disebabkan regresi logistik cenderung menihilkan peluang dari kelompok minoritas karena nilai prediksi akan cenderung kepada kategori data yang mayoritas sehingga tingkat ketepatan yang dihasilkan menjadi kurang baik.

Rare Events Weighted Logistic Regression (RE-WLR) dan *Truncated Regularized Prior Correction* (TR-PC) merupakan pengembangan dari regresi logistik yang digunakan untuk mengatasi kelemahan pada kasus *imbalanced data*. Penelitian ini menerapkan metode RE-WLR dan TR-PC dengan menggunakan 14 variabel penelitian dan hasilnya adalah didapatkan 10 variabel yang signifikan dalam model RE-WLR dan TR-PC dan 4 variabel lainnya tidak signifikan dalam kedua model. Indikator yang digunakan untuk mengukur akurasi performa model dalam memprediksi kelas positif (rumah tangga miskin) dan kelas negatif (rumah tangga tidak miskin) dalam penelitian ini adalah nilai AUC (*Area Under Curve*) dan G-mean. Hasil penelitian ini menunjukkan model RE-WLR mampu memberikan hasil akurasi yang lebih baik dibandingkan model TR-PC dan regresi logistik dengan nilai G-mean sebesar 83,47 persen dan nilai AUC adalah 0,836 (kategori baik). Nilai ini di atas nilai G-mean dari model TR-PC 73,68 persen dan nilai AUC dari model TR-PC yaitu 0,762 (kategori cukup).

Kata kunci : Klasifikasi, Kemiskinan, Regresi Logistik, *Imbalanced data*

halaman ini sengaja dikosongkan

COMPARISON PERFORMANCE OF RARE EVENT WEIGHTED LOGISTIC REGRESSION AND TRUNCATED REGULARIZED PRIOR CORRECTION METHODS FOR CLASSIFICATION ON IMBALANCED DATA (Case Study: Welfare Classification in Bali)

Name of Student	: Sony Puji Triasmoro
NRP	: 06211650017002
Supervisors	: Dr. Vita Ratnasari, S.Si., M.Si.
	Dr. Dra. Agnes Tuti Rumiati, M.Sc.

ABSTRACT

Poverty is a problem that remains to be resolved and requires appropriate poverty reduction programs. Error classification in the implementation of poverty reduction programs such as inclusion error (households that are not entitled to receive benefits, but receive benefits) and exclusion error (households who are eligible to receive benefits, but not included as beneficiaries) should be minimized as little as possible. Therefore, methodology is needed in determining the proper classification of household welfare. Logistic regression is an analytical method that can be used to classify data. However, when used in cases of welfare in Indonesia where the data are imbalanced, the usual logistic regression method is not appropriately used. This is due to logistic regression tends to nullify the opportunities of minority groups because predicted values will tend to the category of data that the majority so that the resulting accuracy becomes less good.

Rare Events Weighted Logistic Regression (RE-WLR) and Truncated Regularized Prior Correction (TR-PC) are the development of logistic regression used to overcome weaknesses in the case of imbalanced data. This research applies RE-WLR and TR-PC method using 14 research variables and the result is found 10 significant variables in RE-WLR and TR-PC model and 4 other variables are not significant in both models. Indicators used to measure the accuracy of model performance in predicting positive (poor household) and negative class (non-poor households in this study were AUC and G-mean values.This study showed that RE-WLR model was able to give more accuracy result both compared to TR-PC model and logistic regression with G-mean value of 83.47 percent and AUC value is 0.836 (good category) .This value is above the G-mean value of TR-PC model is 73.68 percent and the AUC value of model TR-PC is 0.762 (enough category).

Kata kunci : Classification, Poverty, Logistic Regression, Imbalanced data

halaman ini sengaja dikosongkan

KATA PENGANTAR

Alhamdulillah, segala puji bagi ALLAH SWT yang telah melimpahkan karunia nikmat dan rahmat kepada penulis sehingga tesis yang berjudul "Perbandingan Metode Rare Event Weighted Logistic Regression dan Truncated Regularized Prior Correction untuk Klasifikasi pada Imbalanced Data (Studi Kasus: Klasifikasi Tingkat Kesejahteraan di Provinsi Bali)" ini dapat diselesaikan sesuai dengan waktu yang diharapkan. Tesis ini disusun dalam rangka memenuhi salah satu syarat untuk meraih gelar Magister Sains di Program Pasca Sarjana Institut Teknologi Sepuluh Nopember (ITS) Surabaya.

Pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih dan penghargaan yang setinggi-tingginya kepada:

1. Ibu Dr. Vita Ratnasari, S.Si., M.Si. dan Ibu Dr. Dra. Agnes Tuti Rumiati, M.Sc. selaku pembimbing atas segala bimbingan, arahan dan motivasi dalam penyusunan tesis ini.
2. Bapak Prof. Dr. Drs. I Nyoman Budiantara, M.Si. dan Dr. Siti Muchlisoh, M.Si. selaku dosen penguji yang telah memberikan saran dan koreksi atas penulisan tesis ini.
3. Bapak. Dr. Suhartono, M.Sc. selaku Kepala Departemen Statistika ITS dan Bapak Dr.rer.pol. Heri Kuswanto, M.Si. selaku Ketua Program Studi Pascasarjana atas segala arahan dan bantuannya selama penulis menempuh pendidikan.
4. Bapak dan Ibu dosen Statistika ITS yang telah mencurahkan ilmu dan pengalaman yang bermanfaat kepada penulis serta segenap karyawan keluarga besar Departemen Statistika ITS Surabaya, atas segala dukungan dan bantuan yang diberikan selama penulis menjadi bagian dari sistem.
5. Kepala Badan Pusat Statistik (BPS) RI beserta jajarannya, Kepala Pusdiklat BPS beserta jajarannya, Kepala BPS Provinsi Bali beserta jajarannya, dan Kepala BPS Badung beserta jajarannya, yang telah memberi kesempatan, dukungan, kepercayaan dan beasiswa kepada penulis untuk melajutkan studi program S2 di ITS.

6. Istri tercinta, Umi Halimatus Sakdiyah, atas doa, dukungan, pengertian, kesabaran dan pengorbanan yang luar biasa. Kakak Raqila Awahita Farzana dan Adik Ayyash Shaka Dirandra Hamizan, yang turut memberikan semangat selama masa studi.
7. Ibu Sulasih dan Bapak Eko Bayu Kusnanto tercinta, Ibu dan Bapak mertua, Mas Fajar Puji Rianto dan Fery Puji Nova Kurniawan, dan seluruh keluarga atas doa dan semangat yang diberikan.
8. Teman-teman seperjuangan BPS-ITS Batch 10: Mas Fendi, Mas Topik, Ratih, Mbak Ratna, Mbak Tika, Mbak Imra, Mbak Reni, Mbak Mike, Mas Prapto, Bang Fael, Mas Umam, Bang Fieldri, Mbak Prih dan Mas Aniq. Terima kasih atas kerja sama dan pengalaman bersama selama menempuh studi di kampus tercinta.
9. Teman-teman reguler Pascasarjana angkatan 2016 yang telah berjuang bersama-sama mencari ilmu dan menghadapi ujian baik dari kelas B dan A yang tidak bisa disebutkan satu per satu. Terima kasih dan sampai jumpa kembali di lain waktu.
10. Semua pihak yang telah membantu Penulis yang tidak dapat disebutkan satu per satu.

Penulis menyadari bahwa tesis ini masih jauh dari kata sempurna, sehingga saran dan kritik membangun dari semua pihak sangat penulis harapkan untuk perbaikan selanjutnya.

Surabaya, Januari 2018

Penulis

DAFTAR ISI

ABSTRAK ..... **iii**
ABSTRACT ..... **v**
KATA PENGANTAR ..... **vii**
DAFTAR ISI ..... **ix**
DAFTAR TABEL ..... **xi**
DAFTAR GRAFIK ..... **xiii**
DAFTAR LAMPIRAN ..... **xv**
BAB 1 PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Rumusan Masalah 6
1.3 Tujuan Penelitian 7
1.4 Manfaat Penelitian 7
1.5 Batasan Penelitian 8
BAB 2 TINJAUAN PUSTAKA **9**
2.1 Regresi Logistik 9
2.2 *Regularization* 11
2.3 *Iteratively Re-Weighted Least Squares* (IRLS) 12
2.4 *Rare Event Weighted Logistic Regression* (RE-WLR) 14
2.5 *Truncated-Regularized Prior Correction* 18
2.6 Evaluasi Performa Klasifikasi 21
2.7 Kemiskinan 23
2.7.1 Konsep dan Definisi 23
2.7.2 Garis Kemiskinan 24
2.7.3 Indikator Kemiskinan 26
2.7.4 Kajian tentang Karakteristik Kemiskinan 26
BAB 3 METODE PENELITIAN **29**
3.1 Tahapan Mendapatkan Estimasi Parameter RE-WLR dan TR-PC 29
3.2 Penerapan Metode RE-WLR dan TR-PC pada Kasus Klasifikasi Tingkat Kesejahteraan di Provinsi Bali 31

3.2.1 Variabel Penelitian 31
3.2.2 Sumber Data 34
3.2.3 Penyiapan Data *Training* dan Data *Testing* 35
3.2.4 Tahapan Pemodelan RE-WLR dan TR-PC untuk klasifikasi Tingkat Kesejahteraan di Provinsi Bali 36
BAB 4 HASIL DAN PEMBAHASAN 40
4.1 Kajian Estimasi Parameter RE-WLR dan TR-PC 41
4.2 Penerapan Metode RE-WLR dan TR-PC pada Klasifikasi Tingkat Kesejahteraan di Provinsi Bali 57
4.2.1 Deskripsi Karakteristik Rumah Tangga Menurut Variabel Penelitian 57
4.2.2 Klasifikasi Tingkat Kesejahteraan Rumah Tangga dengan RE-WLR 61
4.2.3 Klasifikasi Tingkat Kesejahteraan Rumah Tangga dengan TR-PC 67
4.3 Perbandingan Ketepatan Klasifikasi antara Model RE-WLR dan TR-PC 72
BAB 5 KESIMPULAN DAN SARAN 75
5.1 Kesimpulan 75
5.2 Saran 75
DAFTAR PUSTAKA 77
LAMPIRAN 81
BIOGRAFI PENULIS 99

DAFTAR TABEL

Tabel 2.1 *Confusion matrix* untuk hasil klasifikasi biner 21
Tabel 2.2 Kategori Pengklasifikasian berdasarkan Nilai AUC 23
Tabel 3.1 Variabel penelitian 31
Tabel 3.2 Struktur data penelitian 34
Tabel 3.3 Jumlah rumah tangga sampel Susenas 2015 Provinsi Bali 35
Tabel 3.4 Skema data *training* dan data *testing* 36
Tabel 4.1 Kondisi literasi, pendidikan, dan lapangan usaha KRT menurut tingkat kesejahteraan (dalam persentase) 58
Tabel 4.2 Kondisi demografi rumah tangga menurut tingkat kesejahteraan (dalam persentase) 59
Tabel 4.3 Kondisi keadaan rumah menurut tingkat kesejahteraan (dalam persentase) 59
Tabel 4.4 Kondisi sumber air minum, sanitasi dan wilayah tempat tinggal menurut tingkat kesejahteraan (dalam persentase) 60
Tabel 4.5 Kondisi aset yang dimiliki menurut tingkat kesejahteraan (dalam persentase) 61
Tabel 4.6 Rata-Rata kinerja klasifikasi RE-WLR pada data *testing* 62
Tabel 4.7 Ukuran performa Klasifikasi model RE-WLR terbaik 63
Tabel 4.8 Estimasi Parameter RE-WLR 64
Tabel 4.9 Rata-Rata kinerja klasifikasi TR-PC pada data *testing* 67
Tabel 4.10 Ukuran performa Klasifikasi model TR-PC terbaik 68
Tabel 4.11 Estimasi Parameter TR-PC 69
Tabel 4.12 Perbandingan kinerja klasifikasi antara RE-WLR, TR-PC dan Reglog 72

halaman ini sengaja dikosongkan

DAFTAR GRAFIK

Gambar 2.1 Kurva ROC 22
Gambar 2.2 Kerangka Pikir Penelitian 27
Gambar 3.1 Flowchart Langkah Estimasi Parameter RE-WLR dan TR-PC 29
Gambar 3.2 Flowchart Langkah Pemodelan RE-WLR dan TR-PC 39
Gambar 4.1 Persentase rumah tangga miskin dan tidak miskin di Provinsi Bali Tahun 2015 57

halaman ini sengaja dikosongkan

DAFTAR LAMPIRAN

Lampiran 1 Algoritma RE-WLR menggunakan IRLS 81

Lampiran 2 Algoritma TR-PC menggunakan IRLS 82

Lampiran 3 Algoritma Menghitung $\hat{\boldsymbol{\beta}}$ menggunakan CG linier.....83

Lampiran 4 Algoritma Menghitung B($\hat{\boldsymbol{\beta}}$) menggunakan CG linier 84

Lampiran 5 Data yang digunakan 85

Lampiran 6 Syntax R untuk model RE-WLR.....86

Lampiran 7 Syntax R untuk model TR-PC.....89

Lampiran 8 Contoh Output R untuk model RE-WLR 92

Lampiran 9 Contoh Output R untuk model TR-PC 94

Lampiran 10 Hasil Klasifikasi Model RE-WLR dan TR-PC Data Training 96

Lampiran 11 Rata-Rata Hasil Klasifikasi Model RE-WLR dan TR-PC Pada Data Testing.....97

halaman ini sengaja dikosongkan

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Peningkatan kesejahteraan penduduk merupakan suatu tujuan yang ingin dicapai oleh berbagai pihak di dunia dan hal ini erat kaitannya dengan melakukan program pengentasan kemiskinan. Hal ini karena peningkatan kesejahteraan suatu masyarakat, diawali dari bagaimana menghapus kemiskinan sebagai standar minimum suatu penduduk untuk memenuhi kebutuhan hidup. Penyelesaian masalah kemiskinan dituangkan pada tujuan pertama dari delapan tujuan *Millenium Development Goals* (MDG's) yaitu memberantas kemiskinan dan kelaparan ekstrem. Kelanjutan dari MDG's yakni *Sustainable Development Goals* (SDG's) pun tetap memberikan porsi pengentasan kemiskinan menjadi tujuan pertama dari tujuh belas tujuan yakni tanpa kemiskinan melalui pengentasan segala bentuk kemiskinan di semua tempat (*United Nation*, 2015).

Program pengentasan kemiskinan tentu harus memiliki tolok ukur yang jelas. Salah satu tolok ukur tersebut adalah tersedianya data kemiskinan yang dapat dipercaya sehingga dapat menjadi instrumen bagi pengambil kebijakan dalam melakukan kebijakan yang tepat sasaran. Selain itu data kemiskinan yang baik dapat digunakan untuk mengevaluasi kebijakan pemerintah terhadap kemiskinan dan membandingkan kemiskinan antar daerah dan antar waktu. Data kemiskinan di Indonesia dihasilkan oleh Badan Pusat Statistik (BPS) dengan menggunakan pendekatan *basic need approach* (pendekatan kebutuhan dasar). Dengan pendekatan ini kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran sehingga timbul istilah garis kemiskinan. Penduduk miskin adalah penduduk yang memiliki rata-rata pengeluaran per bulan di bawah garis kemiskinan (BPS, 2016).

Keadaan kemiskinan di Indonesia berdasarkan data Badan Pusat Statistik terlihat bahwa dalam satu dekade yaitu periode 2006-2016 tingkat kemiskinan di Indonesia mengalami penurunan baik itu secara persentase maupun jumlah

absolut. Pada periode ini jumlah penduduk miskin turun sebanyak 11,29 juta jiwa yaitu dari 39,30 juta jiwa pada tahun 2006 menjadi 28,01 juta jiwa pada 2016. Secara persentase juga terjadi penurunan dari 17,75 persen pada tahun 2006 menjadi 10,86 persen pada 2016 (BPS, 2016). Meskipun dari data terlihat upaya pengentasan kemiskinan terkesan berhasil, akan tetapi yang masih perlu diingat bahwa terdapat 1 dari 10 penduduk Indonesia yang masih berada di bawah garis kemiskinan sehingga upaya-upaya pengentasan kemiskinan tetap harus dilakukan.

Provinsi Bali sebagai salah satu provinsi di Indonesia juga tidak terlepas dari permasalahan kemiskinan. Berdasarkan data Badan Pusat Statistik dengan menggunakan pendekatan *basic need approach*, pada tahun 2016 tercatat 178,18 ribu penduduk Bali masih berada di bawah garis kemiskinan atau secara relatif berjumlah 4,25 persen. Angka ini tentu ada kaitan dengan predikat Bali sebagai tujuan pariwisata dunia mendongkrak ekonomi Bali sehingga pertumbuhan ekonomi Bali dalam 5 tahun terakhir stabil berada di atas pertumbuhan ekonomi nasional. Akan tetapi pertumbuhan ekonomi tersebut kurang dapat terdistribusi dengan merata sehingga belum mampu untuk memberikan kesejahteraan bagi seluruh lapisan masyarakat bahkan tetap menyisakan kemiskinan bagi sebagian masyarakat.

Program pengentasan kemisikinan di Bali yang dilakukan pemerintah daerah telah secara gencar dilakukan yaitu dengan meluncurkan berbagai program yang menyasar langsung ke penduduk miskin. Program tersebut antara lain program jaminan kesehatan Bali mandara, beasiswa bagi penduduk miskin, gerbangsadu mandara, sistem pertanian terintegrasi dan jaminan kredit Bali mandara. Sebagian program tersebut menyasar langsung terhadap rumah tangga yang dianggap miskin. Akan tetapi pengukuran kemiskinan berdasarkan pendekatan absolut tidak dapat diimplementasikan secara praktis ketika ingin menetapkan rumah tangga miskin (BPS, 2016). Sehingga diperlukan suatu pendekatan non moneter dalam melihat kemiskinan.

Pendekatan non moneter dilaksanakan dengan mendasarkan pada karakteristik rumah tangga dalam penentuan rumah tangga miskin. Karakteristik rumah tangga tersebut bisa berupa kondisi tempat tinggal, sanitasi, kepemilikan aset dan lain-lain sehingga didapatkan suatu kriteria yang menentukan apakah

suatu rumah tangga tergolong miskin atau tidak. Pada Tahun 2000, BPS pernah melakukan Studi Penentuan Kriteria Penduduk Miskin (SKPM 2000) untuk mengetahui karakteristik-karakteristik rumah tangga yang mampu mencirikan kemiskinan secara konseptual. Studi ini dilakukan untuk mengidentifikasi rumah tangga miskin di lapangan sehingga didapatkan informasi yang berguna untuk penentuan sasaran rumah tangga program pengentasan kemiskinan (intervensi program). Studi dengan cakupan wilayah tujuh provinsi ini menghasilkan 8 variabel yang dianggap layak dan operasional untuk penentuan rumah tangga miskin di lapangan. Kedelapan variabel tersebut diperoleh dengan menggunakan metode *stepwise logistic regression* dan menghasilkan tingkat misklasifikasi sekitar 17 persen (BPS, 2016).

Ketidaktepatan variabel yang mencirikan kemiskinan tentu akan mengakibatkan terjadinya kebijakan yang tidak tepat. Oleh karena itu, dalam rangka mendapatkan karakteristik rumah tangga miskin yang tepat tentunya proses klasifikasi data menjadi proses yang sangat penting. Klasifikasi sendiri dapat diartikan sebagai proses menemukan sekumpulan model atau fungsi yang menggambarkan dan membedakan konsep atau kelas-kelas data, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek atau data (Han, Kamber dan Pei, 2001). Beberapa metode klasifikasi yang umum digunakan, antara lain analisis diskriminan, regresi logistik, *Chi-square Automatic Interaction Detection* (CHAID), *Classification and Regression Trees* (CART), *Multivariate Adaptive Regression Spline* (MARS) dan *Support Vector Machine* (SVM). Masing-masing metode klasifikasi tersebut mempunyai kelebihan dan kekurangan sehingga tidak terdapat satu metode pun yang *powerfull* terhadap semua data.

Penelitian tentang karakteristik dan klasifikasi tingkat kesejahteraan rumah tangga telah cukup banyak dilakukan. Nopriansyah, Junaidi dan Umiyati (2015) melakukan penelitian tentang determinan kemiskinan rumah tangga di Provinsi Jambi. Dalam penelitian tersebut metode analisis yang digunakan adalah metode regresi logistik dengan objek penelitian rumah tangga sampel pada Susenas (Survei Sosial Ekonomi Nasional) Tahun 2012 di Provinsi Jambi. Dari penelitian tersebut dihasilkan kesimpulan bahwa variabel yang mempengaruhi

kemiskinan rumah tangga di Provinsi Jambi adalah klasifikasi desa/kelurahan, jenis kelamin kepala rumah tangga, pendidikan kepala rumah tangga, pekerjaan kepala rumah tangga, jumlah anggota rumah tangga dan bantuan kredit usaha. Tingkat ketepatan klasifikasi yang dihasilkan adalah 93,10 persen. Namun yang perlu dijadikan perhatian dalam penelitian tersebut adalah nilai *specificity* yang dihasilkan tinggi dan *sensitivity* rendah. Hal ini dikarenakan penggunaan regresi logistik terhadap *imbalanced data* (data tak seimbang) dapat mengurangi ketepatan klasifikasi pada kelas yang minoritas (*rare event*).

Mirawanti (2012) membandingkan ketepatan metode regresi logistik ordinal dengan jaringan saraf tiruan fungsi radial basis dalam kasus klasifikasi rumah tangga miskin di Kota Pasuruan. Hasil dari penelitian tersebut adalah metode *radial basis function* memberikan tingkat akurasi yang lebih baik bila dibandingkan regresi logistik ordinal. Akan tetapi akurasi terbaik yang dihasilkan oleh model tersebut hanya sekitar 67,08 persen (untuk fungsi radial basis) dan 38 persen untuk regresi logistik ordinal. Sedangkan Arleina dan Otok (2014) dengan menggunakan metode *Bootstrap Aggregating Multivariate Adaptive Regression Splines* (Bragging MARS) mengklasifikasikan rumah tangga miskin di Kabupaten Jombang. Hasil penelitian tersebut menghasilkan 14 variabel penciri rumah tangga miskin dan ketepatan akurasi yang dihasilkan adalah 69,40 persen untuk data *training* dan 49,40 untuk data *testing*.

Metode regresi logistik sendiri dalam beberapa kasus cukup baik dalam melakukan analisis klasifikasi data (Pohlman dan Leitner, 2003). Metode regresi logistik digunakan untuk analisis data respon kategorik dengan variabel-variabel bebas dapat berupa kategorik maupun kontinyu (Hosmer dan Lemeshow, 2000). Metode ini memliliki beberapa kelebihan dan sering digunakan dalam berbagai penelitian. Selain metode ini telah dipelajari secara ekstensif, Hastie, Tibshirani, dan Friedman (2013) menyatakan bahwa regresi logistik mampu menjadi *classifier* yang tangguh dan mampu memberikan probabilitas dan mencakup masalah klasifikasi multi kelas. Regresi logistik tidak harus memenuhi asumsi kenormalan dan kesamaan varian, tetapi seperti halnya pada analisis regresi, asumsi multikolinieritas tetap harus dipenuhi dalam penggunaannya. Meskipun metode regresi logistik ini cukup baik apabila digunakan untuk klasifikasi data,

namun sayangnya metode ini tidak tepat digunakan pada kondisi *imbalanced data* (data tidak seimbang). King dan Zeng (2001) menyatakan bahwa ketika regresi logistik digunakan pada kasus *imbalanced data*, maka *classifier* cenderung menihilkan peluang dari kelompok minoritas karena nilai prediksi akan cenderung kepada kategori data yang mayoritas, sehingga tingkat ketepatan yang dihasilkan menjadi kurang baik.

Penanganan kasus *imbalanced data* pada regresi logistik terus menerus dilakukan dengan cara mencari metode yang tepat dan menghasilkan ketepatan klasifikasi yang baik utamanya terhadap kelas yang minoritas. King dan Zeng (2001) membuat suatu pengembangan metode regresi logistik pada *rare event* data. Owen (2007) juga mengembangkan metode dan menamakannya *Infinitely Imbalanced Logistic Regression* yang juga dirancang untuk menangani kasus *imbalanced data*. Maalouf dan Trafalis (2011) memperkenalkan metode *Rare Event Weighted Kernel Logistic Regression* (RE-WKLR) yang menggabungkan antara *weighting*, regularisasi, pendekatan metode numerik, kernelisasi, koreksi bias dan implementasi efisiensi yang kesemuanya itu dapat menjadikan metode RE-WKLR tersebut menjadi metode yang efektif, dan bagus untuk memprediksi kasus *rare event*. Akan tetapi kelemahan dari metode ini yaitu hanya disarankan untuk data yang berukuran kecil sampai menengah.

Maalouf dan Siddiqi (2014) mengembangkan suatu metode yang mengakomodir kelemahan regresi logistik dalam kasus *imbalanced data* untuk set data yang besar dan menyebut metode tersebut dengan nama *Rare Event Weighted Logistic Regression* (RE-WLR). RE-WLR merupakan pengembangan dari metode *Rare Event Weighted Kernel Logistic Regression* (RE-WKLR). Dari hasil penelitian Maalouf dan Siddiqi (2014) yang menggunakan enam set data dengan masing-masing jumlah sampel, jumlah variabel dan tingkat *rarity* yang beragam, diperoleh kesimpulan bahwa penggunaan RE-WLR lebih baik secara akurasi maupun kecepatan pengolahan dibandingkan metode *Support Vector Machine* (SVM) dan *Truncated Regularized Iteratively Re-Weighted Least Square* (TR-IRLS) dalam menangani kasus *imbalanced data* dengan jumlah data yang cukup besar.

Penggunaan metode RE-WLR juga digunakan oleh Sulasih (2016) dalam penelitiannya untuk mengklasifikasikan desa tertinggal di Provinsi Jawa Timur. Hasil penelitiannya menunjukkan bahwa metode RE-WLR secara deskriptif mampu memberikan kinerja klasifikasi yang lebih baik dibandingkan TR-IRLS. Sehingga dalam kesimpulannya metode RE-WLR dapat memprediksi kelas minoritas dalam kasus ini yakni desa tertinggal dengan baik.

Selain dengan menggunakan pendekatan penimbang (*weighting*), salah satu pendekatan yang dikemukakan oleh King dan Zeng (2001) adalah dengan pendekatan *prior correction*. Metode ini dikembangkan lagi oleh Maalouf, Homouz, dan Trafalis (2017) dengan memperkenalkan yang disebut *Truncated Regularised Prior Correction* (TR-PC). Dari hasil penelitian yang dilakukan oleh mereka disimpulkan bahwa metode TR-PC hampir sama baiknya dalam memberikan ketepatan klasifikasi pada kasus *imbalanced data* dalam jumlah yang besar.

Berdasarkan uraian di atas, maka penelitian ini akan menerapkan metode RE-WLR dan TR-PC dalam kasus pengklasifikasian tingkat kesejahteraan di Provinsi Bali dan membandingkan performa dari metode tersebut dengan metode regresi logistik biasa. Kedua metode ini digunakan dalam kasus klasifikasi tingkat kesejahteraan di Provinsi Bali karena terdapat perbedaan yang cukup jauh dan tak seimbang (*imbalanced*) pada data tingkat kesejahteraan, yakni antara proporsi rumah tangga tidak miskin ($\pm$ 95 persen) dan rumah tangga miskin ($\pm$ 5 persen) sehingga kasus ini termasuk ke dalam kategori *imbalanced data.*

1.2 Rumusan Masalah

Regresi logistik merupakan salah satu metode untuk klasifikasi data yang sering digunakan. Akan tetapi estimasi parameter yang dihasilkan dalam regresi logistik melalui metode *Maximum Likelihood Estimation* adalah bias baik untuk data berukuran kecil atau *imbalanced data*. Bias yang dihasilkan semakin besar apabila proporsi kelompok minoritas terlalu kecil dibandingkan kelompok mayoritas, sehingga perlu dilakukan modifikasi untuk menghasilkan tingkat ketepatan yang lebih baik. Penggunaan pendekatan *weighting* (penimbang) dan *prior correction* pada regresi logistik oleh King dan Zeng (2000) dan

dikembangkan oleh Maalouf, Homouz, dan Trafalis (2017) melalui metode RE-WLR dan TR-PC dianggap tepat dalam membentuk model yang dapat menghasilkan tingkat klasifikasi yang baik.

Sehingga dalam penelitian ini secara ringkas yang menjadi rumusan masalahnya adalah sebagai berikut:

1. Bagaimana mendapatkan estimasi parameter pada model RE-WLR dan TR-PC?
2. Bagaimana pemodelan klasifikasi tingkat kesejahteraan di Provinsi Bali dengan menggunakan metode RE-WLR dan TR-PC?
3. Apakah performa dari metode RE-WLR dan TR-PC itu jika dibandingkan dengan regresi logistik pada kasus *imbalanced data* mampu memberikan performa klasifikasi lebih baik?

1.3 Tujuan Penelitian

Berdasarkan rumusan permasalahan yang ada, maka tujuan yang ingin dicapai dari penelitian ini dapat dituliskan sebagai berikut:

1. Mengkaji estimasi parameter pada model RE-WLR dan TR-PC.
2. Memodelkan RE-WLR dan TR-PC untuk klasifikasi tingkat kesejahteraan di Provinsi Bali.
3. Membandingkan performa dan ketepatan hasil pengklasifikasian tingkat kesejahteraan di Provinsi Bali dengan menggunakan model RE-WLR, TR-PC dan regresi logistik.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Memberikan rekomendasi metode yang tepat untuk mengatasi permasalahan klasifikasi untuk kasus *imbalanced data.*
2. Menghasilkan karakteristik rumah tangga miskin yang lebih baik sehingga dapat digunakan oleh pemerintah dan pemangku kepentingan terkait untuk meningkatkan validitas terkait program pengentasan kemiskinan di Indonesia.

1.5 Batasan Penelitian

Dalam penelitian ini, dibuat batasan yakni sumber data yang digunakan hanya dari Susenas yang dikumpulkan BPS pada tahun 2015, sehingga pemilihan variabel karakteristik kemiskinan terbatas pada variabel yang terdapat dalam kuesioner Susenas 2015.

BAB 2
TINJAUAN PUSTAKA

Materi yang akan dibahas dalam bab ini dibedakan menjadi tinjauan pustaka di bidang statistik dan non statistik. Materi-materi untuk bidang statistik tersebut antara lain mengenai regresi logistik, RE-WLR, TR-PC, dan evaluasi performa klasifikasi. Sedangkan untuk materi non statistik adalah kajian mengenai kemiskinan dan karakteristiknya.

2.1 Regresi Logistik

Regresi logistik adalah salah satu metode analisis statistik yang digunakan dalam analisis data kategorik. Menurut Hosmer dan Lemeshow (2000) regresi logistik digunakan untuk set data dengan variabel respon berskala biner (dikotomus) atau hanya terdiri dari dua kategori, sedangkan variabel penjelasnya dapat berbentuk variabel kontinyu dan atau variabel dengan dua ataupun lebih dari dua kategori. Dalam kasus variabel respon biner, regresi logistik juga dapat digunakan untuk melakukan klasifikasi atau menentukan variabel yang berpengaruh terhadap variabel respon (Agresti, 2007).

Regresi logistik digunakan untuk menganalisa hubungan variabel independen X_1, X_2, ... , X_d terhadap variabel respon Y. Variabel respon Y terdiri dari dua kategori (misalnya Y=1 menyatakan sukses dan Y=0 menyatakan gagal), sehingga variabel respon Y mengikuti distribusi Bernoulli. Notasi matriksnya adalah

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{id} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}=\begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_i^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}, \qquad \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}.$$

Bentuk umum model logistik adalah sebagai berikut:

$$\mathrm{p_i}=\frac{\exp(\beta_0+\beta_1 x_{i1}+...+\beta_d x_{id})}{1+\exp(\beta_0+\beta_1 x_{i1}+...+\beta_d x_{id})}=\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}=\frac{1}{1+e^{-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}. \qquad (2.1)$$

Keterangan:

p_i = peluang sukses untuk data ke-i

d = banyaknya parameter

β_0 = *intercept*

β_d = koefisien untuk variabel independen ke-d

Dengan menggunakan transformasi logit dari p_i, maka model logistik dapat disebut model logit yang ditunjukkan oleh:

$$\text{logit}[p_i] = g(x_i) = \ln\left[\frac{p_i}{1-p_i}\right] = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \tag{2.2}$$

Dalam model regresi logistik, variabel respon mengikuti distribusi Bernoulli dengan fungsi densitas peluang

$$f(y_i) = [p_i]^{y_i}\,[1-p_i]^{1-y_i}. \tag{2.3}$$

Metode estimasi parameter yang umum dan sering dipakai adalah metode *Maximum Likelihood* (ML). Metode ini didasarkan kepada *joint probability density* dari data yang diamati, dan bertindak sebagai fungsi dari parameter yang tidak diketahui dalam model (Garthwaite, Jolliffe dan Byron, 2002). Sehingga dengan mengasumsikan antar observasi diasumsikan saling bebas, dengan Y_i ; i = 1,2,...,n adalah sampel acak yang sudah terambil, maka diperoleh fungsi likelihood sebagai berikut:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} [p_i]^{y_i}\,[1-p_i]^{1-y_i}. \tag{2.4}$$

Dengan mensubtitusikan (2.1) ke dalam persamaan (2.4) maka didapatkan persamaan fungsi likelihood sebagai berikut:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\beta_0 + \sum_{d=1}^{D}\beta_d x_{id})}{1+\exp(\beta_0 + \sum_{d=1}^{D}\beta_d x_{id})}\right]^{y_i}\left[\frac{1}{1+\exp(\beta_0 + \sum_{d=1}^{D}\beta_d x_{id})}\right]^{1-y_i}. \tag{2.5}$$

Sehingga log-likelihood dari persamaan (2.5) adalah

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i \ln\left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) + (1-y_i)\ln\left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right)\right). \tag{2.6}$$

Akan tetapi, estimasi parameter dengan metode ML tidak menghasilkan solusi yang berbentuk *closed form*, sehingga estimasi parameter untuk regresi logistik dengan metode ML ini diperoleh dengan menggunakan metode optimisasi numerik (*numerical optimisation*). Salah satu metode numerik yang biasa dipakai adalah menggunakan metode Newton-Rhapson.

2.2 *Regularization*

Dalam penggunaan regresi logistik untuk klasifikasi, umumnya dengan menggunakan data *training* dan data *testing* untuk menguji keakuratan model yang dibentuk. Permasalahan bisa muncul yakni terjadinya *over fitting* pada data *training* yaitu suatu kondisi dimana model yang dihasilkan bagus untuk data *training* tapi mempunyai performa jelek untuk data baru (data *testing*), utamanya ketika data berdimensi tinggi (Hosmer dan Lemeshow, 2000). Dalam mengatasi permasalahan tersebut, salah satu pendekatan yang dapat mengurangi *over fitting* tersebut adalah melalui *regularization* (Bielza, Robles, dan Larranaga, 2011; Bishop, 2006). Sehingga *regularized log likelihood* yang dihasilkan untuk model *regularized logistic regression* (Maalouf dan Siddiqi, 2014) adalah

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) + (1-y_i) \ln\left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right] - \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{2.7}$$

$$= \sum_{i=1}^{n} \ln\left[\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right] - \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2. \tag{2.8}$$

Dengan nilai $\|\boldsymbol{\beta}\|^2 = \sum_{d=0}^{D} \beta_d^2$ adalah *penalty* dan $\lambda > 0$ adalah parameter regularisasi. Penambahan unsur $\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$ adalah sebagai regularisasi atau penalti yang digunakan agar meningkatkan kinerja generalisasi dan mengatasi masalah *over fitting*.

Perlunya suatu regularisasi ditambahkan ke dalam model adalah terletak kepada penentuan titik temu antara bias-varians dari estimator. Ketika λ bernilai sangat kecil, maka bias berkurang tapi varians bertambah, sebaliknya nilai λ yang besar akan mengakibatkan bias bertambah tapi varians berkurang. Selain itu

penambahan unsur regularisasi menghindarkan terjadinya kompleksitas pada model.

Pada regresi logistik, terdapat istilah Deviance (DEV), yaitu suatu statistik yang memainkan peran penting pada beberapa pendekatan untuk menilai *goodness of fit* (Hosmer dan Lemeshow, 2000). Deviance dirumuskan sebagai negatif log likelihood seperti pada formula berikut

$$\mathrm{DEV}(\hat{\boldsymbol{\beta}}) = -2\ln \mathrm{L}(\hat{\boldsymbol{\beta}}) \tag{2.9}$$

Dengan meminimumkan nilai $\mathrm{DEV}(\hat{\boldsymbol{\beta}})$ adalah equivalen dengan memaksimumkan nilai log likelihoodnya. Maalouf dan Siddiqi (2014) menyatakan bahwa fungsi *deviance* di atas non linier terhadap $\boldsymbol{\beta}$ sehingga untuk meminimumkannya membutuhkan metode numerikal dalam rangka mencari *Maximum Likelihood Estimator* (MLE) dari $\boldsymbol{\beta}$. Metode numerikal tersebut salah satunya adalah dengan metode IRLS (*Iteratively Re-Weighted Least Squares*).

2.3 *Iteratively Re-Weighted Least Squares* (IRLS)

Salah satu teknik yang digunakan untuk mendapatkan MLE dari $\boldsymbol{\beta}$ adalah dengan menggunakan teknik IRLS, yang menggunakan algoritma Newton-Rhapson. Tiap iterasi pada teknik ini mendapatkan perkiraan *weighted least squares* (WLS) untuk setiap set penimbang tertentu, yang digunakan untuk membentuk sebuah set penimbang yang baru (Garthwaite dkk., 2002). Vektor gradien dan matriks Hessian juga digunakan dalam teknik ini, yang didapatkan dari turunan pertama dan kedua dari *regularised* log-likelihood pada persamaan (2.7). Vektor Gradien yang diperoleh adalah

$$\frac{\partial}{\partial \beta_j}\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i\left(\frac{x_{ij}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) + (1-y_i)\left(\frac{-x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right)\right) - \lambda\beta_j \tag{2.10}$$

$$= \sum_{i=1}^{n}\left(x_{ij}\left(y_i - p_i\right)\right) - \lambda\beta_j = 0. \tag{2.11}$$

Dimana $j = 0,...,d$ dan d adalah jumlah parameter. Sehingga bila dituliskan dalam bentuk matriks adalah

$$\frac{\partial}{\partial \boldsymbol{\beta}}\ln \mathrm{L}(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{p}) - \lambda\boldsymbol{\beta} = 0. \tag{2.12}$$

Sedangkan matriks Hessian yang diperoleh adalah

$$\frac{\partial^2}{\partial\beta_j\beta_k}\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{-x_{ij}x_{ik}e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)}\right)-\lambda \tag{2.13}$$

$$= \sum_{i=1}^{n}\left(-x_{ij}x_{ik}\left(p_i\left(1-p_i\right)\right)\right)-\lambda. \tag{2.14}$$

Sehingga bila kita definisikan v_i adalah $p_i(1-p_i)$ dan $\mathbf{V} = diag(v_1,...,v_n)$, maka matriks Hessian dapat ditulis sebagai berikut

$$H(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{V}\mathbf{X} - \lambda\mathbf{I} \tag{2.15}$$

dimana $\mathbf{I}$ adalah matriks identitas berukuran $(d+1)\times(d+1)$.

Dengan diperolehnya turunan pertama dan kedua, maka formula untuk update Newton-Rhapson pada iterasi ke c+1 adalah

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\left(\mathbf{y}-\mathbf{p}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right). \tag{2.16}$$

Karena $\hat{\boldsymbol{\beta}}^{(c)} = \left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c)}$, maka persamaan (2.16) dapat ditulis menjadi

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\left(\mathbf{V}\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{y}-\mathbf{p}\right)\right) \tag{2.17}$$

$$= \left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}^{(c)}, \tag{2.18}$$

dimana $\mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{p}\right)$ disebut juga dengan *adjusted response* (Hastie dkk, 2013).

Dibalik keuntungan yang didapatkan dalam penggunaan parameter regularisasi (λ), tetapi akibatnya akan mengakibatkan perhitungan iterasi komputasi yang sangat lamban (Komarek, 2004). Sehingga dibutuhkan jalan tengah antara kecepatan konvergensi dan akurasi dari perhitungan. Metode yang dapat menampung masalah tersebut dikenal sebagai metode Truncated Newton.

Komarek dan Moore (2005) membuat suatu modifikasi linear CG untuk memperkirakan *Newton direction* dalam IRLS regresi logistik dan menamakannya TR-IRLS (Truncated Regularized-IRLS). Algoritma dari TR-IRLS terdiri dari dua bagian, yang pertama adalah iterasi untuk mencari solusi dari *WLS problem* dan menghentikannya ketika selisih relatif Deviance dari dua iterasi yang berurutan

tidak lebih dari batas yang telah ditentukan (ε_1). Sedangkan bagian kedua yaitu iterasi untuk mencari solusi bagi persamaan WLS subproblem pada bagian pertama dengan cara metode linier CG dan menghentikan iterasi ketika nilai residual

$$\mathbf{r}^{(c+1)} = \left(\mathbf{X}^T\mathbf{V}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c+1)} - \mathbf{X}^T\mathbf{V}\mathbf{z}^{(c)} \tag{2.19}$$

tidak melebihi dari batas yang telah ditentukan (ε_2).

Akan tetapi metode IRLS ini tidak didesain untuk menangani kasus *imbalanced data* sehingga digunakan pendekatan regresi logistik dengan menggunakan *weighting* dan *prior correction* dalam menangani kasus tersebut.

2.4 *Rare Event Weighted Logistic Regression* (RE-WLR)

Dalam rangka mengatasi kelemahan regresi logistik terhadap kasus *imbalanced data*, maka Maalouf dan Siddiqi (2014) mengembangkan metode RE-WLR. Metode RE-WLR ini merupakan pengembangan dari RE-WKLR guna mengatasi permasalahan untuk data yang berukuran besar. RE-WLR ini menerapkan regularisasi, *weighting*, dan *bias correction* pada regresi logistik. Secara definisi, yang dimaksud *rare events* adalah suatu kejadian yang frekuensi terjadinya secara substansial lebih kecil daripada frekuensi kejadian pada umumnya.

Menurut King dan Zeng (2001) untuk mengatasi kasus *imbalanced data* atau *rare events data* maka dapat dilakukan dengan cara pembobotan. Dalam pengambilan data training pada akhirnya akan diketahui proporsi kejadian sukses (*event*) dalam sampel (data training) yang dilambangkan dengan $\bar{y}$. Sedangkan proporsi kejadian sukses (*event*) dalam populasi (seluruh data) dilambangkan τ.

Pembobotan dilakukan karena proporsi suatu kelompok terlalu kecil dibandingkan kelompok yang lain sehingga diperlukan suatu *treatment* pada data *training* yang memungkinkan suatu model dapat menangkap informasi lebih tepat dibandingkan ketika pengambilan sampel pada data *training* hanya menggunakan metode acak. Pembobotan ini dilakukan saat memodelkan dengan menggunakan

data *training*, sehingga fungsi log likelihood untuk *weighted logistic regression* adalah dengan menambahkan bobot w_i pada log likelihood regresi logistik biasa.

Sehingga persamaan log likelihoodnya menjadi

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) \tag{2.20}$$

dengan $w_i = \left(\frac{\tau}{\bar{y}}\right) y_i + \left(\frac{1-\tau}{1-\bar{y}}\right)(1-y_i)$ (2.21)

dimana $\tau =$ proporsi kejadian sukses (*event*) dalam seluruh data.

$\bar{y} =$ proporsi kejadian sukses (*event*) dalam data *training*.

Dapat dilihat dari formula (2.20) dan (2.21) bahwa ketika proporsi terjadinya *event* pada data *training* dan proporsi terjadinya *event* pada populasi nilainya sama maka fungsi log likelihood yang dihasilkan adalah sama dengan log likelihood regresi logistik biasa tanpa penimbang.

Pada WLR, gradien dan Hessian didapatkan dengan cara menurunkan *regularized weighted log-likelihood* terhadap masing-masing β.

$$\ln \mathrm{L}_W(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln \frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2. \tag{2.22}$$

Sehingga bila dituliskan dalam bentuk matriks, didapatkan persamaan gradien adalah

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathrm{L}_W(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}(\mathbf{y}-\mathbf{p}) - \lambda\boldsymbol{\beta}, \tag{2.23}$$

dimana $\mathbf{W} = diag(w_i)$ dan $\mathbf{p}$ adalah vektor peluang seperti pada persamaan (2.1).

Sedangkan Matriks Hessiannya adalah

$$H(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{D}\mathbf{X} - \lambda\mathbf{I}. \tag{2.24}$$

Update Newton-Rhapson untuk β pada iterasi ke-(c+1) adalah

$$\hat{\boldsymbol{\beta}}^{(c+1)} = (\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{D}\mathbf{z}^{(c)}, \tag{2.25}$$

dengan $\mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}(\mathbf{y}-\mathbf{p})$ (2.26)

adalah variabel dependen yang disesuaikan.

Sehingga diperoleh *subproblem* untuk *Weighted least squares* adalah

$$(\mathbf{X}^T\mathbf{D}\mathbf{X}+\lambda\mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)}=\mathbf{X}^T\mathbf{D}\mathbf{z}^{(c)}. \quad (2.27)$$

Permasalahan proses iteratif di atas dapat diatasi dengan menggunakan metode Conjugate Gradient (CG) dengan memberikan nilai limit tertentu pada jumlah iterasi CG guna menghindari perhitungan komputer yang terlalu lama.

Selain penambahan unsur pembobot, King dan Zeng (2001) menambahkan unsur *bias corrections*, sebagaimana yang diberikan oleh McCullagh dan Nelder (1989) dimana tujuan utamanya adalah mengurangi bias dan varians. Lazim diketahui bahwa penduga parameter dalam regresi logistik melalui metode MLE adalah bias untuk data yang kecil atau untuk kasus *imbalanced data*. Contoh kasus seperti yang diberikan oleh King dan Zeng (2001), jika $p_i=\frac{1}{1+e^{-\beta_0+x_i}}$, maka besaran biasnya adalah $\mathrm{E}\left[\hat{\beta}_0-\beta_0\right]\approx\frac{\bar{p}-0,5}{n\bar{p}(1-\bar{p})}$, dimana $\bar{p}$ adalah proporsi kejadian pada sampel. Sehingga terlihat bahwa selama $\bar{p}$ tidak sama dengan 0,5 dan atau jumlah n kecil maka bias tidak akan sama dengan nol. Hal ini mengindikasikan bahwa penting untuk menambahkan unsur koreksi bias pada penduga parameter.

Formula untuk vektor bias (McCullagh dan Nelder , 1989) pada regresi logistik adalah

$$bias(\hat{\boldsymbol{\beta}})=(\mathbf{X^TVX})^{-1}\mathbf{X^TV}\xi, \quad (2.28)$$

dengan $\xi_i=Q_{ii}(\hat{p}_i-\frac{1}{2})$, dan Q_{ii} adalah elemen diagonal dari matriks

$$\mathbf{Q}=\mathbf{X}(\mathbf{X^TVX})^{-1}\mathbf{X}^T \quad (2.29)$$

yaitu perkiraan matriks kovarian dari fungsi link logistik η. Sedangkan $\mathbf{V}=diag\left(v_1,...,v_n\right)$ dengan nilai dari v_i adalah

$$v_i=p_i(1-p_i). \quad (2.30)$$

Sehingga estimator untuk parameter dengan koreksi bias pada regresi logistik adalah

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}-bias(\hat{\boldsymbol{\beta}}). \quad (2.31)$$

Dengan menggunakan pendekatan *weighting* maka didapatkan vektor bias untuk WLR dapat ditulis dengan

$$\mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)=\left(\mathbf{X^T D X}\right)^{-1}\mathbf{X^T D}\boldsymbol{\xi}, \tag{2.32}$$

dimana elemen ke-i vektor ξ pada saat ini sama dengan

$$\xi_i = 0,5Q_{ii}\left(1+w_1 p_i - w_1\right), \tag{2.33}$$

dengan Q_{ii} adalah elemen diagonal dari matriks **Q**, yang sekarang menjadi $\mathbf{Q}=\mathbf{X}\left(\mathbf{X^T D X}\right)^{-1}\mathbf{X}^T$, dan

$$\mathbf{D} = diag\left(v_i w_i\right) \tag{2.34}$$

untuk $i=1,2,...,n$. Sehingga didapatkan estimator parameter dengan bias terkoreksi untuk WLR adalah

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}-\mathbf{B}(\hat{\boldsymbol{\beta}}). \tag{2.35}$$

Setelah mendapatkan $\tilde{\beta}$ yaitu estimator parameter terkoreksi dengan algoritma linier CG, maka dengan memasukkan estimator tersebut ke dalam persamaan logit maka didapatkan nilai prediksi peluang sebagai berikut:

$$\tilde{p}_i = \frac{e^{\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}}}. \tag{2.36}$$

Untuk menguji estimasi parameter yang dihasilkan dalam model RE-WLR secara parsial, dapat menggunakan statistik uji Wald (Hosmer dan Lemeshow, 2000). Secara umum, hipotesis dan rumus yang digunakan untuk menguji estimasi parameter secara parsial dengan menggunakan uji wald adalah sebagai berikut:

$\mathrm{H}_0 : \beta_j = 0$

$\mathrm{H}_1 : \beta_j \neq 0,$ dengan j=1,2,...,p.

Statistik ujinya adalah:

$$\mathrm{W}_j = \left[\frac{\tilde{\beta}_j}{\mathrm{Se}\left(\tilde{\beta}_j\right)}\right]^2 \text{ ;dengan j=1,2,...,p.}$$

dimana statistik uji W_j mengikuti sebaran χ^2 dengan derajat bebas satu. Formula $\mathrm{Se}\left(\tilde{\boldsymbol{\beta}}\right)$ untuk model RE-WLR (Maalouf dkk, 2014) adalah

$$\mathrm{Se}\left(\tilde{\boldsymbol{\beta}}\right) = \sqrt{\left(\frac{n}{n+d}\right)^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\right)^{-1}}$$

Pada uji Wald, H_0 ditolak apabila nilai W_j lebih dari nilai $\chi^2_{(1)}$ atau apabila nilai p-value kurang dari α.

2.5 *Truncated-Regularized Prior Correction*

Selain dengan metode pembobotan (*weighting*), King dan Zeng (2000) memberikan solusi lain dalam mengatasi permasalahan pada *imbalanced data*, yakni dengan metode *Prior Correction*. Maalouf, Homouz dan Trafalis (2017) menggunakan teknik ini dan mengembangkannya dengan metode Truncated Newton sehingga didapatkan metode regresi logistik *Truncated Regularized Prior Correction* (TR-PC). Secara struktur algoritma, metode ini hampir sama dengan algoritma pada RE-WLR karena sama-sama menggunakan algoritma TR-IRLS yang didasarkan pada metode linier Conjugate Gradient (CG).

Penggunaan *prior correction* melibatkan regresi logistik biasa dan mengkoreksi perhitungannya berdasar pada informasi prior untuk mengatasi kasus *imbalanced data*. Informasi prior tersebut berupa rasio dari terjadinya *event* pada populasi (dilambangkan τ) dan rasio terjadinya *event* pada sampel (data *training*) yang dilambangkan $\bar{y}$. Kedua informasi itu membentuk ukuran

$$\delta_1 = \left(\frac{\bar{y}}{\tau}\right) \quad \text{dan} \quad \delta_0 = \left(\frac{1-\bar{y}}{1-\tau}\right).$$

King dan Zeng (2000) mendefinisikan *modified logit probability* dari regresi logistik dengan *prior correction* adalah sebagai berikut

$$\tilde{p}_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}, \quad \text{untuk } i = 1,...,n. \tag{2.37}$$

Sehingga dari formula tersebut dan menambahkannya dengan regularisasi, Maalouf dkk (2017) mendapatkan *regularized* log likelihood untuk TR-PC adalah sebagai berikut:

$$\ln \mathrm{L}_{PC}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right) - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right] - \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 . \quad (2.38)$$

Untuk memaksimalkan nilai likelihood di atas maka formula di atas diturunkan terhadap masing-masing β. Sehingga dari turunan pertama, didapatkan vektor gradiennya adalah

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln \mathrm{L}_{PC}(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \tilde{\mathbf{p}}) - \lambda \boldsymbol{\beta}, \quad (2.39)$$

dengan $\tilde{\mathbf{p}}$ adalah vektor probabilitas terkoreksi pada persamaan (2.37). Sedangkan turunan kedua *regularized* log likelihood TR-PC akan didapatkan matriks Hessian sebagai berikut:

$$H(\boldsymbol{\beta}) = -\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X} - \lambda \mathbf{I}. \quad (2.40)$$

dengan

$$\mathbf{D} = diag(v_i) \quad (2.41)$$

dimana

$$v_i = (\tilde{p}_i)(1-\tilde{p}_i). \quad (2.42)$$

Dan update Newton-Rhapson untuk β pada iterasi ke-(c+1) adalah

$$\hat{\boldsymbol{\beta}}^{(c+1)} = (\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{D}\mathbf{z}^{(c)}, \quad (2.43)$$

dengan

$$\mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}(\mathbf{y} - \tilde{\mathbf{p}}) \quad (2.44)$$

adalah variabel dependen yang disesuaikan.

Sehingga diperoleh *subproblem* untuk *Weighted least squares* adalah

$$(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda \mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{z}^{(c)}. \quad (2.45)$$

Sama halnya dengan subproblem pada RE-WLR, subproblem pada (2.45) dapat diselesaikan melalui iteratif menggunakan metode CG dengan Truncated Newton.

Senada dengan RE-WLR, metode TR-PC menggunakan *bias correction* guna mengurangi bias dan varians. Formula untuk menghitung vektor bias pada TR-PC adalah

$$\mathbf{B}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}\xi, \quad (2.46)$$

dengan $\mathbf{D} = diag\left(\left(\tilde{p}_i\right)\left(1-\tilde{p}_i\right)\right),$ dan bias responsenya adalah

$$\xi_i = Q_{ii}(\tilde{p}_i - \frac{1}{2}) \tag{2.47}$$

dengan Q_{ii} adalah elemen diagonal dari matriks

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \tag{2.48}$$

Sehingga didapatkan estimasi parameter terkoreksi adalah

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}(\hat{\boldsymbol{\beta}}). \tag{2.49}$$

Setelah mendapatkan $\tilde{\beta}$ yaitu estimator parameter terkoreksi dengan algoritma linier CG, maka dengan memasukkan estimator tersebut ke dalam persamaan logit maka didapatkan nilai prediksi peluang sebagai berikut:

$$\tilde{p}_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}}}. \tag{2.50}$$

Untuk menguji estimasi parameter yang dihasilkan dalam model TR-PC secara parsial, juga dapat menggunakan statistik uji Wald (Hosmer dan Lemeshow, 2000). Secara umum, hipotesis dan rumus yang digunakan untuk menguji estimasi parameter secara parsial dengan menggunakan uji wald adalah sebagai berikut:

$\mathrm{H}_0 : \beta_j = 0$

$\mathrm{H}_1 : \beta_j \neq 0,$ dengan j=1,2,...,p.

Statistik ujinya adalah:

$$\mathrm{W}_j = \left[\frac{\tilde{\beta}_j}{\mathrm{Se}\left(\tilde{\beta}_j\right)}\right]^2 \text{ ;dengan j=1,2,...,p.}$$

dimana statistik uji W_j mengikuti sebaran χ^2 dengan derajat bebas satu. Formula $\mathrm{Se}\left(\tilde{\boldsymbol{\beta}}\right)$ untuk model TR-PC (Maalouf dkk, 2017) adalah

$$\mathrm{Se}\left(\tilde{\boldsymbol{\beta}}\right) = \sqrt{\left(\frac{n}{n+d}\right)^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\right)^{-1}}$$

Pada uji Wald, H_0 ditolak apabila nilai W_j lebih dari nilai $\chi^2_{(1)}$ atau apabila nilai p-value kurang dari α.

2.6 Evaluasi Performa Klasifikasi

Guna mengevaluasi performa suatu metode klasifikasi adalah dengan menggunakan *confusion matrix*. Dari *confusion matrix* tersebut dapat dihitung beberapa ukuran untuk mengevaluasi performa klasifikasi. *Confusion matrix* adalah alat yang sangat berguna untuk menganalisis seberapa baik suatu metode klasifikasi mengenali objek dari kelas yang berbeda (Han, Kamber, dan Pei, 2011). *Confusion matrix* untuk dua kelas dengan respon biner sesuai dengan jumlah kelas dapat dilihat pada Tabel 2.1 berikut:

Tabel 2.1 *Confusion matrix* untuk hasil klasifikasi biner

Kelas Sebenarnya	**Kelas Prediksi**		
	Positif	**Negatif**	**Total**
Positif	TP	FN	TP+FN
Negatif	FP	TN	FP+TN
Total	TP+FP	FN+TN	TP+FN+FP+TN

Keterangan:

TP = *True Positive* (Sebenarnya positif dan diklasifikasikan positif).

TN = *True Negative* (Sebenarnya negatif dan diklasifikasikan negatif).

FP = *False Positive* (Sebenarnya negatif tetapi diklasifikasikan positif).

FN = *False Negative* (Sebenarnya positif tetapi diklasifikasikan negatif).

Terdapat beberapa ukuran yang dapat digunakan dalam menilai kelayakan suatu metode dalam pengklasifikasian menurut Fawcett (2006), yaitu ukuran *accuracy* (akurasi), *sensitivity* (sensitivitas), dan *specificity* (spesifisitas).

- *Accuracy* merupakan nilai persentase ketepatan klasifikasi secara keseluruhan. Nilai akurasi yang semakin mendekati 100% menunjukkan performa klasifikator semakin tinggi/baik.

$$Accuracy(\%) = \frac{TP + TN}{TN + FP + FN + TN} \tag{2.51}$$

  Akan tetapi penggunaan indikator ini tidak tepat jika digunakan pada kasus *imbalanced data*.

- *Sensitivity* menunjukkan nilai persentase kebenaran dari kejadian-kejadian yang diinginkan terletak pada kelompok tertentu. Semakin tinggi nilai sensitivitas maka semakin baik pula performa klasifikator.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.52}$$

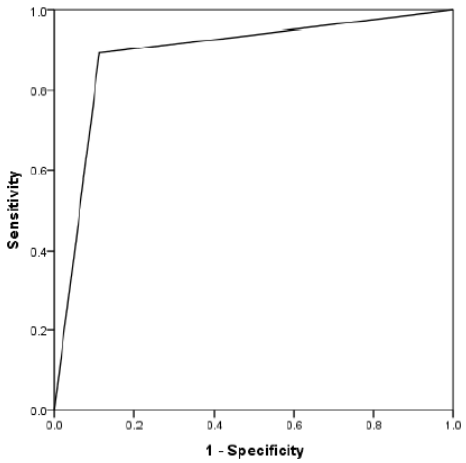
- *Specificity* menunjukkan nilai persentase kebenaran dari kejadian-kejadian yang tidak diinginkan. Semakin tinggi nilai spesifisitas maka semakin baik pula performa klasifikator.

$$Specificity = \frac{TN}{TN + FP} \tag{2.53}$$

Sebagai indikator yang dapat mengakomodir akurasi pada kelas positif (*sensitivity*) dan kelas negatif (*specificity*) dalam kasus *imbalanced data*, maka kinerja klasifikasi dihitung juga menggunakan indikator G-Mean dengan formula

$$G - mean = \sqrt{sensitivity\ x\ specificity} \tag{2.54}$$

Selain dengan menggunakan G-mean, dalam melihat apakah suatu model dikatakan tepat dalam melakukan prediksi maka dapat digunakan alat bantu yang bernama kurva ROC (*Receiver Operating characteristics*). Kurva ROC merupakan grafik perbandingan antara *sensitivity* pada sumbu vertikal dan *specificity* pada sumbu horizontal (Fawcett, 2006). Kurva ROC adalah gambaran dua dimensi dari kinerja suatu klasifikator. Alat ukur yang digunakan adalah AUC (*Area Under the Receiving Operating Characteristics Curve*) yaitu luas daerah di bawah kurva ROC.



Gambar 2.1 Kurva ROC

Tabel kategori pengklasifikasian menurut nilai AUC (Mohanty et.al, 2011) dapat dilihat pada tabel 2.2 berikut.

Tabel 2.2 Kategori Pengklasifikasian berdasarkan Nilai AUC

Nilai AUC	Model Diklasifikasikan Sebagai
$0,9 \leq AUC \leq 1,0$	*Excellent*
$0,8 \leq AUC < 0,9$	*Good*
$0,7 \leq AUC < 0,8$	*Fair*
$0,6 \leq AUC < 0,7$	*Poor*
$0,5 \leq AUC < 0,6$	*Fail*

2.7 Kemiskinan

Berikut ini akan dijelaskan beberapa hal terkait konsep kemiskinan, yaitu definisi kemiskinan, pengukuran garis kemiskinan, indikator kemiskinan dan kajian mengenai karakteristik kemiskinan.

2.7.1 Konsep dan Definisi

Kemiskinan berhubungan erat dengan arti kesejahteraan. Berdasarkan Rancangan Undang-undang tentang Sistem Kesejahteraan Sosial Nasional, kesejahteraan sosial adalah kondisi sosial ekonomi yang memungkinkan bagi setiap warga negara untuk dapat memenuhi kebutuhan yang bersifat jasmani, rohani, dan sosial sesuai dengan harkat dan martabat manusia. Kesejahteraan merupakan kondisi seseorang dapat memenuhi kebutuhan dasar baik material maupun non material yang mencakup aspek gizi dan kesehatan, pengetahuan, dan kekayaan materi.

Konsep dan definisi tentang kesejahteraan di atas tentunya hal yang bertolak belakang dengan pengertian kemiskinan menurut *World Bank* (2000) yang didefinisikan sebagai "*poverty is pronounced deprivation in well-being*" yang bermakna bahwa kemiskinan adalah suatu keadaan kehilangan kesejahteraan. Batasan kesejahteraan di Indonesia sendiri dapat didekati dengan istilah "fakir miskin" yang tertuang di dalam UU No.13 Tahun 2011 tentang penanganan fakir miskin. Fakir miskin adalah orang yang sama sekali tidak mempunyai sumber mata pencaharian dan atau mempunyai sumber mata

pencaharian, tetapi tidak mempunyai kemampuan memenuhi kebutuhan dasar yang layak bagi kehidupan dirinya dan atau keluarganya. Kebutuhan dasar yang dimaksud meliputi kebutuhan pangan, sandang, perumahan, kesehatan, pendidikan, pekerjaan, dan atau pelayanan sosial.

Sedangkan Bappenas (2004) mendefinisikan kemiskinan sebagai kondisi dimana seseorang atau sekelompok orang, laki-laki dan perempuan, tidak mampu memenuhi hak-hak dasarnya untuk mempertahankan dan mengembangkan kehidupan yang bermartabat. Hak-hak dasar tersebut antara lain terpenuhinya kebutuhan pangan, kesehatan, pendidikan, pekerjaan, perumahan, air bersih, pertanahan, sumber daya alam dan lingkungan hidup, rasa aman dari perlakuan atau ancaman tindak kekerasan dan hak untuk berpartisipasi dalam kegiatan sosial politik, baik bagi laki-laki maupun perempuan.

Menurut BPS (2016), secara konseptual, kemiskinan dapat dibedakan menurut kemiskinan relatif dan kemiskinan absolut, dimana perbedaannya terletak pada standar penilaiannya. Dalam kemiskinan relatif, standar penilaian kehidupannya ditentukan dan ditetapkan secara subyektif oleh masyarakat setempat dan bersifat lokal sehingga mereka yang berada di bawah standar tersebut dikategorikan sebagai miskin secara relatif. Sedangkan kemiskinan absolut mendefinisikan standar penilaian kehidupan minimum yang dibutuhkan untuk memenuhi kebutuhan dasar yang diperlukan baik makanan dan non makanan. Standar kehidupan minimum tersebut yang disebut sebagai garis kemiskinan, sedangkan penduduk yang pendapatannya di bawah garis kemiskinan dikategorikan sebagai penduduk miskin.

2.7.2 Garis Kemiskinan

Di dalam pengukuran tentang indikator kemiskinan, garis kemiskinan menjadi suatu komponen penting yang digunakan dalam penghitungan kemiskinan itu sendiri. Terdapat beberapa pendekatan dalam menetapkan garis kemiskinan. Salah satu pendekatan garis kemiskinan yang digunakan oleh Bank Dunia adalah dengan menggunakan perkiraan konsumsi yang dikonversikan ke

Dollar Amerika dengan menggunakan paritas (kesetaraan) daya beli (*Purchasing Power Parity*, PPP) per hari, bukan dengan nilai tukar US Dollar resmi. Angka konversi PPP adalah banyaknya rupiah yang dikeluarkan untuk membeli sejumlah kebutuhan barang dan jasa yang dikeluarkan untuk membeli sejumlah kebutuhan barang dan jasa dimana jumlah yang sama tersebut dapat dibeli sebesar US$ 1 di Amerika Serikat. Berdasarkan penelitian yang telah dilakukan oleh Bank Dunia, ditetapkan bahwa garis kemiskinan internasional adalah sebesar 1,25 dollar AS PPP per kapita per hari.

Sedangkan untuk pendekatan garis kemiskinan di Indonesia, BPS mendefinisikan garis kemiskinan sebagai nilai rupiah yang harus dikeluarkan seseorang dalam sebulan supaya dapat memenuhi kebutuhan dasar asupan kalori sebesar 2.100 kkal/hari per kapita (garis kemiskinan makanan), ditambah dengan kebutuhan minimum non makanan yang merupakan kebutuhan dasar seseorang, yaitu papan, sandang, sekolah, dan transportasi serta kebutuhan individu dan rumah tangga dasar lainnya (garis kemiskinan non makanan). Sehingga melalui pendekatan ini, kemiskinan diartikan sebagai ketidakmampuan dari dimensi ekonomi dalam memenuhi kebutuhan dasar baik makanan dan non makanan yang diukur dari sisi pengeluaran, yang kemudian batasan dari sisi pengeluaran tersebut disebut sebagai garis kemiskinan.

Ukuran 2.100 kkal/hari per kapita yang menjadi standar minimum sendiri merupakan rekomendasi dari Widyakara Nasional Pangan dan Gizi Tahun 1978 dan angka tersebut telah juga menjadi kesepakatan dunia. Hal ini sejalan dengan hasil pertemuan di Roma Tahun 2001, FAO (*Food and Agriculture Organization*) dan WHO (*World Health Organization*) yang menyatakan bahwa berdasarkan hasil kajian mendalam oleh para pakar, batas minimal kebutuhan manusia untuk mampu bertahan hidup dan mampu bekerja adalah sekitar 2.100 kilo kalori ditambah dengan kebutuhan paling mendasar bukan makanan (Hasbullah, 2012).

2.7.3 Indikator Kemiskinan

Terdapat 3 ukuran indikator kemiskinan BPS yang merujuk pada Foster, Greer dan Thorbecke (1984) yang sering dijadikan tolok ukur dalam program pengentasan kemiskinan. Indikator tersebut antara lain:

1. Persentase Penduduk Miskin (*Head Count Index* - P_0) adalah persentase penduduk miskin yang berada di bawah Garis Kemiskinan (GK).
2. Indek Kedalaman Kemiskinan (*Poverty Gap Index* - P_1) adalah ukuran rata-rata kesenjangan pengeluaran masing-masing penduduk miskin terhadap garis kemiskinan. Semakin tinggi nilai indeks, maka semakin jauh rata-rata pengeluaran penduduk dari garis kemiskinan.
3. Indeks Keparahan Kemiskinan (*Poverty Severity Index* - P_2), yaitu indeks yang memberikan gambaran mengenai penyebaran pengeluaran di antara penduduk miskin.

2.7.4 Kajian tentang Karakteristik Kemiskinan

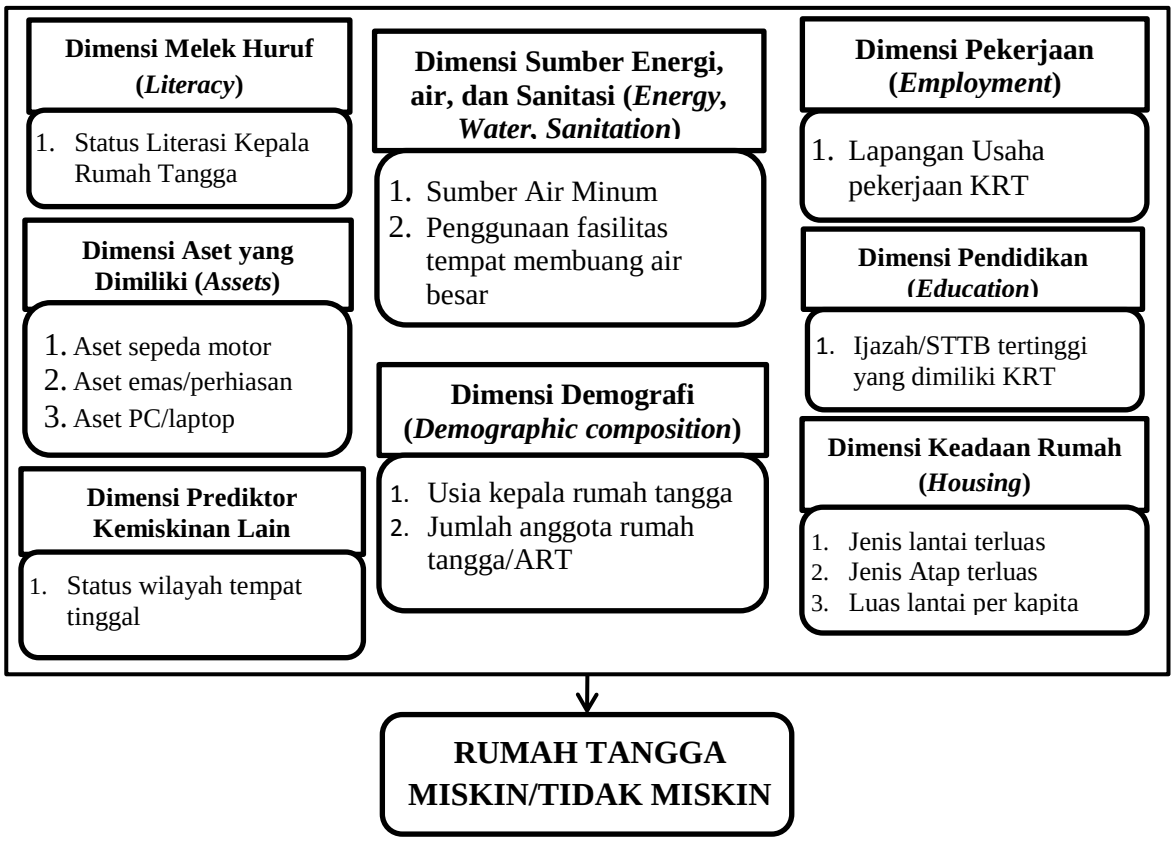
Dalam mengidentifikasi variabel penciri kemiskinan, Mulia (2005) menjabarkan karakteristik rumah tangga miskin berdasarkan penelitiannya di Kabupaten Kupang, Nusa Tenggara Timur. Dengan mengetahui karakteristik rumah tangga miskin, maka akan lebih memudahkan perencanaan program dan target kegiatan pada kelompok sasaran (rumah tangga miskin). Terdapat beberapa faktor yang dapat disimpulkan dalam mengetahui karakteristik rumah tangga miskin diantaranya faktor kependudukan (demografi), karakteristik pendidikan, karakteristik rumah tinggal, karakteristik ekonomi, karakteristik konsumsi, karakteristik ketenagakerjaan dan pendapatan.

Mathiassen dan Roll-Halsen (2007) melakukan penelitian dengan menganalisis rumah tangga berdasarkan model pengeluarannya dan mengidentifikasi karakteristiknya. Terdapat 8 dimensi yang dapat mencirikan kemiskinan pada rumah tangga yaitu dimensi melek huruf (*literacy*), pendidikan (*education*), pekerjaan (e*mployment*), aset yang dimiliki (*assets*), sumber energi,

air dan sanitasi (*energy, water, sanitation*), keadaan rumah (*housing*), kondisi demografi (*demographic composition*), dan dimensi prediktor kemiskinan lain.

Putri (2013) dalam penelitiannya menyimpulkan bahwa faktor umur kepala rumah tangga, faktor pendidikan, jenis pekerjaan mempunyai peran yang signifikan dalam mencirikan rumah tangga miskin di Provinsi Bali.

Dari kajian tentang karakteristik kemiskinan tersebut maka dalam penelitian ini digunakan variabel penelitian yang dituangkan dalam kerangka pikir penelitian sesuai dengan gambar 2.2 berikut.



Gambar 2.2 Kerangka Pikir Penelitian

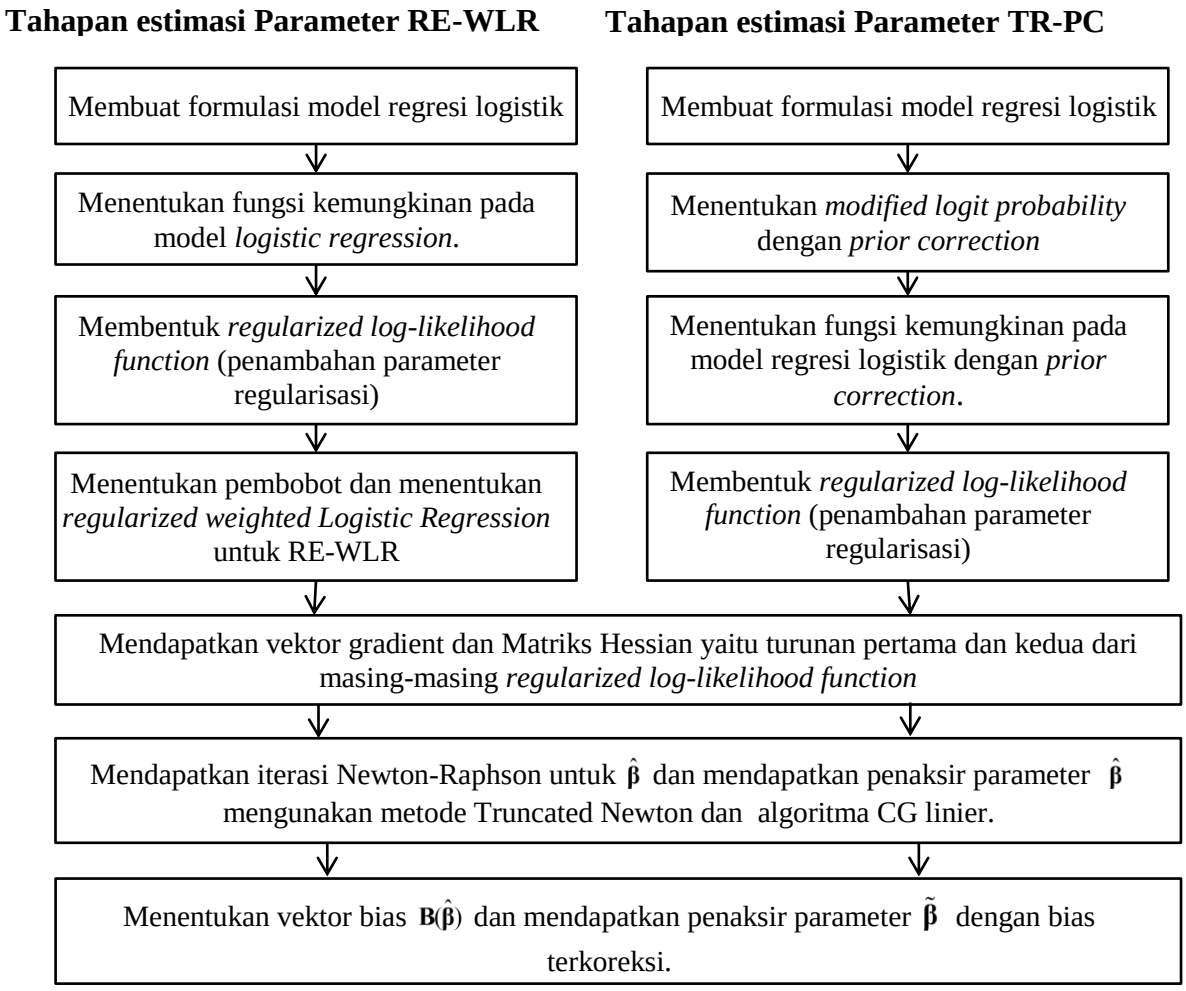
halaman ini sengaja dikosongkan

BAB 3
METODE PENELITIAN

Bab ini terdiri dari 2 sub bab yaitu metodologi untuk mencapai tujuan pertama penelitian yakni kajian estimasi parameter metode RE-WLR dan TR-PC dan metodologi untuk mencapai tujuan kedua dan ketiga penelitian yaitu memodelkan RE-WLR dan TR-PC untuk klasifikasi tingkat kesejahteraan di Provinsi Bali dan membandingkan performa dan ketepatan hasil pengklasifikasian antara metode RE-WLR, TR-PC dan regresi logistik.

3.1 Tahapan Mendapatkan Estimasi Parameter RE-WLR dan TR-PC

Dalam bentuk gambar, tahapan kajian estimasi parameter model RE-WLR dan TR-PC dapat digambarkan seperti berikut:



Gambar 3.1 Flowchart Langkah Estimasi Parameter RE-WLR dan TR-PC

Secara lengkap, langkah-langkah untuk mendapatkan bentuk estimasi parameter model RE-WLR dapat dituliskan sebagai berikut:

1. Formulasikan model regresi logistik sesuai dengan persamaan (2.2).
2. Mendapatkan fungsi kemungkinan (*likelihood function*) pada model *logistic regression* sesuai dengan persamaan (2.6)
3. Membentuk *regularized log-likelihood function* yang diperoleh dengan menentukan *log-likelihood function* pada langkah 2 dan menambahkan parameter regularisasi sesuai dengan persamaan (2.8).
4. Menentukan pembobot untuk RE-WLR.
5. Mendapatkan fungsi kemungkinan *(likelihood function*) pada model *Weighted Logistic Regression* $\left(\ln \mathrm{L}_w(\boldsymbol{\beta})\right)$ untuk RE-WLR sesuai dengan persamaan (2.20).
6. Mendapatkan vektor *gradient* dengan mencari turunan parsial pertama dari *regularized weighted log-likelihood function*.
7. Mendapatkan matriks *Hessian* dengan mencari turunan parsial kedua dari *regularized weighted log-likelihood function*.
8. Mendapatkan iterasi Newton-Raphson untuk $\hat{\boldsymbol{\beta}}$ menggunakan turunan pertama dan turunan kedua dari langkah (6) dan (7).
9. Mendapatkan penaksir parameter $\hat{\boldsymbol{\beta}}$ dengan mengunakan metode Truncated Newton untuk subproblem *WLS* dengan algoritma *CG* linier (lampiran 1).
10. Mendapatkan vektor bias $\mathbf{B}(\hat{\boldsymbol{\beta}})$.
11. Mendapatkan penaksir parameter $\tilde{\boldsymbol{\beta}}$ dengan bias terkoreksi.

Sedangkan tahapan untuk mendapatkan estimasi parameter dari model TR-PC adalah:

1. Formulasikan model regresi logistik sesuai dengan persamaan (2.2).
2. Mendapatkan *modified logit probability* dengan *prior correction* sesuai dengan persamaan (2.37).
3. Menentukan fungsi kemungkinan (*likelihood function*) pada model regresi logistik dengan *prior correction.*

4. Membentuk *regularized log-likelihood function* $\left(\ln \mathrm{L}_{PC}(\boldsymbol{\beta})\right)$ yang diperoleh dengan menentukan *log-likelihood function* pada langkah 1 dan menambahkan parameter regularisasi sesuai dengan persamaan (2.38).
5. Mendapatkan vektor *gradient* dengan mencari turunan parsial pertama dari *regularized log-likelihood function*.
6. Mendapatkan matriks *Hessian* dengan mencari turunan parsial kedua dari *regularized log-likelihood function*.
7. Mendapatkan iterasi Newton-Raphson untuk $\hat{\boldsymbol{\beta}}$ menggunakan turunan pertama dan turunan kedua dari langkah (5) dan (6).
8. Mendapatkan penaksir parameter $\hat{\boldsymbol{\beta}}$ dengan mengunakan metode Truncated Newton untuk subproblem *WLS* dengan menggunakan algoritma *CG* linier (lampiran 2) .
9. Menentukan vektor bias $\mathbf{B}(\hat{\boldsymbol{\beta}})$.
10. Mendapatkan penaksir parameter $\tilde{\boldsymbol{\beta}}$ dengan bias terkoreksi.

3.2 Penerapan Metode RE-WLR dan TR-PC pada Kasus Klasifikasi Tingkat Kesejahteraan di Provinsi Bali

3.2.1 Variabel Penelitian

Variabel-variabel yang digunakan dalam penelitian ini terdiri atas 1 variabel respon (Y) dan 14 variabel prediktor (X) yang dijabarkan sebagai berikut:

Tabel 3.1 Variabel penelitian

Variabel	Nama	Kategori
Y	Status rumah tangga	Kode 0 : Ruta tidak miskin Kode 1 : Ruta miskin
X_1	Status Buta Huruf KRT	Kode 0 : Melek Huruf. Kode 1 : Buta Huruf.
X_2	Usia KRT	Kode 1 : Usia 35 tahun ke bawah. Kode 2 : Usia 36-45 tahun. Kode 3 : Usia 46-55 tahun. Kode 4 : Usia 56 tahun ke atas.

Lanjutan Tabel 3.1 Variabel Penelitian

Variabel	Nama	Kategori
X_3	Jumlah ART	Kode 1 : 2 ART ke bawah. Kode 2 : 3 ART. Kode 3 : 4 ART. Kode 4 : 5 ART ke atas.
X_4	Ijazah/STTB tertinggi yang dimiliki KRT	Kode 1 : Perguruan tinggi. Kode 2 : Berijazah SMA/sederajat. Kode 3 : Berijazah SMP/sederajat. Kode 4 : Berijazah SD/sederajat ke bawah.
X_5	Lapangan usaha utama pekerjaan KRT	Kode 1 : Sektor pertanian. Kode 2 : Sektor non pertanian. Kode 3 : Tidak bekerja atau penerima pendapatan.
X_6	Jenis atap terluas	Kode 1 : Beton atau genteng Kode 2 : Asbes atau seng Kode 3 : Bambu, Kayu/sirap, jerami atau lainnya.
X_7	Jenis lantai terluas	Kode 1 : Marmer/granit, keramik, parket, ubin/tegel/teraso. Kode 2 : Semen/bata merah. Kode 3 : Bambu, kayu/papan kualitas rendah, tanah atau lainnya.
X_8	Luas lantai per kapita	Kode 0 : 10 m^2 ke atas Kode 1 : Kurang dari 10 m^2.
X_9	Penggunaan fasilitas BAB	Kode 1 : Sendiri. Kode 2 : Bersama. Kode 3 : Umum/tidak ada.
X_{10}	Sumber air minum	Kode 1 : Air kemasan bermerk/isi ulang Kode 2 : Ledeng. Kode 3 : Sumber terlindung. Kode 4 : Sumber tak terlindung.
X_{11}	aset PC/laptop	Kode 0 : Memiliki. Kode 1 : Tidak memiliki.
X_{12}	aset emas/ perhiasan	Kode 0 : Memiliki. Kode 1 : Tidak memiliki.
X_{13}	aset sepeda motor	Kode 0 : Memiliki. Kode 1 : Tidak memiliki.
X_{14}	Status wilayah tempat tinggal	Kode 0 : Perkotaan. Kode 1 : Perdesaan.

Definisi operasional dari sebagian variabel penelitian yang digunakan dalam penelitian ini dapat dijelaskan sebagai berikut:

1. Status rumah tangga (Y)
   - Rumah tangga tidak miskin adalah rumah tangga dengan pengeluaran per kapita sebulan di atas garis kemiskinan kabupaten/kota tempat tinggalnya.
   - Rumah tangga miskin adalah rumah tangga dengan pengeluaran per kapita sebulan di bawah garis kemiskinan kabupaten/kota tempat tinggalnya.
2. Status Literasi kepala rumah tangga (X_1)

   Buta huruf adalah tidak dapat membaca dan menulis. Dapat membaca dan menulis adalah dapat membaca dan menulis surat/kalimat sederhana atau dapat membaca dan menulis Huruf Braile. Orang cacat yang pernah bisa membaca dan menulis digolongkan dapat membaca dan menulis.
3. Usia kepala rumah tangga (X_2)

   Usia adalah lama hidup yang telah dijalani oleh seseorang dan dihitung berdasarkan ulang tahun terakhir (pembulatan ke bawah) dalam satuan tahun.
4. Ijazah/STTB tertinggi yang dimiliki kepala rumah tangga (X_4)

   Ijazah/STTB adalah lembaran atau tanda bukti kelulusan yang diberikan kepada seseorang yang sudah menyelesaikan semua persyaratan akademik pada suatu jenjang pendidikan tertentu.
5. Lapangan usaha utama pekerjaan kepala rumah tangga (X_5)

   Lapangan usaha adalah bidang kegiatan dari pekerjaan/usaha/ perusahaan/kantor tempat KRT bekerja.
6. Jenis atap terluas (X_6)

   Atap adalah penutup bagian atas bangunan sehingga yang mendiami di bawahnya terlindung dari terik matahari, hujan dan sebagainya.
7. Jenis lantai terluas (X_7)

   Lantai adalah bagian bawah/dasar/alas suatu ruangan, baik terbuat dari marmer/keramik/granit, tegel/traso, semen, tanah, dan lainnya.
8. Luas lantai per kapita (X_8)

Luas lantai adalah luas lantai yang ditempati dan digunakan untuk keperluan sehari-hari dan untuk bangunan bertingkat adalah jumlah luas dari semua tingkat yang ditempati.

9. Penggunaan fasilitas tempat membuang air besar (X_9)

   Fasilitas tempat buang air besar adalah ketersediaan jamban/kloset yang dapat digunakan oleh rumah tangga responden.

   - Sendiri, jika hanya digunakan oleh rumah tangga responden saja.
   - Bersama, jika digunakan oleh rumah tangga responden bersama dengan beberapa rumah tangga tertentu.
   - Umum/tidak ada, jika dapat digunakan oleh setiap rumah tangga (umum) atau tidak mempunya fasilitas membuang air besar.

10. Sumber air minum (X_{10})

    Jika menggunakan air yang berasal dari beberapa sumber air, maka pilih salah satu sumber air yang volume airnya paling banyak digunakan oleh rumah tangga.

Sedangkan struktur data dalam penelitian ini tersaji dalam Tabel 3.2 berikut ini:

Tabel 3.2 Struktur data penelitian

Unit Observasi (Rumah Tangga)	**Y**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**...**	**X_{14}**
1	y_1	$x_{1.1}$	$x_{2.1}$	$x_{3.1}$	$x_{4.1}$	$x_{5.1}$	...	$x_{14.1}$
2	y_2	$x_{1.2}$	$x_{2.2}$	$x_{3.2}$	$x_{4.2}$	$x_{5.2}$	...	$x_{14.2}$
3	y_3	$x_{1.3}$	$x_{2.3}$	$x_{3.3}$	$x_{4.3}$	$x_{5.3}$	...	$x_{14.3}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.639	y_{5639}	$x_{1.5639}$	$x_{2.5639}$	$x_{3.5639}$	$x_{4.5639}$	$x_{5.5639}$	...	$x_{14.5639}$

3.2.2 Sumber Data

Variabel-variabel yang digunakan untuk klasifikasi tingkat kesejahteraan rumah tangga tersebut didapatkan dari data yang berasal dari hasil pendataan Survei Sosial Ekonomi Nasional (Susenas) yang dilaksanakan oleh Badan Pusat Statistik pada tahun 2015 dengan jumlah sampel sebanyak 5.639 rumah tangga yang mencakup seluruh kabupaten/kota di Provinsi Bali. Unit observasi yang akan

diteliti dalam penelitian ini adalah rumah tangga yang tersebar di 9 kabupaten/kota di Provinsi Bali dengan rincian sebagai berikut:

Tabel 3.3 Jumlah rumah tangga sampel Susenas 2015 Provinsi Bali

Kabupaten	Jumlah Rumah Tangga Sampel	Persentase (%)
1. Jembrana	546	9,7
2. Tabanan	623	11,0
3. Badung	664	11,8
4. Gianyar	638	11,3
5. Klungkung	508	9,0
6. Bangli	550	9,8
7. Karangasem	630	11,2
8. Buleleng	713	12,6
9. Denpasar	767	13,6
TOTAL	5.639	100,0

Susenas adalah survei yang diselenggarakan secara rutin tiap tahun di Indonesia yang bertujuan untuk mengumpulkan data kesejahteraan rakyat di bidang kesehatan, pendidikan, fertilitas, keluarga berencana, perumahan, dan kondisi sosial ekonomi lainnya. Selain itu Susenas mengumpulkan data konsumsi dan pengeluaran yang menjadi dasar bagi penghitungan angka kemiskinan.

3.2.3 Penyiapan Data *Training* dan Data *Testing*

Pada kasus *imbalanced data*, sangat penting untuk mendapatkan model yang dapat memprediksi kelas minoritas dengan baik yang diukur dari nilai sensitivitas yang tinggi. Untuk itu perlu digunakan teknik sampling khusus dalam penyiapan data *training* agar model yang didapatkan dapat memprediksi dengan tepat kelas minoritas pada data *testing*. Teknik yang digunakan dalam penelitian ini adalah dengan menggunakan *choice-based sampling* yang merupakan salah satu bentuk sampling terstratifikasi. Pada teknik ini, data distratifikasi berdasarkan kelompok pada variabel respon, kemudian sampel (data *training*) diambil dari tiap strata dengan proporsi yang telah ditentukan sehingga kelas minoritas lebih terwakili pada sampel.

Data *training* yang digunakan pada penelitian ini berjumlah lima set data dengan masing-masing set data *training* terdiri dari perbandingan kelompok yang

ditentukan. Dari 298 rumah tangga miskin, dipilih 150 rumah tangga secara acak menjadi data *training* sedangkan sisanya 148 rumah tangga sebagai data *testing*. Satu set data *testing* diambil dari 25 persen dari total observasi (1.483 rumah tangga) dengan proporsi rumah tangga miskin ditetapkan sebanding dengan proporsi kemiskinan total (5,2 persen) yang diambil secara random. Data rumah tangga yang digunakan sebagai data *testing* merupakan rumah tangga yang berbeda dengan yang digunakan sebagai data *training* (tidak *overlap*). Untuk melihat konsistensi dari model yang dihasilkan maka set data *testing* yang digunakan berjumlah 10 set dengan masing-masing set diambil dari data observasi secara random dan proporsi rumah tangga miskin sebesar 5,2 persen.

Tabel 3.4 Skema data *training* dan data *testing*

Data Training				**Data Testing (Per Set Data)**		
Skema (Miskin:Tidak Miskin)	Ruta Miskin	Ruta Tidak Miskin	Total Data Training	Ruta Miskin	Ruta Tidak Miskin	Total
1:3	150	450	600	74	1.409	1.483
1:4	150	600	750	74	1.409	1.483
1:5	150	750	900	74	1.409	1.483
1:6	150	900	1.050	74	1.409	1.483
1:10	150	1.500	1.650	74	1.409	1.483

3.2.4 Tahapan Pemodelan RE-WLR dan TR-PC untuk klasifikasi Tingkat Kesejahteraan di Provinsi Bali

Tujuan kedua dalam penelitian ini adalah memodelkan RE-WLR dan TR-PC untuk klasifikasi tingkat kesejahteraan rumah tangga di Provinsi Bali. Adapun tahapan penelitian untuk mencapai tujuan tersebut adalah sebagai berikut:

1. Melakukan analisis deskriptif mengenai karakteristik rumah tangga di Bali dihubungkan dengan menggunakan variabel respon dan variabel prediktor.
2. Mendapatkan model RE-WLR terbaik untuk klasifikasi tingkat kesejahteraan di Provinsi Bali dengan tahapan sebagai berikut:
   a. Mendapatkan penaksir parameter model RE-WLR dengan tahapan sebagai berikut:

1. Hitung probabiliti $\hat{p}_i$ sesuai dengan persamaan (2.1).
2. Hitung nilai varians v_i sesuai dengan persamaan (2.30).
3. Tentukan nilai pembobot w_i sesuai dengan persamaan (2.21).
4. Hitung *Adjusted Response* sesuai dengan persamaan (2.26).
5. Hitung matriks kovarians (Q) sesuai dengan persamaan (2.29).
6. Tentukan nilai Q_{ii} .
7. Hitung *bias response* sesuai dengan persamaan (2.33).
8. Tentukan matriks pembobot D sesuai dengan persamaan (2.34)
9. Hitung estimator untuk $\hat{\beta}$ dengan algoritma 2.
10. Hitung $bias(\hat{\beta})$ dengan persamaan (2.32) dengan algoritma 3.
11. Hitung estimator terkoreksi $\tilde{\beta}$ sesuai dengan persamaan (2.35).

b. Hitung probabilita optimal $(\tilde{p}_i)$ dari masing-masing observasi sesuai dengan persamaan (2.36).

c. Mendapatkan *confusion matrix* dari data training dan data testing sehingga dapat menghitung nilai *sensitivity* (persamaan (2.52)), *specificity* (persamaan (2.53)), G-mean (persamaan (2.54)) dan nilai AUC dari model RE-WLR.

d. Memilih skema data training terbaik dengan berdasar nilai AUC dan G-mean tertinggi.

e. Memilih parameter regularisasi (λ) terbaik dengan berdasar nilai AUC dan G-mean tertinggi.

f. Menetapkan model RE-WLR terbaik dengan menggunakan nilai AUC tertinggi.

g. Menginterpretasikan model RE-WLR terbaik dan membuat kesimpulan dari model tersebut.

3. Mendapatkan model TR-PC terbaik untuk klasifikasi tingkat kesejahteraan di Provinsi Bali dengan tahapan sebagai berikut:

a. Menentukan penaksir parameter model TR-PC dengan tahapan sebagai berikut:

1. Hitung probabiliti $\tilde{p}_i$ sesuai dengan persamaan (2.37).

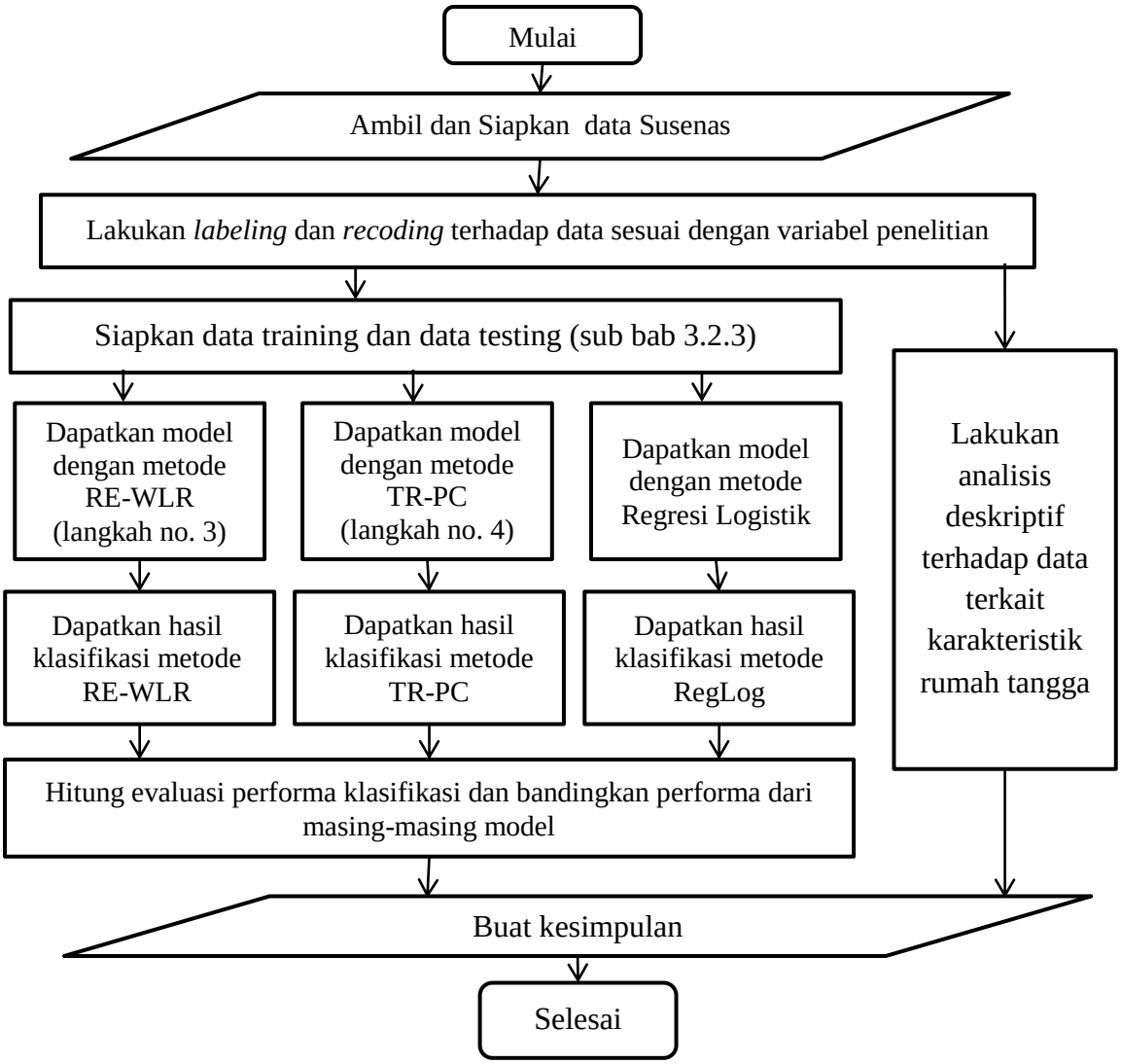
2. Hitung nilai varians v_i sesuai dengan persamaan (2.42).
3. Hitung *Adjusted Response* sesuai dengan persamaan (2.44).
4. Hitung matriks kovarians (Q) sesuai dengan persamaan (2.48).
5. Tentukan nilai Q_{ii} .
6. Hitung *bias response* sesuai dengan persamaan (2.47).
7. Tentukan matriks pembobot D sesuai dengan persamaan (2.41).
8. Hitung estimator untuk $\hat{\beta}$ dengan algoritma 2.
9. Hitung $bias(\hat{\beta})$ dengan persamaan (2.46) dengan algoritma 3.
10. Hitung estimator terkoreksi $\tilde{\beta}$ sesuai dengan persamaan (2.49).

b. Hitung probabilita optimal $(\tilde{p}_i)$ dari masing-masing observasi sesuai dengan persamaan (2.50).
c. Mendapatkan *confusion matrix* dari data training dan data testing sehingga dapat menghitung nilai *sensitivity* (persamaan (2.52)), *specificity* (persamaan (2.53)), G-mean (persamaan (2.54)) dan nilai AUC dari model TR-PC.
d. Memilih skema data training terbaik dengan berdasar nilai AUC dan G-mean.
e. Memilih parameter regularisasi (λ) terbaik dengan berdasar nilai AUC dan G-mean.
f. Menetapkan model TR-PC terbaik dengan menggunakan nilai AUC tertinggi.
g. Menginterpretasikan model TR-PC terbaik dan membuat kesimpulan dari model tersebut.

Setelah tujuan kedua penelitian tercapai yakni mendapatkan model RE-WLR dan TR-PC terbaik pada kasus data tingkat kesejahteraan rumah tangga di Provinsi Bali, maka untuk menjawab tujuan ketiga yakni membandingkan performa dari metode RE-WLR, TR-PC dan regresi logistik maka langkah-langkahnya adalah sebagai berikut:

a. Bandingkan performa klasifikasi kedua model tersebut dan bandingkan juga dengan hasil klasifikasi dari metode regresi logistik sebagai *benchmark* dengan menggunakan ukuran nilai AUC dan G-mean.
b. Membuat kesimpulan dari hasil analisis tersebut

Adapun langkah-langkah secara umum yang dilakukan untuk menjawab tujuan penelitian dapat ditunjukkan oleh Gambar 3.2 berikut ini.



Gambar 3.2 Flowchart Langkah Pemodelan RE-WLR dan TR-PC

halaman ini sengaja dikosongkan

BAB 4
HASIL DAN PEMBAHASAN

Pada bab ini terdiri dari tiga sub bab, dimana sub bab pertama akan membahas mengenai kajian estimasi parameter RE-WLR dan TR-PC, sub bab kedua membahas tentang penerapan RE-WLR dan TR-PC pada klasifikasi *imbalanced data* dengan studi kasus klasifikasi tingkat kesejahteraan rumah tangga di Provinsi Bali, dan sub bab ketiga membahas tentang perbandingan ketepatan klasifikasi antara metode RE-WLR dan TR-PC.

4.1 Kajian Estimasi Parameter RE-WLR dan TR-PC

Estimasi parameter untuk metode RE-WLR dan TR-PC dalam penelitian ini menggunakan metode *Maximum Likelihood Estimation* (MLE). Metode MLE adalah suatu metode untuk mendapatkan estimasi parameter dengan cara memaksimumkan fungsi *likelihood*. Model RE-WLR dan TR-PC merupakan pengembangan dari model regresi logistik sederhana dimana dalam regresi logistik, variabel respon y mengikuti distribusi Bernoulli dengan fungsi sebaran peluangnya adalah:

$$P\left(x_i, y_i \mid \beta\right) = \begin{cases} p_i & ,\text{untuk } y = 1 \\ 1 - p_i & ,\text{untuk } y = 0 \end{cases}$$

$$P\left(\mathbf{x}_i, y_i \mid \boldsymbol{\beta}\right) = \left(p_i\right)^{y_i} \left(1 - p_i\right)^{(1-y_i)}$$

4.1.1 Kajian Estimasi Parameter untuk metode RE-WLR

Metode RE-WLR ini menerapkan regularisasi, *weighting*, *bias correction* dan optimasi *Conjugate Gradient* pada regresi logistik. Langkah-langkah untuk mendapatkan bentuk estimasi parameter model RE-WLR adalah:

1. Menentukan fungsi kemungkinan (*likelihood function*) pada model *logistic regression.*

   Bentuk fungsi likelihood pada regresi logistik untuk n sampel independen adalah

   $$L\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{n} \left(p_i\right)^{y_i} \left(1 - p_i\right)^{(1-y_i)}$$, karena $p_i = \dfrac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$ maka

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right)^{y_i} \left(1 - \frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right)^{(1-y_i)} = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right)^{(1-y_i)}$$

Untuk memudahkan dalam mencari estimasi parameter yang dapat memaksimumkan fungsi likelihood maka lebih mudah menggunakan bentuk logaritma natural dari fungsi likelihood yang disebut fungsi log likelihood.

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \ln \left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right).$$

2. Membentuk *regularized log-likelihood function* yang diperoleh dengan menentukan *log-likelihood function* pada langkah 1 dan memberikan unsur parameter regularisasi ($\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$).

   Fungsi Regularized log likelihood yang terbentuk adalah

$$\begin{aligned}
\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i \ln \left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} \left(y_i \left(\ln \left(\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) - \ln \left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right) + \ln \left(\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} \left(y_i \left(\ln \left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) - \ln \left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) - \ln 1 + \ln \left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) \right) + \ln 1 - \ln \left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} \left(y_i \left(\ln \left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) - 0 \right) + 0 - \ln \left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} \left(\ln \left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right)^{y_i} - \ln \left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \right) \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \\
&= \sum_{i=1}^{n} \ln \left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}} \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2
\end{aligned}$$

3. Menentukan pembobot untuk RE-WLR.

   King dan Zheng merekomendasikan dua metode yaitu *weighting* dan *prior correction*. Dalam metode RE-WLR yang digunakan adalah *weighting* (pembobot). *Joint distribution* dari **y** dan **X** dalam sampel adalah

   $$f_s(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\beta}) = P_s(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\beta}) P_s(\mathbf{y})$$

dimana $\boldsymbol{\beta}$ adalah parameter yang tidak diketahui dan akan diestimasi. $\mathbf{X}$ adalah matriks variabel prediktor, maka probabilita bersyarat $\mathbf{X}$ dalam sampel adalah sama dengan probabilita bersyarat $\mathbf{X}$ dalam populasi atau $P_s(\mathbf{X}\,|\,\mathbf{y},\boldsymbol{\beta}) = P(\mathbf{X}\,|\,\mathbf{y},\boldsymbol{\beta})$. Probabilita bersyarat dalam populasi yaitu

$$P(\mathbf{X}\,|\,\mathbf{y},\boldsymbol{\beta}) = \frac{f(\mathbf{y},\mathbf{X}\,|\,\boldsymbol{\beta})}{P(\mathbf{y})},$$

dan $f(\mathbf{y},\mathbf{X}\,|\,\boldsymbol{\beta}) = P(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta})P(\mathbf{X})$. Sehingga

$$\begin{aligned} f(\mathbf{y},\mathbf{X}\,|\,\boldsymbol{\beta}) &= \frac{P_s(\mathbf{y})}{P(\mathbf{y})} P(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta}) P(\mathbf{X}) \\ &= \frac{H}{Q} P(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta}) P(\mathbf{X}) \end{aligned}$$

dimana $\frac{H}{Q} = \frac{P_s(\mathbf{y})}{P(\mathbf{y})}$ dengan H adalah proporsi y dalam sampel dan Q adalah proporsi y dalam populasi. Selanjutnya fungsi likelihood yaitu

$$L_{Endogenous} = \prod_{i=1}^{n} \frac{H_i}{Q_i} P(y_i\,|\,\mathbf{x}_i,\boldsymbol{\beta}) P(\mathbf{x}_i),$$

dimana $\frac{H_i}{Q_i} = \left(\frac{\tau}{\bar{y}}\right) y_i + \left(\frac{1-\tau}{1-\bar{y}}\right)(1-y_i),$ dengan $\bar{y}$ adalah proporsi kejadian dalam sampel dan τ adalah proporsi kejadian dalam populasi.

Fungsi Log Likelihoodnya adalah

$$\begin{aligned} \ln L(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X}) &= \sum_{i=1}^{n} \frac{H_i}{Q_i} P(y_i\,|\,\mathbf{x}_i,\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \frac{H_i}{Q_i} \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}\right) \\ &= \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}\right), \end{aligned}$$

dimana $w_i = \frac{H_i}{Q_i}$. Maka pembobot $w_i = \left(\frac{\tau}{\bar{y}}\right) y_i + \left(\frac{1-\tau}{1-\bar{y}}\right)(1-y_i),$ sehingga jika $y_i = 1$ maka $w_i = \frac{\tau}{\bar{y}} = w_1$ dan jika $y_i = 0$ maka $w_i = \frac{1-\tau}{1-\bar{y}} = w_0.$

4. Menentukan fungsi kemungkinan *(likelihood function*) pada model *Weighted Logistic Regression* untuk RE-WLR

$$\ln L_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) - \frac{\lambda}{2}\left(\sqrt{\beta_0^2+\beta_1^2+\beta_2^2+\ldots+\beta_d^2}\right)^2$$

$$= \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) - \frac{\lambda}{2}\left(\beta_0^2+\beta_1^2+\beta_2^2+\ldots+\beta_d^2\right).$$

5. Mendapatkan vektor *gradient* dengan mencari turunan parsial pertama dari *regularized weighted log-likelihood function*.

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial}{\partial\beta_0}\ln L_w(\boldsymbol{\beta}) \\ \frac{\partial}{\partial\beta_1}\ln L_w(\boldsymbol{\beta}) \\ \vdots \\ \frac{\partial}{\partial\beta_j}\ln L_w(\boldsymbol{\beta}) \\ \vdots \\ \frac{\partial}{\partial\beta_d}\ln L_w(\boldsymbol{\beta}) \end{bmatrix}$$

- $\frac{\partial}{\partial\beta_0}\ln L_w(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_0}\left(\sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right) - \frac{\lambda}{2}\left(\beta_0^2+\beta_1^2+\beta_2^2+\ldots+\beta_d^2\right)\right)$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right)\left(\frac{y_i e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right) - e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^2}\right)\right) - \frac{\lambda}{2}.2\beta_o$$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}\right)\left(\frac{e^{y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(y_i\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right) - \left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^2}\right)\right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} w_i \left(\frac{y_i\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right) - \left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}\right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} w_i \left(y_i - \frac{\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}\right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} w_i \left(y_i - p_i \right) - \lambda \beta_o$$

- $\dfrac{\partial}{\partial \beta_1} \ln L_w(\boldsymbol{\beta}) = \dfrac{\partial}{\partial \beta_1} \left(\sum_{i=1}^{n} w_i \ln \left(\dfrac{e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}} \right) - \dfrac{\lambda}{2} \left(\beta_0^2 + \beta_1^2 + \beta_2^2 + \ldots + \beta_d^2 \right) \right)$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}{e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}} \right) \left(\frac{y_i x_{i1} e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right) - e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \left(x_{i1} e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)}{\left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)^2} \right) \right) - \frac{\lambda}{2} . 2 \beta_1$$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}}{e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}} \right) \left(\frac{x_{i1} e^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \left(y_i \left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right) - \left(e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right) \right)}{\left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)^2} \right) \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} w_i \left(x_{i1} \left(\frac{y_i \left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right) - \left(e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)}{\left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)} \right) \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} w_i x_{i1} \left(y_i - \frac{\left(e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)}{\left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right)} \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} w_i x_{i1} \left(y_i - p_i \right) - \lambda \beta_1$$

$\vdots$

- $\dfrac{\partial}{\partial \beta_j} \ln L_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i x_{ij} \left(y_i - p_i \right) - \lambda \beta_j$

$\vdots$

- $\dfrac{\partial}{\partial \beta_d} \ln L_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i x_{id} \left(y_i - p_i \right) - \lambda \beta_d$

Sehingga apabila dituliskan dalam bentuk matriks, vektor gradien yang terbentuk adalah

$\mathbf{g}(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}} \mathbf{W} (\mathbf{y} - \mathbf{p}) - \lambda \boldsymbol{\beta}$,

dengan $\mathbf{W} = diag\left(w_1, w_2, \cdots, w_n \right)$ dan $\mathbf{p}$ adalah vektor probabilitas.

6. Mendapatkan matriks *Hessian* dengan mencari turunan parsial kedua dari *regularized weighted log-likelihood function*.

$$\mathbf{H}(\boldsymbol{\beta})=\begin{pmatrix} \frac{\partial^2}{\partial\beta_0\partial\beta_0}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_0\partial\beta_j}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_0\partial\beta_d}\ln L_w(\boldsymbol{\beta}) \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial^2}{\partial\beta_j\partial\beta_0}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_j\partial\beta_j}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_j\partial\beta_d}\ln L_w(\boldsymbol{\beta}) \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\beta_d\partial\beta_0}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_d\partial\beta_j}\ln L_w(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_d\partial\beta_d}\ln L_w(\boldsymbol{\beta}) \end{pmatrix}$$

- $$\frac{\partial^2}{\partial\beta_j\partial\beta_j}\ln L_w(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\frac{\partial}{\partial\beta_j}\ln L_w(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}w_i x_{ij}\left(y_i-\frac{\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}\right)-\lambda\beta_j\right)$$

$$=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}w_i x_{ij}y_i-w_i x_{ij}\frac{\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}-\lambda\beta_j\right)$$

$$=\sum_{i=1}^{n}-w_i x_{ij}\frac{x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)-x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^2}-\lambda$$

$$=\sum_{i=1}^{n}-w_i x_{ij}\frac{x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}-e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^2}-\lambda$$

$$=\sum_{i=1}^{n}-w_i x_{ij}x_{ij}\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}\frac{1}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}-\lambda$$

$$=\sum_{i=1}^{n}-w_i x_{ij}x_{ij}\left(p_i\left(1-p_i\right)\right)-\lambda$$

- $$\frac{\partial^2}{\partial\beta_j\partial\beta_k}\ln L_w(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\frac{\partial}{\partial\beta_d}\ln L_w(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}w_i x_{id}\left(y_i-\frac{\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}\right)-\lambda\beta_d\right)$$

$$=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}w_i x_{id}y_i-w_i x_{id}\frac{\left(e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)}-\lambda\beta_d\right)$$

$$=\sum_{i=1}^{n}-w_i x_{id}\frac{x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)-x_{ij}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}}{\left(1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)^2}-0$$

$$= \sum_{i=1}^{n} -w_i x_{id} \frac{x_{ij} e^{\mathbf{x}_i^T \boldsymbol{\beta}} \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}} - e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)}{\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)^2} - 0$$

$$= \sum_{i=1}^{n} -w_i x_{id} x_{ij} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)} \frac{1}{\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)} - 0$$

$$= \sum_{i=1}^{n} -w_i x_{id} x_{ij} \left(p_i \left(1 - p_i\right)\right) - 0$$

Sehingga bila dituliskan dalam bentuk matriks maka bentuknya adalah

$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} - \lambda\mathbf{I}$

dimana $\mathbf{D} = diag\left(w_i p_i \left(1 - p_i\right)\right)$

7. Mendapatkan iterasi Newton-Raphson untuk $\hat{\boldsymbol{\beta}}$ menggunakan vektor gradien dan matriks Hessian dari langkah (5) dan (6).

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \hat{\boldsymbol{\beta}}^{(c)} - \left(\mathbf{H}(\boldsymbol{\beta})\right)^{-1}\left(\mathbf{g}(\boldsymbol{\beta})\right)$$
$$= \hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$

Karena $\hat{\boldsymbol{\beta}}^{(c)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c)}$ maka

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$
$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)}\right) + \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$
$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)^{-1}\lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right) - \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)^{-1}\lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)}\right]$$
$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right)\right], \text{ karena } \mathbf{D} = \mathbf{W}\mathbf{V}, \text{maka}$$
$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{V}^{-1}\mathbf{W}^{-1}\mathbf{W}\left(\mathbf{y} - \mathbf{p}\right)\right]$$
$$= \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{p}\right)\right]$$

Sehingga update Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) yaitu

$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{z}^{(c)}$, dimana $\mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{p}\right)$.

8. Mendapatkan penaksir parameter $\hat{\boldsymbol{\beta}}$ dengan mengunakan metode Truncated Newton untuk subproblem *WLS* dengan algoritma *CG* linier.

Dalam update Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) terdapat problem numerik yaitu mendapatkan invers matriks Hessian berukuran besar sehingga pemrosesan menjadi sangat lama. Sehingga untuk mengatasinya adalah menggunakan Truncated Newton dengan CG linier sebagai algoritma truncated inner. Iterasi Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) adalah

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X^T D X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X^T D z}^{(c)}$$
$$\left(\mathbf{X^T D X} + \lambda \mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X^T D z}^{(c)}$$

Tahapan untuk memperoleh estimasi parameter dengan CG linier adalah sebagai berikut:

a. Langkah awal hitung $\mathbf{A} = \mathbf{X^T D X} + \lambda \mathbf{I}$, $\mathbf{r}^{(0)} = \mathbf{X^T D z} - \left(\mathbf{X^T D X} + \lambda \mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(0)}$ dan $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$, mulai iterasi dengan t=0

b. Hitung

   Panjang langkah optimal $s^{(t)} = \dfrac{\mathbf{r}^{(t)T}\mathbf{r}^{(t)}}{\mathbf{d}^{(t)T}\mathbf{A}\mathbf{d}^{(t)}}$

   Hitung perkiraan solusi $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + s^{(t)}\mathbf{d}^{(t)}$

   Update *residual* $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - s^{(t)}\mathbf{A}\mathbf{d}^{(t)}$

   Update A-*conjugacy* $a^{(t)} = \dfrac{\mathbf{r}^{(t+1)T}\mathbf{r}^{(t+1)}}{\mathbf{r}^{(t)T}\mathbf{r}^{(t)}}$

   Update arah pencarian $\mathbf{d}^{(t+1)} = \mathbf{r}^{(t+1)} + a^{(t)}\mathbf{d}^{(t)}$

c. Iterasi dihentikan jika $\left\|\mathbf{r}^{(t)}\right\|^2 < \varepsilon$ dimana ε adalah bilangan kecil mendekati nol (pada penelitian ini ε sama dengan 0,005).

9. Menentukan vektor bias $\mathbf{B}(\hat{\boldsymbol{\beta}})$.

   Vektor bias untuk WLR adalah $\mathbf{B}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X^T D X}\right)^{-1}\mathbf{X^T D}\boldsymbol{\xi}$.

10. Mendapatkan penaksir parameter $\tilde{\boldsymbol{\beta}}$ dengan bias terkoreksi.

    $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$.

4.1.2 Kajian Estimasi Parameter untuk metode TR-PC

Metode TR-PC ini menerapkan regularisasi, *prior correction*, *bias correction* dan optimasi *Conjugate Gradient* pada regresi logistik. Langkah-langkah untuk mendapatkan bentuk estimasi parameter model TR-PC adalah:

1. Menentukan *modified logit probability* dengan *prior correction*

   King dan Zeng (2000) memodifikasi *logit probability* dari regresi logistik dengan menggunakan prinsip Bayesian. Misalkan keseluruhan data dengan N observasi dengan τ adalah proporsi terjadinya *event*/kejadian dan $1-\tau$ adalah proporsi terjadinya *non-event*. Dalam data *training* (sampel) dengan n observasi dengan $\bar{y}$ adalah proporsi terjadinya *event* dan $1-\bar{y}$ adalah proporsi terjadinya *non-event*. Dengan $\hat{p}$ adalah probabiliti terjadinya event pada sampel maka bayesian formula (Palepu, 1986; Cramer, 2005) adalah

$$\hat{p} = P\left(y=1 \middle| s=1\right) = \frac{P\left(s=1 \middle| y=1\right)P\left(y=1\right)}{P\left(s=1 \middle| y=1\right)P\left(y=1\right) + P\left(s=1 \middle| y=0\right)P\left(y=0\right)}$$

$$= \frac{\left(\frac{\bar{y}}{\tau}\right)\tilde{p}}{\left(\frac{\bar{y}}{\tau}\right)\tilde{p} + \left(\frac{1-\bar{y}}{1-\tau}\right)\left(1-\tilde{p}\right)}$$

$$= \frac{\delta_1 \tilde{p}}{\delta_1 \tilde{p} + \delta_0\left(1-\tilde{p}\right)}$$

   Sehingga fungsi Oddnya adalah

$$Odd = \frac{\hat{p}}{1-\hat{p}} = \left(\frac{\delta_1 \tilde{p}}{\delta_1 \tilde{p} + \delta_0\left(1-\tilde{p}\right)}\right) \Bigg/ \left(1 - \frac{\delta_1 \tilde{p}}{\delta_1 \tilde{p} + \delta_0\left(1-\tilde{p}\right)}\right)$$

$$= \frac{\delta_1 \tilde{p}}{\delta_0\left(1-\tilde{p}\right)} = \frac{\delta_1}{\delta_0}\frac{\tilde{p}}{\left(1-\tilde{p}\right)}$$

   log fungsi Odd adalah

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \ln\left(\frac{\delta_1}{\delta_0}\frac{\tilde{p}}{\left(1-\tilde{p}\right)}\right)$$

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \ln\left(\frac{\delta_1}{\delta_0}\right) + \ln\left(\frac{\tilde{p}}{\left(1-\tilde{p}\right)}\right)$$

$$\mathbf{x}\hat{\boldsymbol{\beta}} = \ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}\tilde{\boldsymbol{\beta}}$$

$$\mathbf{x}\tilde{\boldsymbol{\beta}} = \mathbf{x}\hat{\boldsymbol{\beta}} - \ln\left(\frac{\delta_1}{\delta_0}\right)$$

sehingga modified logit probability menjadi

$$\tilde{p}_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}$$

2. Menentukan fungsi kemungkinan (*likelihood function*) pada model regresi logistik dengan *prior correction.*

   Bentuk fungsi likelihood pada regresi logistik untuk n sampel independen adalah

   $L_{PC}(\boldsymbol{\beta}) = \prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{(1-y_i)}$, karena $p_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}$ maka

   $$L_{PC}(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{(1-y_i)}$$

   $$= \prod_{i=1}^{n}\left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)^{(1-y_i)}$$

   Untuk memudahkan dalam mencari estimasi parameter yang dapat memaksimumkan fungsi likelihood maka lebih mudah menggunakan bentuk logaritma natural dari fungsi likelihood yang disebut fungsi log likelihood.

   $$\ln L_{PC}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i \ln\left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right) + (1-y_i)\ln\left(\frac{e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}\right)\right).$$

3. Membentuk *regularized log-likelihood function* yang diperoleh dengan menentukan *log-likelihood function* pada langkah 2 dan memberikan unsur parameter regularisasi.

   Fungsi Regularized log likelihood yang terbentuk adalah

$$\ln L_{PC}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i \ln\left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}} \right) + (1-y_i)\ln\left(\frac{e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}} \right)\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\left(y_i\left(\ln\left(\frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}} \right) - \ln\left(\frac{e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}}} \right)\right) + \ln\left(\frac{1}{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}} \right)\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\left(y_i\left(\ln(1) - \ln\left(1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}} \right) - \ln\left(e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}} \right) + \ln\left(1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}} \right)\right) + \ln(1) - \ln\left(1+e^{-1\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}} \right)\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\left(y_i\left(-\ln\left(e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i^T\boldsymbol{\beta}} \right)\right) - \ln\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}} \right)\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\left(\ln\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}} \right)^{y_i} - \ln\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}} \right)\right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}\ln\left(\frac{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}}{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}} \right) - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$$

4. Mendapatkan vektor *gradient* dengan mencari turunan parsial pertama dari *regularized log-likelihood function*.

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial}{\partial\beta_0}\ln L_{PC}(\boldsymbol{\beta}) \\ \frac{\partial}{\partial\beta_1}\ln L_{PC}(\boldsymbol{\beta}) \\ \vdots \\ \frac{\partial}{\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta}) \\ \vdots \\ \frac{\partial}{\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) \end{bmatrix}$$

- $\frac{\partial}{\partial\beta_0}\ln L_{PC}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_0}\left(\sum_{i=1}^{n}\ln\left(\frac{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}}{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}} \right) - \frac{\lambda}{2}\left(\beta_0^2+\beta_1^2+\beta_2^2+\ldots+\beta_k^2\right)\right)$

$$= \sum_{i=1}^{n} \left(\left(\frac{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}} \right) \left(\frac{y_i e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right) - e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2} \right) \right) - \frac{\lambda}{2}.2\beta_o$$

$$= \sum_{i=1}^{n} \left(\left(\frac{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}} \right) \left(\frac{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}\left(y_i\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right) - \left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2} \right) \right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} \left(\frac{y_i\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right) - \left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)} \right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} \left(y_i - \frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)} \right) - \lambda\beta_o$$

$$= \sum_{i=1}^{n} \left(y_i - \tilde{p}_i \right) - \lambda\beta_o$$

- $$\frac{\partial}{\partial\beta_1} \ln L_{PC}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_1} \left(\sum_{i=1}^{n} \ln \left(\frac{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}}{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}} \right) - \frac{\lambda}{2}\left(\beta_0^2+\beta_1^2+\beta_2^2+\ldots+\beta_k^2\right) \right)$$

$$= \sum_{i=1}^{n} \left(\left(\frac{1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}} \right) \left(\frac{y_i x_{i1} e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right) - e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}\right)}\left(x_{i1}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2} \right) \right) - \frac{\lambda}{2}.2\beta_1$$

$$= \sum_{i=1}^{n} \left(\left(\frac{1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}}}{e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}\right)}} \right) \left(\frac{x_{i1} e^{y_i\left(-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}\right)} \left(y_i \left(1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right) - \left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right) \right)}{\left(1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right)^2} \right) \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} \left(x_{i1} \left(\frac{y_i \left(1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right) - \left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right)}{\left(1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right)} \right) \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} x_{i1} \left(y_i - \frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right)}{\left(1 + e^{-\ln\left(\frac{\delta_1}{\delta_0}\right) + \mathbf{x}_i^T \boldsymbol{\beta}} \right)} \right) - \lambda \beta_1$$

$$= \sum_{i=1}^{n} x_{i1} \left(y_i - \tilde{p}_i \right) - \lambda \beta_1$$

$$\vdots$$

- $\dfrac{\partial}{\partial \beta_j} \ln L_{PC}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij} \left(y_i - \tilde{p}_i \right) - \lambda \beta_j$

$$\vdots$$

- $\dfrac{\partial}{\partial \beta_k} \ln L_{PC}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ik} \left(y_i - \tilde{p}_i \right) - \lambda \beta_k$

Sehingga apabila dituliskan dalam bentuk matriks, vektor gradien yang terbentuk adalah

$\mathbf{g}(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}} (\mathbf{y} - \tilde{\mathbf{p}}) - \lambda \boldsymbol{\beta}$,

dengan $\tilde{\mathbf{p}}$ adalah vektor probabilitas.

5. Mendapatkan matriks *Hessian* dengan mencari turunan parsial kedua dari *regularized log-likelihood function*.

$$\mathbf{H}(\boldsymbol{\beta})=\begin{pmatrix} \frac{\partial^2}{\partial\beta_0\partial\beta_0}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_0\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_0\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial^2}{\partial\beta_j\partial\beta_0}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_j\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_j\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\beta_k\partial\beta_0}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_k\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial\beta_k\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) \end{pmatrix}$$

- $$\frac{\partial^2}{\partial\beta_j\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\frac{\partial}{\partial\beta_j}\ln L_{PC}(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}x_{ij}\left(y_i-\frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}\right)-\lambda\beta_j\right)$$

$$=\frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n}x_{ij}y_i-x_{ij}\frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}-\lambda\beta_j\right)$$

$$=\sum_{i=1}^{n}-x_{ij}\frac{x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)-x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2}-\lambda$$

$$=\sum_{i=1}^{n}-x_{ij}\frac{x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}-e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2}-\lambda$$

$$=\sum_{i=1}^{n}-x_{ij}x_{ij}\frac{e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}\frac{1}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}-\lambda$$

$$=\sum_{i=1}^{n}-x_{ij}x_{ij}\left(\tilde{p}_i\left(1-\tilde{p}_i\right)\right)-\lambda$$

- $$\frac{\partial^2}{\partial\beta_j\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_j}\frac{\partial}{\partial\beta_k}\ln L_{PC}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n} x_{ik}\left(y_i - \frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}\right) - \lambda\beta_k\right)$$

$$= \frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^{n} x_{ik}y_i - x_{ik}\frac{\left(e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)} - \lambda\beta_k\right)$$

$$= \sum_{i=1}^{n} -x_{ik}\frac{x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right) - x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2} - 0$$

$$= \sum_{i=1}^{n} -x_{ik}\frac{x_{ij}e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}} - e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)^2} - 0$$

$$= \sum_{i=1}^{n} -x_{ik}x_{ij}\frac{e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)}\frac{1}{\left(1+e^{-\ln\left(\frac{\delta_1}{\delta_0}\right)+\mathbf{x}_i^T\boldsymbol{\beta}}\right)} - 0$$

$$= \sum_{i=1}^{n} -x_{ik}x_{ij}\left(\tilde{p}_i\left(1-\tilde{p}_i\right)\right) - 0$$

  Sehingga bila dituliskan dalam bentuk matriks maka bentuknya adalah

  $$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X} - \lambda\mathbf{I}$$

  dimana $\mathbf{D} = diag\left(\tilde{p}_i\left(1-\tilde{p}_i\right)\right)$

6. Mendapatkan iterasi Newton-Raphson untuk $\hat{\boldsymbol{\beta}}$ menggunakan vektor gradien dan matriks Hessian dari langkah (4) dan (5).

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \hat{\boldsymbol{\beta}}^{(c)} - \left(\mathbf{H}(\boldsymbol{\beta})\right)^{-1}\left(\mathbf{g}(\boldsymbol{\beta})\right)$$
$$= \hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \tilde{\mathbf{p}}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$

Jika $\hat{\boldsymbol{\beta}}^{(c)} = \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c)}$ maka

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \tilde{\mathbf{p}}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$
$$= \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X^T D X}\hat{\boldsymbol{\beta}}^{(c)} + \lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)}\right) + \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \tilde{\mathbf{p}}\right) - \lambda\hat{\boldsymbol{\beta}}^{(c)}\right)$$
$$= \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X^T D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \left(\mathbf{X^T D}\right)^{-1}\lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}\left(\mathbf{y} - \tilde{\mathbf{p}}\right) - \left(\mathbf{X^T D}\right)^{-1}\lambda\mathbf{I}\hat{\boldsymbol{\beta}}^{(c)}\right]$$
$$= \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X^T D}\right)\left[\mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}\left(\mathbf{y} - \tilde{\mathbf{p}}\right)\right]$$

Sehingga update Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) yaitu

$$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X^T D z}^{(c)}, \qquad \text{dimana } \mathbf{z}^{(c)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D}^{-1}\left(\mathbf{y} - \tilde{\mathbf{p}}\right).$$

7. Mendapatkan penaksir parameter $\hat{\boldsymbol{\beta}}$ dengan mengunakan metode Truncated Newton untuk subproblem *WLS* dengan algoritma *CG* linier.

   Dalam update Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) terdapat problem numerik yaitu mendapatkan invers matriks Hessian berukuran besar sehingga pemrosesan menjadi sangat lama. Sehingga untuk mengatasinya adalah menggunakan Truncated Newton dengan CG linier sebagai algoritma truncated inner. Iterasi Newton-Raphson untuk $\boldsymbol{\beta}$ pada iterasi ke (c+1) adalah

   $$\hat{\boldsymbol{\beta}}^{(c+1)} = \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X^T D z}^{(c)}$$
   $$\left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X^T D z}^{(c)}$$

   Tahapan untuk memperoleh estimasi parameter dengan CG linier adalah sebagai berikut:

   a. Langkah awal hitung $\mathbf{A} = \mathbf{X^T D X} + \lambda\mathbf{I}$, $\mathbf{r}^{(0)} = \mathbf{X^T D z} - \left(\mathbf{X^T D X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(0)}$

      dan $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$, mulai iterasi dengan t=0

   b. Hitung

      Panjang langkah optimal $s^{(t)} = \dfrac{\mathbf{r}^{(t)T}\mathbf{r}^{(t)}}{\mathbf{d}^{(t)T}A\mathbf{d}^{(t)}}$

      Hitung perkiraan solusi $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + s^{(t)}\mathbf{d}^{(t)}$

Update *residual* $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - s^{(t)}\mathbf{A}\mathbf{d}^{(t)}$

Update A-*conjugacy* $a^{(t)} = \dfrac{\mathbf{r}^{(t+1)T}\mathbf{r}^{(t+1)}}{\mathbf{r}^{(t)T}\mathbf{r}^{(t)}}$

Update arah pencarian $\mathbf{d}^{(t+1)} = \mathbf{r}^{(t+1)} + a^{(t)}\mathbf{d}^{(t)}$

c. Iterasi dihentikan jika $\left\|\mathbf{r}^{(t)}\right\|^2 < \varepsilon$ dimana ε adalah bilangan kecil mendekati nol (pada penelitian ini ε ditetapkan sebesar 0,005)

8. Menentukan vektor bias $\mathbf{B}(\hat{\boldsymbol{\beta}})$.

   Vektor bias untuk TR-PC adalah $\mathbf{B}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{D}\boldsymbol{\xi}$.

9. Mendapatkan penaksir parameter $\tilde{\boldsymbol{\beta}}$ dengan bias terkoreksi.

   $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$.

4.2 Penerapan Metode RE-WLR dan TR-PC pada Klasifikasi Tingkat Kesejahteraan di Provinsi Bali

4.2.1 Deskripsi Karakteristik Rumah Tangga Menurut Variabel Penelitian

Pengkategorian rumah tangga pada penelitian ini didasarkan pada tingkat kesejahteraan rumah tangga (ruta) yaitu ruta miskin dan ruta tidak miskin. Pengelompokan tersebut didasarkan pada pengeluaran per kapita rumah tangga yang dibandingkan dengan garis kemiskinan yang ada.

Gambar 4.1 Persentase rumah tangga miskin dan tidak miskin di Provinsi Bali Tahun 2015

Berdasarkan Gambar 4.1 terlihat bahwa persentase rumah tangga miskin berdasarkan sampel Susenas 2015 di Provinsi Bali adalah sebesar 5,28 persen (298 ruta). Sedangkan persentase rumah tangga tidak miskin adalah sebesar 94,72 persen (5.341 ruta). Hal ini mengindikasikan bahwa data tingkat kesejahteraan rumah tangga di Provinsi Bali adalah tergolong *imbalanced data* yang cukup besar karena proporsi rumah tangga miskin jauh di bawah proporsi rumah tangga tidak miskin.

Dalam rangka analisa awal terkait karakteristik rumah tangga miskin di Provinsi Bali, maka perlu dilakukan analisis deskriptif terhadap data yang dihubungkan dengan variabel penelitian. Gambaran ini dapat membantu memberikan informasi awal terkait kondisi rumah tangga di Provinsi Bali berdasarkan variabel pada penelitian ini. Berikut ini ditampilkan analisis deskriptif mengenai kondisi literasi, pendidikan dan pekerjaan KRT.

Tabel 4.1 Kondisi literasi, pendidikan, dan lapangan usaha KRT menurut tingkat kesejahteraan (dalam persentase)

Status	Literasi KRT		Ijazah tertinggi KRT				Lap. Usaha KRT		
	Melek Huruf	Buta Huruf	PT	SMA	SMP	SD ke bawah	Pertanian	Non Tani	Lainnya
Tidak Miskin	95,3	86,7	99,9	97,2	93,7	91,8	91,5	96,5	89,1
Miskin	4,7	13,3	0,1	2,8	6,3	8,2	8,5	3,5	10,9
Jumlah	100	100	100	100	100	100	100	100	100

Berdasarkan Tabel 4.1 terlihat bahwa pada variabel pendidikan kepala rumah tangga mengindikasikan bahwa semakin tinggi pendidikan kepala rumah tangga, maka persentase kemiskinan semakin kecil, dan sebaliknya semakin rendah tingkat pendikan KRT maka persentase kemiskinannya pun semakin tinggi. KRT dengan ijazah SD ke bawah memiliki persentase miskin yang paling tinggi pada kelompoknya dan sebaliknya KRT dengan ijazah Perguruan Tinggi yang berstatus miskin sangat kecil yakni 0,1 persen. Hal ini menunjukkan bahwa pendidikan KRT memiliki hubungan dengan tingkat kesejahteraan yang terjadi. Senada dengan pendidikan, variabel literasi juga menunjukkan hal yang serupa. Pada ruta miskin, ruta dengan KRT yang berstatus buta huruf lebih tinggi

persentasenya (13,3 persen) dibandingkan dengan ruta dengan KRT melek huruf (4,7 persen) dengan perbedaan yang cukup besar. Sedangkan pada dimensi ketenagakerjaan, rumah tangga miskin dengan KRT yang bekerja di sektor pertanian lebih tinggi persentasenya dibandingkan KRT yang bekerja di sektor non pertanian. Akan tetapi persentase yang tertinggi adalah rumah tangga dengan KRT yang tidak bekerja atau penerima pendapatan yaitu sebesar 10,9 persen.

Tabel 4.2 Kondisi demografi rumah tangga menurut tingkat kesejahteraan (dalam persentase)

Status	Usia KRT				Jumlah ART			
	<35	36-45	46-55	> 56	<3	3	4	>4
Tidak Miskin	94,92	94,42	96,72	92,90	96,31	96,60	94,93	92,16
Miskin	5,08	5,58	3,28	7,10	3,69	3,40	5,07	7,84
Jumlah	100	100	100	100	100	100	100	100

Dalam dimensi demografi seperti yang terlihat pada Tabel 4.2, semakin besar ukuran rumah tangga menunjukkan semakin besar pula persentase kemiskinannya. Rumah tangga dengan jumlah ART lebih dari 4 lebih besar persentase kemiskinannya dibanding ruta yang berukuran 3 ART ke bawah. Kesimpulan untuk variabel jumlah ART ini adalah semakin banyak ARTnya maka semakin besar persentase rumah tangga miskin. Sedangkan dari segi usia KRT terlihat bahwa KRT dengan usia 56 tahun ke atas memiliki persentase kemiskinan tertinggi dibandingkan kelompok umur lainnya.

Tabel 4.3 Kondisi keadaan rumah menurut tingkat kesejahteraan (dalam persentase)

Status	Jenis Atap			Jenis Lantai			Luas Lantai per kapita	
	Genteng	Asbes/ Seng	Lainnya	Keramik	Semen	Lainnya	>10 meter	<10 meter
Tidak Miskin	96,2	90,6	45,8	97,5	89,1	64,0	97,3	83,4
Miskin	3,8	9,4	54,2	2,5	10,9	36,0	2,7	16,6
Jumlah	100	100	100	100	100	100	100	100

Kondisi bangunan rumah yang ditempati oleh suatu rumah tangga dapat menggambarkan sekilas tentang kondisi kesejahteraan rumah tangga tersebut. Hal

ini terlihat dari informasi Tabel 4.3 yang menunjukkan bahwa rumah tangga dengan jenis atap lainnya (bambu, jerami, kayu, dll) dan jenis lantai tanah memiliki persentase kemiskinan yang paling tinggi jika dibandingkan kategori lain pada masing-masing variabel. Selain itu luas lantai per kapita yang sempit juga memberikan gambaran kemiskinan pada rumah tangga. Rumah tangga miskin dengan luas lantai per kapita kurang dari 10 meter persegi persentasenya jauh lebih tinggi dibanding ruta miskin dengan luas lantai per kapitanya lebih dari 10 meter persegi . Hal ini menambah kesimpulan bahwa semakin rendah kualitas dari bagian rumah dan ukuran per kapita rumah, maka semakin tinggi pula persentase kemiskinan yang terjadi.

Tabel 4.4 Kondisi sumber air minum, sanitasi dan wilayah tempat tinggal menurut tingkat kesejahteraan (dalam persentase)

Status	Fasilitas BAB			Sumber Air Minum				Wilayah	
	Sendiri	Bersama	Tidak Ada	Air kemasan	Ledeng	Sumber lindung	Sumber tak lindung	Perdesaan	Perkotaan
Tidak Miskin	97,1	93,8	79,2	97,2	95,4	93,2	88,3	95,1	94,2
Miskin	2,9	6,2	20,8	2,8	4,6	6,8	11,7	4,9	5,8
Jumlah	100	100	100	100	100	100	100	100	100

Dari dimensi sumber air minum dan sanitasi (fasilitas buang air besar) pada Tabel 4.4 terlihat bahwa semakin bersih sumber air minum yang digunakan oleh rumah tangga, maka semakin kecil proporsi kemiskinannya. Rumah tangga miskin yang menggunakan air minum kemasan hanya sebesar 2,8 persen sedangkan rumah tangga miskin yang menggunakan sumber tak terlindung sebagai sumber minumnya memiliki proporsi paling tinggi yakni 11,7 persen.

Sedangkan dari segi sanitasi pun menunjukkan hal yang senada. Rumah tangga miskin yang memiliki fasilitas buang air besar sendiri memiliki proporsi yang paling rendah yakni hanya sebesar 2,9 persen, berbanding terbalik dengan yang tidak memiliki fasilitas buang air besar yakni sebesar 20,8 persen. Hal ini menunjukkan bahwa ketidakadaan fasilitas buang air besar pada rumah tangga miskin cukup besar.

Tabel 4.5 Kondisi aset yang dimiliki menurut tingkat kesejahteraan (dalam persentase)

Status	Aset PC/Laptop		Aset Emas perhiasan		Aset Sepeda Motor	
	Punya	Tidak	Punya	Tidak	Punya	Tidak
Tidak Miskin	99,5	93,0	99,4	92,6	96,4	85,5
Miskin	0,5	7,0	0,6	7,4	3,6	14,5
Jumlah	100	100	100	100	100	100

Aset yang dimiliki suatu rumah tangga dapat memberikan gambaran tentang kondisi ekonomi di rumah tangga tersebut. Rumah tangga yang tidak memiliki aset sama sekali biasanya merupakan rumah tangga dengan keadaan ekonomi yang cukup rendah. Sebagaimana terlihat pada Tabel 4.5 yang menunjukkan pola bahwa rumah tangga yang memiliki aset berupa PC/laptop atau emas/perhiasan atau sepeda motor memiliki kecenderungan yang lebih kecil untuk teridentifikasi menjadi rumah tangga miskin. Rumah tangga miskin yang memiliki aset-aset tersebut persentasenya jauh sangat kecil dibandingkan rumah tangga tidak miskin. Hal ini tentunya merupakan fenomena yang wajar karena pada rumah tangga miskin, jangankan untuk memiliki aset-aset tersebut, untuk mencukupi kebutuhan sehari-hari mereka tentunya sudah sulit.

4.2.2 Klasifikasi Tingkat Kesejahteraan Rumah Tangga dengan RE-WLR

Model RE-WLR terbaik dibentuk dengan menggunakan nilai parameter λ yang telah ditentukan (*pre-defined*) untuk mendapatkan model yang paling optimal. Ukuran ketepatan klasifikasi yang digunakan adalah AUC (*Area Under Curve*) dan G-mean karena ukuran ini lebih tepat digunakan pada kasus *imbalanced data.* Baik AUC maupaun G-mean mampu memberikan informasi tentang seberapa baik model yang dihasilkan dalam memprediksi baik kelas positif (*sensitivity*) maupun kelas negatif (*specificity*), sehingga tingkat akurasi yang dihasilkan cocok diterapkan pada kasus *imbalanced data* dimana proporsi kelas yang satu lebih tinggi jauh di atas kelas lainnya. Berikut ini adalah hasil

ukuran performa klasifikasi dari model RE-WLR yang telah dibentuk oleh masing-masing data training dengan menggunakan nilai lambda tertentu.

Tabel 4.6 Rata-Rata kinerja klasifikasi RE-WLR pada data *testing*

λ	G-mean (dalam persen)					AUC				
	1:3	1:4	1:5	1:6	1:10	1:3	1:4	1:5	1:6	1:10
0,01	78,06	80,13	83,12	78,15	68,54	0,786	0,799	0,832	0,787	0,722
0,1	77,93	79,98	83,17	79,00	68,55	0,785	0,797	0,833	0,795	0,722
0,5	78,34	80,29	83,21	78,81	67,57	0,790	0,801	0,833	0,793	0,715
0,6	78,36	80,32	83,27	78,83	67,57	0,790	0,801	0,834	0,794	0,715
0,7	78,39	80,33	83,47	78,61	67,59	0,790	0,801	0,836	0,792	0,716
0,8	78,33	80,28	83,46	78,37	67,61	0,789	0,801	0,835	0,789	0,716
0,9	78,32	80,24	83,46	78,37	67,67	0,790	0,801	0,835	0,789	0,717
1	78,23	80,05	83,46	78,03	67,58	0,789	0,801	0,835	0,787	0,716
1,1	78,21	80,02	83,45	78,48	67,58	0,788	0,800	0,835	0,791	0,716
1,2	78,20	79,99	83,43	78,95	67,59	0,788	0,800	0,835	0,795	0,716
1,3	78,22	80,27	83,18	78,95	67,60	0,789	0,803	0,832	0,795	0,717
2	78,68	81,34	82,68	77,25	65,37	0,795	0,814	0,827	0,780	0,702
3	78,71	82,26	82,39	76,53	64,45	0,796	0,823	0,824	0,774	0,696
4	78,83	82,17	81,95	75,78	63,91	0,799	0,822	0,820	0,759	0,692
5	78,81	82,05	81,53	75,58	63,98	0,799	0,821	0,816	0,766	0,694

Berdasarkan Tabel 4.6 diketahui bahwa lambda optimal untuk model RE-WLR terbaik saat diterapkan pada kasus tingkat kesejahteraan di Provinsi Bali adalah pada saat nilai lambda bernilai sama dengan 0,7 dan saat diterapkan pada data training dengan skema 1:5. Dengan skema tersebut menghasilkan model RE-WLR terbaik dengan nilai G-mean adalah 83,47 persen dan nilai AUC adalah 0,836. Selain itu dapat disimpulkan bahwa nilai G-mean dan AUC meningkat seiring dengan penambahan skema perbandingan ruta miskin:ruta tidak miskin dan mencapai puncaknya pada saat perbandingannya adalah 1:5 dan kemudian menurun seiring penambahan komposisi rumah tangga. Berikut ini adalah hasil lengkap ukuran klasifikasi dari model terbaik RE-WLR dengan nilai lambda 0,7 dan skema data training 1:5.

Tabel 4.7 Ukuran performa Klasifikasi model RE-WLR terbaik

Akurasi	*Sensitivity*	*Specificity*	G-Mean	AUC
84,11%	82,97%	84,17%	83,47%	0,836

Ukuran *sensitivity* dapat menunjukkan tingkat ketepatan klasifikasi model dalam memprediksi kelas positif dalam hal ini rumah tangga miskin. Pada model RE-WLR ini rata-rata nilai *sensitivity* yang diperoleh adalah sebesar 82,97 persen, artinya model RE-WLR ini mampu memprediksi rumah tangga miskin dengan tepat sebesar 82,97 persen. Sedangkan nilai *specificity* adalah ukuran yang menunjukkan tingkat ketepatan model dalam memprediksi kelas negatif dalam hal ini rumah tangga tidak miskin. Nilai *specificity* pada model RE-WLR sebesar 84,17 persen, artinya model RE-WLR ini mampu memprediksi rumah tangga tidak miskin dengan tepat sebesar 84,12 persen. Untuk menentukan apakah model klasifikasi yang dihasilkan tergolong baik maka dapat menggunakan ukuran AUC. Nilai AUC pada model ini adalah sebesar 0,836 dimana apabila nilai AUC ini dibandingkan dengan kategori hasil klasifikasi pada Tabel 2.2 maka hasil pengklasifikasian dengan menggunakan metode RE-WLR ini tergolong dalam kategori baik (*good*).

Dalam rangka pemaknaan terhadap interpretasi variabel karakteristik rumah tangga yang telah dibentuk dengan menggunakan model RE-WLR perlu untuk diketahui variabel yang signifikan dalam model. Model yang ingin diestimasi nilai parameter pada model RE-WLR ini adalah

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{26} x_{14i}.$$

Dalam mendapatkan variabel yang signifikan dalam model maka perlu dilakukan pengujian parameter secara parsial dengan hipotesis yang digunakan adalah sebagai berikut:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0,$ dengan j=0,1,2,...,d.

Pengujian secara parsial dilakukan dengan cara membandingkan nilai statistik uji W dengan nilai χ^2 dengan derajat bebas 1. Keputusan untuk menolak H_0 dilakukan apabila nilai $W_{hitung} > \chi^2_{(1)}$ atau *p-value* $< \alpha$. Pada Tabel 4.8 berikut ditampilkan estimasi parameter secara keseluruhan dengan menggunakan skema data *training* 1:5 dan nilai lambda adalah 0,7.

Tabel 4.8 Estimasi Parameter RE-WLR

Koefisien	Estimasi	Std. Error	W_{hitung}	P-Value	Odds Ratio
(Intercept)	-5,05	0,67	56,57	0,00	0,01
X_1	0,49	0,28	2,93	0,09	1,63
$X_{2(1)}$	0,19	0,23	0,64	0,42	1,21
$X_{2(2)}$	-0,37	0,27	1,89	0,17	0,69
$X_{2(3)}$	0,11	0,30	0,14	0,71	1,12
$X_{3(1)}$	0,32	0,29	1,24	0,27	1,38
$X_{3(2)}$	0,89	0,28	9,86	0,00	2,43
$X_{3(3)}$	1,91	0,28	46,76	0,00	6,74
$X_{4(1)}$	0,94	0,58	2,64	0,10	2,56
$X_{4(2)}$	0,88	0,60	2,12	0,15	2,41
$X_{4(3)}$	0,54	0,60	0,82	0,37	1,72
$X_{5(1)}$	-1,14	0,22	26,82	0,00	0,32
$X_{5(2)}$	0,89	0,28	10,33	0,00	2,44
$X_{6(1)}$	0,02	0,19	0,01	0,93	1,02
$X_{6(2)}$	1,47	0,83	3,24	0,03	4,34
$X_{7(1)}$	0,41	0,18	5,24	0,02	1,51
$X_{7(2)}$	1,17	0,44	7,06	0,01	3,22
X_8	1,77	0,19	91,78	0,00	5,89
$X_{9(1)}$	0,19	0,20	0,89	0,35	1,20
$X_{9(2)}$	0,69	0,25	7,85	0,01	1,99
$X_{10(1)}$	0,01	0,24	0,00	0,97	1,01
$X_{10(2)}$	0,05	0,23	0,05	0,82	1,05
$X_{10(3)}$	0,28	0,32	0,75	0,39	1,32
X_{11}	1,55	0,30	26,60	0,00	4,73
X_{12}	1,31	0,25	27,12	0,00	3,72
X_{13}	1,56	0,25	37,45	0,00	4,76
X_{14}	-0,98	0,20	24,10	0,00	0,37

Dengan menggunakan nilai alpha 5 persen, diperoleh 4 variabel penelitian yang tidak signifikan dalam model yakni variabel X_1 (Literasi Kepala Rumah Tangga), X_2 (Usia Kepala Rumah Tangga), X_4 (Ijazah yang ditamatkan KRT), dan X_{10} (Sumber Air Minum). Interpretasi estimasi parameter dalam regresi logistik lebih mudah dengan menggunakan nilai *odds ratio* yang dapat

menggambarkan seberapa besar suatu kategori variabel independen memiliki kecenderungan untuk mengklasifikasikan suatu rumah tangga sebagai rumah tangga miskin yang dibandingkan dengan *base level* (dalam hal ini adalah kategori dengan kode terkecil pada variabel independen). Interpretasi masing-masing variabel yang signifikan pada model RE-WLR adalah sebagai berikut:

- Variabel jumlah anggota rumah tangga (X_3) dengan nilai *odds* sebesar 1,38 untuk kategori 2 (3 ART), 2,43 untuk kategori 3 (4 ART) dan 6,74 untuk kategori 4(>4 ART).

  Nilai *odds* tersebut menunjukkan kecenderungan semakin tinggi jumlah ART, maka kecenderungan rumah tangga untuk terklasifikasi sebagai rumah tangga miskin semakin besar. Hal ini bisa terlihat dari *odds* untuk kategori 4 yang memiliki nilai paling tinggi, yang dapat diinterpretasikan bahwa rumah tangga dengan ART lebih dari 4 memiliki kecenderungan 6,74 kali lebih besar dibanding dengan rumah tangga dengan 1 atau 2 ART untuk terklasifikasi sebagai rumah tangga miskin. Demikian pula interpretasi untuk kategori 2 dan 3.

- Variabel lapangan usaha pekerjaan KRT (X_5) dengan nilai *odds* sebesar 0,32 untuk kategori 2 (sektor non pertanian) dan 2,44 untuk kategori 3 (tidak bekerja atau penerima pendapatan).

  Interpretasi untuk *odds* kategori 2 adalah rumah tangga dengan KRT bekerja di sektor pertanian memiliki kecenderungan 3,12 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding dengan ruta dengan KRT yang bekerja di sektor non pertanian. Sedangkan ruta dengan KRT yang tidak bekerja memiliki kecenderungan 2,44 kali lebih besar terklasifikasi sebagai ruta miskin dibanding ruta dengan KRT yang bekerja di sektor pertanian.

- Variabel jenis atap terluas (X_6) dengan nilai *odds* untuk kategori 3 (atap jenis bambu, kayu, jerami atau lainnya) adalah 4,34. Hal ini menunjukkan bahwa ruta dengan atap jerami atau kayu memiliki kecenderungan 4,34 kali lebih besar dibandingkan rumah tangga dengan atap rumah berupa genteng untuk terklasifikasi sebagai ruta miskin.

- Variabel jenis lantai terluas (X_7) dengan nilai *odds* untuk kategori 2 (semen/bata merah) adalah 1,51 dan nilai *odds* untuk kategori 3 (tanah, bambu atau lainnya) adalah 3,22.

  Semakin rendah kualitas lantai pada rumah yang ditempati, maka semakin besar kecenderungan rumah tangga tersebut untuk terklasifikasi sebagai ruta miskin. Hal ini terlihat dari nilai *odds* kategori 3 lebih besar dari nilai odds kategori 2 dengan interpretasi nilai *odds* adalah rumah tangga dengan jenis lantai berupa tanah memiliki kecenderungan 3,22 kali lebih besar dibandingkan rumah tangga dengan jenis lantai keramik untuk terklasifikasi sebagai rumah tangga miskin.
- Variabel luas lantai per kapita (X_8) dengan nilai *odds* 5,89 yang berarti bahwa rumah tangga dengan luas lantai per kapita kurang dari 10 meter persegi memiliki kecenderungan 5,89 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding rumah tangga dengan luas lantai per kapita lebih dari 10 meter persegi.
- Variabel fasilitas tempat buang air besar (X_9) dengan nilai *odds* untuk kategori 3 (umum/tidak ada) sebesar 1,99. Hal ini menunjukkan bahwa rumah tangga yang tidak memiliki fasilitas buang air besar memiliki kecenderungan sebesar 1,99 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding rumah tangga yang memiliki fasilitas buang air besar.
- Variabel kepemilikan aset PC/laptop (X_{11}) dengan nilai *odds* 4,73 yang berarti bahwa rumah tangga yang tidak memiliki laptop/PC memiliki kecenderungan 4,73 lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibandingkan rumah tangga yang memiliki PC/laptop.
- Variabel kepemilikan aset emas/perhiasan minimal 10 gram (X_{12}) dengan nilai *odds* 3,72 yang berarti bahwa rumah tangga yang tidak memiliki aset berupa emas perhiasan dengan nominal minimal 10 gram memiliki kecenderungan 3,72 kali lebih besar untuk terklasifikasi menjadi rumah tangga miskin dibanding rumah tangga yang memiliki aset tersebut.
- Variabel kepemilikan aset sepeda motor (X_{13}) dengan nilai *odds* 4,76 yang menunjukkan bahwa rumah tangga yang tidak memiliki sepeda motor memiliki

kecenderungan 4,76 kali lebih besar untuk terklasifikasi menjadi rumah tangga miskin dibanding rumah tangga yang memiliki sepeda motor.

- Variabel status wilayah tempat tinggal (X_{14}) dengan nilai *odds* 0,37 yang mempunyai arti bahwa rumah tangga yang tinggal di perkotaan memiliki kecenderungan 2,7 kali lebih besar terklasifikasi sebagai rumah tangga miskin dibandingkan rumah tangga yang tinggal di wilayah pedesaan.

4.2.3 Klasifikasi Tingkat Kesejahteraan Rumah Tangga dengan TR-PC

Seperti halnya pada model RE-WLR, pembentukan model TR-PC terbaik juga dibentuk dengan menggunakan nilai parameter λ yang telah ditentukan (*pre-defined*) untuk mendapatkan model yang paling optimal. Model dengan simulasi skema data *training* dan nilai parameter λ yang menghasilkan ketepatan klasifikasi yang paling tinggi merupakan model TR-PC yang terbaik. Berikut ini adalah hasil ukuran performa klasifikasi dari model TR-PC yang telah dibentuk oleh masing-masing data training dengan menggunakan nilai lambda tertentu.

Tabel 4.9 Rata-Rata kinerja klasifikasi TR-PC pada data *testing*

λ	G-mean (dalam persen)					AUC				
	1:3	1:4	1:5	1:6	1:10	1:3	1:4	1:5	1:6	1:10
0,01	70,948	70,633	73,652	64,272	61,985	0,7331	0,7366	0,7614	0,6954	0,6851
0,02	70,948	70,623	73,652	64,272	61,992	0,7331	0,7366	0,7614	0,6954	0,6851
0,03	70,942	70,657	73,666	64,272	62,001	0,7331	0,7371	0,7616	0,6954	0,6852
0,04	70,942	70,291	73,680	63,466	62,001	0,7331	0,7344	0,7618	0,6901	0,6852
0,05	70,937	69,921	73,677	63,670	61,374	0,7330	0,7323	0,7617	0,6914	0,6812
0,06	70,937	69,939	73,677	63,670	61,363	0,7330	0,7319	0,7617	0,6914	0,6810
0,1	70,700	69,970	73,422	63,687	61,363	0,7313	0,7323	0,7597	0,6913	0,6810
1	67,054	65,473	69,465	62,130	59,720	0,7068	0,7020	0,7335	0,6834	0,6717
2	64,790	62,313	69,445	62,233	57,904	0,6923	0,6817	0,7311	0,6849	0,6617
3	63,934	62,847	67,610	62,301	56,165	0,6868	0,6865	0,7212	0,6861	0,6521
4	63,452	62,314	66,404	61,667	52,464	0,6844	0,6838	0,7135	0,6831	0,6331
5	62,003	61,033	65,175	60,866	51,849	0,6754	0,6766	0,7058	0,6784	0,6302

Berdasarkan Tabel 4.9 diketahui bahwa lambda optimal untuk model TR-PC terbaik adalah pada saat nilai lambda bernilai sama dengan 0,04 dan saat diterapkan pada data *training* dengan skema 1:5. Dengan skema tersebut

menghasilkan model TR-PC terbaik dengan nilai G-mean adalah 73,68 persen dan nilai AUC adalah 0,762. Selain itu pola yang dapat dapat disimpulkan adalah bahwa nilai G-mean dan AUC menurun seiring dengan penambahan nilai lambda, sedangkan titik tertinggi nilai AUC dan G-mean terjadi saat skema data *training* adalah 1:5 kemudian terus menurun seiring perbandingan ruta miskin dan ruta tidak miskin yang semakin besar. Berikut ini adalah hasil lengkap ukuran klasifikasi dari model terbaik TR-PC dengan nilai lambda 0,04 dan skema data *training* 1:5.

Tabel 4.10 Ukuran performa Klasifikasi model TR-PC terbaik

Akurasi	*Sensitivity*	*Specificity*	G-Mean	AUC
93,41%	57,03%	95,32%	73,68%	0,7618

Ukuran *sensitivity* dapat menunjukkan tingkat ketepatan klasifikasi model dalam memprediksi kelas positif dalam hal ini rumah tangga miskin. Pada model TR-PC ini rata-rata nilai *sensitivity* sebesar 57,03 persen, artinya model TR-PC ini mampu memprediksi rumah tangga miskin dengan tepat sebesar 57,03 persen. Sedangkan nilai *specificity* adalah ukuran yang menunjukkan tingkat ketepatan model dalam memprediksi kelas negatif dalam hal ini rumah tangga tidak miskin. Nilai *specificity* pada model TR-PC sebesar 95,32 persen, artinya model TR-PC ini mampu memprediksi rumah tangga tidak miskin dengan tepat sebesar 95,32 persen. Untuk menentukan apakah model klasifikasi yang dihasilkan tergolong baik maka dapat digunakan ukuran AUC. Nilai AUC pada model ini adalah sebesar 0,7618 dimana apabila nilai AUC ini dibandingkan dengan kategori hasil klasifikasi pada Tabel 2.2 maka hasil pengklasifikasian dengan menggunakan metode TR-PC ini tergolong dalam kategori cukup (*fair*).

Dalam rangka pemaknaan terhadap interpretasi variabel karakteristik rumah tangga yang telah dibentuk dengan menggunakan model TR-PC. Model yang akan diestimasi parameternya menggunakan pada model TR-PC adalah

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{26} x_{14i}, \quad \text{dengan } \hat{p}_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right) - \mathbf{x}_i^T \boldsymbol{\beta}}}.$$

Hipotesis untuk uji parsial dan statistik uji yang digunakan dengan hipotesis dan statistik uji parsial pada model RE-WLR di atas. Sehingga pada Tabel 4.11 berikut ditampilkan estimasi parameter secara keseluruhan dengan menggunakan skema data *training* 1:5 dan nilai lambda adalah 0,04.

Tabel 4.11 Estimasi Parameter TR-PC

Koefisien	Estimasi	Std. Error	W_{hitung}	P-value	Odds Ratio
(Intercept)	-5.62	1.16	23.31	0.00	0.00
X_1	0.56	0.38	2.19	0.14	1.75
$X_{2(1)}$	0.08	0.32	0.07	0.80	1.09
$X_{2(2)}$	-0.39	0.39	1.02	0.31	0.68
$X_{2(3)}$	0.06	0.40	0.02	0.88	1.06
$X_{3(1)}$	0.37	0.42	0.78	0.38	1.45
$X_{3(2)}$	1.00	0.40	6.16	0.01	2.72
$X_{3(3)}$	2.07	0.40	27.13	0.00	7.89
$X_{4(1)}$	1.55	1.06	2.12	0.15	4.71
$X_{4(2)}$	1.52	1.09	1.94	0.16	4.56
$X_{4(3)}$	1.04	1.08	0.91	0.34	2.82
$X_{5(1)}$	-1.07	0.29	13.23	0.00	0.34
$X_{5(2)}$	1.09	0.38	8.34	0.00	2.98
$X_{6(1)}$	0.03	0.26	0.01	0.92	1.03
$X_{6(2)}$	2.49	1.25	3.98	0.03	12.12
$X_{7(1)}$	0.40	0.25	2.41	0.12	1.49
$X_{7(2)}$	1.34	0.57	5.57	0.02	3.84
X_8	1.66	0.25	42.45	0.00	5.24
$X_{9(1)}$	0.17	0.28	0.38	0.54	1.19
$X_{9(2)}$	0.88	0.32	7.38	0.01	2.40
$X_{10(1)}$	-0.04	0.34	0.01	0.91	0.96
$X_{10(2)}$	-0.02	0.32	0.00	0.95	0.98
$X_{10(3)}$	0.19	0.43	0.20	0.65	1.21
X_{11}	1.47	0.48	9.43	0.00	4.34
X_{12}	1.32	0.39	11.54	0.00	3.75
X_{13}	1.51	0.33	20.65	0.00	4.53
X_{14}	-0.92	0.27	11.41	0.00	0.40

Dengan menggunakan nilai alpha 5 persen, diperoleh 4 variabel penelitian yang tidak signifikan dalam model yakni variabel X_1 (Literasi Kepala Rumah Tangga), X_2 (Usia Kepala Rumah Tangga), X_4 (Ijazah yang ditamatkan KRT), dan X_{10} (Sumber Air Minum). Variabel-variabel yang tidak signifikan pada model TR-PC tersebut sama dengan variabel yang tidak signifikan pada model RE-WLR. Interpretasi estimasi parameter dalam regresi logistik lebih mudah

dengan menggunakan nilai *odds ratio* yang dapat menggambarkan seberapa besar suatu kategori variabel independen memiliki kecenderungan untuk mengklasifikasikan suatu rumah tangga sebagai rumah tangga miskin yang dibandingkan dengan *base level* (dalam hal ini adalah kategori dengan kode terkecil pada variabel independen). Interpretasi masing-masing variabel yang signifikan pada model TR-PC adalah sebagai berikut:

- Variabel jumlah anggota rumah tangga (X_3) dengan nilai *odds* sebesar 1,45 untuk kategori 2 (3 ART), 2,72 untuk kategori 3 (4 ART) dan 7,89 untuk kategori 4(>4 ART).

  Nilai *odds* tersebut menunjukkan kecenderungan semakin tinggi jumlah ART, maka kecenderungan rumah tangga untuk terklasifikasi sebagai rumah tangga miskin semakin besar. Hal ini bisa terlihat dari *odds* untuk kategori 4 yang memiliki nilai paling tinggi yang dapat diinterpretasikan bahwa rumah tangga dengan ART lebih dari 4 memiliki kecenderungan 7,89 kali lebih besar dibanding dengan rumah tangga dengan 1 atau 2 ART untuk terklasifikasi sebagai rumah tangga miskin. Demikian interpretasi pula untuk kategori 2 dan 3.
- Variabel lapangan usaha pekerjaan KRT (X_5) dengan nilai *odds* sebesar 0,34 untuk kategori 2 (sektor non pertanian) dan 2,98 untuk kategori 3 (tidak bekerja atau penerima pendapatan).

  Interpretasi untuk *odds* kategori 2 adalah rumah tangga dengan KRT bekerja di sektor pertanian memiliki kecenderungan 2,94 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding dengan ruta dengan KRT yang bekerja di sektor non pertanian. Sedangkan ruta dengan KRT yang tidak bekerja memiliki kecenderungan 2,98 kali lebih besar terklasifikasi sebagai ruta miskin dibanding ruta dengan KRT yang bekerja di sektor pertanian.
- Variabel jenis atap terluas (X_6) dengan nilai *odds* untuk kategori 3 (ata p jenis bambu, kayu, jerami atau lainnya) adalah 12,12. Hal ini menunjukkan bahwa ruta dengan atap jerami atau kayu memiliki kecenderungan 12,12 kali lebih

besar dibandingkan rumah tangga dengan atap rumah berupa genteng untuk terklasifikasi sebagai ruta miskin.

- Variabel jenis lantai terluas (X_7) dengan nilai *odds* untuk kategori 2 (semen/bata merah) adalah 1,49 dan nilai *odds* untuk kategori 3 (tanah, bambu atau lainnya) adalah 3,84.

  Semakin rendah kualitas lantai pada rumah yang ditempati, maka semakin besar kecenderungan rumah tangga tersebut untuk terklasifikasi sebagai ruta miskin. Hal ini terlihat dari nilai *odds* kategori 3 lebih besar dari nilai odds kategori 2 dengan interpretasi nilai *odds* adalah rumah tangga dengan jenis lantai berupa tanah memiliki kecenderungan 3,84 kali lebih besar dibandingkan rumah tangga dengan jenis lantai keramik untuk terklasifikasi sebagai rumah tangga miskin.
- Variabel luas lantai per kapita (X_8) dengan nilai *odds* 5,24 yang berarti bahwa rumah tangga dengan luas lantai per kapita kurang dari 10 meter persegi memiliki kecenderungan 5,24 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding rumah tangga dengan luas lantai per kapita lebih dari 10 meter persegi.
- Variabel fasilitas tempat buang air besar (X_9) dengan nilai *odds* untuk kategori 3 (umum/tidak ada) sebesar 2,40. Hal ini menunjukkan bahwa rumah tangga yang tidak memiliki fasilitas buang air besar memiliki kecenderungan sebesar 2,40 kali lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibanding rumah tangga yang memiliki fasilitas buang air besar.
- Variabel kepemilikan aset PC/laptop (X_{11}) dengan nilai *odds* 4,34 yang berarti bahwa rumah tangga yang tidak memiliki laptop/PC memiliki kecenderungan 4,34 lebih besar untuk terklasifikasi sebagai rumah tangga miskin dibandingkan rumah tangga yang memiliki PC/laptop.
- Variabel kepemilikan aset emas/perhiasan minimal 10 gram (X_{12}) dengan nilai *odds* 3,75 yang berarti bahwa rumah tangga yang tidak memiliki aset berupa emas perhiasan dengan nominal minimal 10 gram memiliki

kecenderungan 3,75 kali lebih besar untuk terklasifikasi menjadi rumah tangga miskin dibanding rumah tangga yang memiliki aset tersebut.

- Variabel kepemilikan aset sepeda motor (X_{13}) dengan nilai *odds* 4,53 yang menunjukkan bahwa rumah tangga yang tidak memiliki sepeda motor memiliki kecenderungan 4,53 kali lebih besar untuk terklasifikasi menjadi rumah tangga miskin dibanding rumah tangga yang memiliki sepeda motor.
- Variabel status wilayah tempat tinggal (X_{14}) dengan nilai *odds* 0,4 yang mempunyai arti bahwa rumah tangga yang tinggal di perkotaan memiliki kecenderungan 2,5 kali lebih besar terklasifikasi sebagai rumah tangga miskin dibandingkan rumah tangga yang tinggal di wilayah pedesaan.

4.3 Perbandingan Ketepatan Klasifikasi antara Model RE-WLR dan TR-PC

Setelah mendapatkan model terbaik dari masing-masing metode, perlu dilakukan perbandingan performa klasifikasi untuk memperoleh kesimpulan tentang model terbaik pada kasus klasifikasi tingkat kesejahteraan di Provinsi Bali. Indikator ketepatan klasifikasi yang digunakan untuk membandingkan kedua metode dalam penelitian ini adalah ukuran G-mean dan AUC. Alasan penggunaan indikator ini karena indikator ini lebih tepat digunakan dalam kasus *imbalanced data* dan mampu mengakomodir akurasi pada kelas positif dan kelas negatif. Dalam konteks klasifikasi tingkat kesejahteraan rumah tangga, dimana proporsi rumah tangga miskin jauh lebih sedikit dibandingkan rumah tangga tidak miskin, maka penggunaan kedua indikator ini lebih tepat digunakan untuk membandingkan tingkat ketepatan klasifikasi. Model RE-WLR dan TR-PC terbaik juga dibandingkan dengan hasil klasifikasi jika menggunakan metode regresi logistik biasa.

Tabel 4.12 Perbandingan kinerja klasifikasi antara RE-WLR, TR-PC dan Reglog

Metode	G-mean	AUC	Hasil Klasifikasi
RE-WLR	83,47%	0,836	*Good*
TR-PC	73,68%	0,762	*Fair*
Regresi Logistik	50,00%	0,623	*Poor*

Dari hasil pengolahan seperti yang tertera pada Tabel 4.12 terlihat bahwa pada metode RE-WLR mampu memberikan hasil yang lebih baik dalam hal keakuratan dalam *imbalanced data* dibandingkan metode TR-PC maupun regresi logistik biasa. Sedangkan hasil klasifikasi dengan metode TR-PC juga memberikan hasil yang berbeda secara signifikan bila dibandingkan dengan regresi logistik biasa. Sehingga dalam kasus klasifikasi tingkat kesejahteraan rumah tangga di Provinsi Bali, metode RE-WLR mampu memberikan hasil yang lebih baik dibandingkan metode TR-PC dan regresi logistik.

Penambahan penimbang dalam model RE-WLR terbukti mampu memberikan hasil klasifikasi pada kasus *imbalanced data* menjadi lebih akurat pada kelas minoritas. Hal ini dikarenakan pemberian bobot yang lebih besar pada kelas minoritas dapat memberikan informasi yang lebih besar dibandingkan pada model tanpa penimbang. Skema *case based sampling* yang digunakan pada penelitian ini membuat model dapat menangkap informasi yang cukup terhadap kedua kelas tanpa memberikan kesimpulan yang bias terhadap populasi dan hal ini dikarenakan peran penimbang yang digunakan pada model RE-WLR. Sehingga akurasi yang dihasilkan menjadi lebih akurat.

halaman ini sengaja dikosongkan

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Beberapa kesimpulan yang dapat diambil berdasarkan pembahasan dan merujuk pada tujuan penelitian adalah sebagai berikut

1. Metode iterasi numerik yang digunakan dalam metode ini adalah Metode Truncated Newton yaitu gabungan antara metode iterasi Newton-Raphson dengan metode Conjugate Gradient linier sehingga dengan penambahan algoritma ini mampu menghasilkan konvergenitas lebih cepat.
2. Dengan menggunakan metode RE-WLR dan TR-PC terbentuk variabel-variabel yang secara signifikan mampu menggambarkan kemiskinan rumah tangga di Provinsi Bali. Variabel yang dihasilkan oleh kedua metode tersebut antara lain variabel jumlah anggota rumah tangga, lapangan usaha pekerjaan kepala rumah tangga, jenis atap terluas, jenis lantai terluas, luas lantai per kapita, penggunaan fasilitas buang air besar, kepemilikan aset PC/laptop, sepeda motor, emas perhiasan minimal 10 gram dan status wilayah tempat tinggal.
3. Dari hasil perbandingan kinerja klasifikasi dengan menggunakan metode RE-WLR, TR-PC dan regresi logistik didapatkan hasil bahwa model RE-WLR lebih baik performanya dibandingkan model TR-PC dan regresi logistik. Model RE-WLR mampu memberikan hasil yang lebih bagus dengan nilai AUC tertinggi adalah 0,84 (kategori baik), sedangkan hasil klasifikasi pada model TR-PC menghasilkan nilai AUC tertingi 0,76 (kategori cukup).

5.2 Saran

Berdasarkan penelitian yang telah dilakukan maka saran yang dapat diberikan adalah sebagai berikut:

1. Dalam penelitian ini, variabel respon yang digunakan masih bersifat biner sehingga untuk mengembangkan lebih jauh dapat mengembangkan model RE-WLR dan TR-PC untuk variabel respon yang bersifat *multi class.*
2. Variabel independen yang digunakan pada penelitian ini masih bersifat tunggal dalam artian bukan berupa gabungan variabel yang saling berhubungan, sehingga untuk penelitian selanjutnya dapat digunakan penambahan variabel independen yang bersifat komposit untuk menambah informasi pada variabel penjelas.
3. Perbandingan metode yang digunakan dalam penelitian ini masih dalam sebatas antar regresi logistik, sehingga untuk penelitian selanjutnya dapat membandingkan performa klasifikasi dengan metode klasifikasi lainnya.

DAFTAR PUSTAKA

Agresti, A., (2007), *An Introduction to Categorical Data Analysis 2nd Edition*, John Wiley & Sons, Inc: New Jersey.

Arleina, O.D. dan Otok, B.W., (2014), "Bootsrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) untuk Mengklasifikasikan Rumah Tangga Miskin di Kabupaten Jombang", *Jurnal Sains dan Seni POMITS*, Vol. 3, No.2, hal 91-96.

Bappenas, (2014), *Strategi Nasional Penanggulangan Kemiskinan*, Badan Perencanaan Pembangunan Nasional, Jakarta.

Bielza, C., Robles, V., dan Larranaga, P., (2011), "Regularized Logistic Regression Without a Penalty Term: an Aplication to Cancer Classification with Microarray Data", *Expert Systems with Applications*, Vol. 38, Hal. 5110-5118.

Bishop, C.M., (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

BPS, (2016), *Penghitungan dan Analisis Kemiskinan Makro Indonesia 2016*, Badan Pusat Statistik, Jakarta.

BPS Prov. Bali, (2016), *Indikator Kesejahteraan Rakyat Provinsi Bali 2015*, Badan Pusat Statistik Provinsi Bali, Denpasar.

Cramer, JS., (2005), "Logit Models From Economics And Other Fields", *Technometrics*, Vol. 47, No.1, Hal. 97-98.

Fawcett, T., (2006), "An Introduction to ROC Analysis", *Pattern Recognition Letters*, Vol. 27, Hal. 861-874.

Foster, J., Greer, J., dan Thorbecke, E., (1984), "A Class of Decomposable Poverty Measures", *Econometrica*, Vol. 52, Issue 3, Hal. 761-766.

Garthwaite, P., Jolliffe, I., dan Byron, J., (2002), *Statistical Inference*, Oxford University Press, Oxford.

Gujarati, D.N., (2003), *Basic Econometrics 4th Edition*, McGraw-Hill, New York, USA.

Han, J., Kamber, M., dan Pei,J., (2011), *Data Mining Concepts and Techniques 3rd Edition* , Waltham: Elsevier Inc.

Hasbullah, J., (2012), *Tangguh dengan Statistik*, Nuansa Cendekia, Bandung.

Hastie, T., Tibshirani, R., dan Friedman, J. (2013), *The Elements of Statistical Learning; Data Mining, Inference, and Prediction 2nd Edition*, Springer, California.

Hosmer, D.W. dan Lemeshow, S., (2000), *Applied Logistic Regression 2nd edition*, John Wiley & Sons, Inc: New Jersey.

King, G. dan Zeng, L., (2001), "Logistic Regression in Rare Events Data.", *Political Analysis 9(2)*, hal. 137-163.

Komarek, P., (2004), *Logistic Regression for Data Mining and High-Dimensional Classification,* Tesis PhD, Carnegie Mellon University, Pittsburgh.

Komarek, P. dan Moore, A., (2005), "Making Logistic Regression a Core Data Mining Tool with TR-IRLS", *Proceedings of the Fifth IEEE Conference on Data Mining*, IEEE Computer Society, Washington DC.

Maalouf, M., (2011), "Logistic Regression in Data Analysis: an Overview", *Int. Jurnal Data Analysis Technique and Strategies*, Vol.3, No.3, hal. 281-299.

Maalouf, M. dan Siddiqi, M., (2014), "Weighted Logistic Regression for Large-Scale Imbalanced and Rare Events Data", *Journal of Knowledge Based Systems*, Vol. 59, hal. 142-148.

Maalouf, M. Homouz, D., dan Trafalis, T.B., (2017), "Logistic Regression in Large Rare Events and Imbalanced Data: A Performance Comparison of Prior Correction and Weighting Metdhods", *Computational Inteligence*, hal. 1-14.

Maalouf, M. dan Trafalis, T.B., (2011), "Robust Weighted Kelrnel Logistic Regression in Imbalanced and Rare Events Data", *Computational Statistics and Data Analysis*, Vol. 55, Hal. 168-183.

Mathiassen, A. dan Roll-Halsen, D., (2007), *Predicting Poverty for Mozambique 2000 to 2005; How Robust are the Models?*, Statistics Norway/Division for Development Cooperation, Oslo.

McCullagh, P. dan Nelder, J.A., (1989), *Generalized Linear Models 2nd Edition*, Chapman and Hall.

Minka, T.P., (2007), "A Comparison of Numerical Optimizers for Logistic Regression", Tech. Rep., Department of Statistics, Carnegie Mellon University.

Mirawanti, Y., (2012), *Perbandingan Metode Regresi Logistik Ordinal dengan Jaringan Syaraf Tiruan Fungsi Radial Basis (Studi Kasus: Klasifikasi Rumah Tangga Miskin Kota Pasuruan Tahun 2008)*, Tesis, Institut Teknologi Sepuluh Nopember, Surabaya.

Mohanty, A.K., Beberta, S., dan Lenka,S.K. (2011). Classifying Benign and Malignant Mass using GLCM and GLRLM based Texture Features form Mammogram, International Journal of Engineering Research and Applications (IJERA), Vol.1, hal 687-693.

Mulia, A., (2005), *Analisis Penetapan Penduduk Miskin (Studi Kasus di Kabupaten Kupang, Nusa Tenggara Timur*, Tesis, Institut Pertanian Bogor, Bogor.

Nopriyansyah, Junaidi, dan Umiyati, E.,, (2015), "Determinan Kemiskinan Rumah Tangga di Provinsi Jambi", *Jurnal Perspektif Pembiayaan dan Pembangunan Daerah*, Vol 2, No. 3, hal. 119-128.

Owen, A.B., (2007), "Infinitely Imbalanced Logistic Regression", *Journal of Machine Learning Research* , Vol. 8, Hal 761-773.

Palepu, K., (1986), Predicting takeover targets: a methodological and empirical analysis, *J Account Econ*, Vol.8 Hal.3‑35.

Pohlman, J.T. dan Leitner, D.W., (2003), A Comparison of Ordinary Least Squares and Logistic Regression, *Ohio Juournal of Science*, Vol 103, Issue 5, Hal. 118-125.

Putri, A.D (2013), "Pengaruh Umur, Pendidikan, Pekerjaan terhadp Pendapatan Rumah Tangga Miskin di Desa Bebandem", *Jurnal Ekonomi Pembangunan Unud*, Vol 2, No. 4, hal. 173-180.

Sulasih, D.E., (2016), *Rare Event Weighted Logistic Regression untuk Klasifikasi Imbalanced Data (Studi Kasus: Klasifikasi Desa Tertinggal di Provinsi Jawa Timur),* Tesis, Institut Teknologi Sepuluh Nopember, Surabaya.

United Nations, (2015), *Transforming Our World: The 2030 Agenda for Sustainable Development*, United Nations, New York.

World Bank, (2000), *World Development Report 2000/2001*, World Bank, Washington DC.

LAMPIRAN

Lampiran 1. Algoritma RE-WLR menggunakan IRLS

Algoritma 1.1 (Maalouf dan Siddiqi, 2014)

Mencari Estimator MLE RE-WLR menggunakan IRLS

Data: $\mathbf{X}, \mathbf{y}, \hat{\beta}^{(0)}, w_1, w_0$

Result: $\hat{\beta}, \mathbf{B}\left(\hat{\beta}\right), \tilde{\beta}, \tilde{\mathrm{p}}_i$

begin
  c = 0
  while $\left|\frac{DEV^{(c)} - DEV^{(c+1)}}{DEV^{(c+1)}}\right| > \varepsilon_1$ **and** $c \leq$ *Max IRLS Iterations* **do**
    for $i \leftarrow 1$ **to** n **do**
      $\hat{p}_i = \frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\beta}}}$ **/*** compute probabilities ***/**
      $v_i = \hat{p}_i\left(1-\hat{p}_i\right)$ **/*** compute variance ***/**
      $w_i = w_1 y_1 + w_0\left(1-y_i\right)$ /* compute weights ***/**
      $z_i = \mathbf{x}_i\hat{\beta}^{(c)} + \frac{\left(y_i-\hat{p}_i\right)}{\hat{p}_i\left(1-\hat{p}_i\right)}$ **/*** compute adjusted response ***/**
    $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ **/*** compute the covariance matrix ***/**
    $\mathbf{Q}_{ii} = diag\left(\mathbf{Q}\right)$ **/* */**
    for $k \leftarrow 1$ **to** n **do**
      $\xi_k = \frac{1}{2}Q_{ii_k}\left(\left(1+w_1\right)\hat{p}_k - w_1\right)$ **/*** compute the bias response ***/**
    $\mathbf{D} = diag\left(v_i w_i\right)$ **/* */**
    $\left(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{z}^{(c)}$ **/*** compute $\hat{\boldsymbol{\beta}}$ via Algorithm 2 (CG) ***/**
    $\left(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{B}\left(\hat{\boldsymbol{\beta}}^{(c+1)}\right) = \mathbf{X}^{\mathrm{T}}\mathbf{D}\xi^{(c)}$ **/*** compute $\mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$ via Algorithm 3 (CG) ***/**
    $c = c+1$
  $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$ **/*** compute the unbiased $\hat{\boldsymbol{\beta}}$ ***/**
  $\tilde{p}_i = \frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\beta}}}$ **/*** compute the optimal probabilities ***/**

Lampiran 2. Algoritma TR-PC menggunakan IRLS

Algoritma 1.2 (Maalouf dkk, 2017)

Mencari Estimator MLE TR-PC menggunakan IRLS

Data: $\mathbf{X}, \mathbf{y}, \hat{\beta}^{(0)}, \delta_0, \delta_1$

Result: $\hat{\boldsymbol{\beta}}, \mathbf{B}\left(\hat{\boldsymbol{\beta}}\right), \tilde{\boldsymbol{\beta}}, \tilde{\mathrm{p}}_i$

begin

  c = 0

  while $\left|\frac{DEV^{(c)} - DEV^{(c+1)}}{DEV^{(c+1)}}\right| > \varepsilon_1$ **and** $c \leq$ *Max IRLS Iterations* **do**

    for $i \leftarrow 1$ **to** n **do**

      $\tilde{p}_i = \frac{1}{1+e^{\ln\left(\frac{\delta_1}{\delta_0}\right)-\mathbf{x}_i\boldsymbol{\beta}}}$ **/*** compute probabilities ***/**

      $v_i = \tilde{p}_i\left(1-\tilde{p}_i\right)$ **/*** compute variance ***/**

      $z_i = \mathbf{x}_i\hat{\beta}^{(c)} + \frac{\left(y_i - \tilde{p}_i\right)}{\tilde{p}_i\left(1-\tilde{p}_i\right)}$ **/*** compute adjusted response ***/**

    $\mathbf{D} = \mathrm{diag}\left(v_i\right)$ **/* */**

    $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ **/*** compute the covariance matrix ***/**

    $\mathbf{Q}_{ii} = diag\left(\mathbf{Q}\right)$ **/* */**

    for $k \leftarrow 1$ **to** n **do**

      $\xi_k = Q_{ii_k}\left(\tilde{p}_k - \frac{1}{2}\right)$ **/*** compute the bias response ***/**

    $\left(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{z}^{(c)}$ **/*** compute $\hat{\boldsymbol{\beta}}$ via Algorithm 2 (CG) ***/**

    $\left(\mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{B}\left(\hat{\boldsymbol{\beta}}^{(c+1)}\right) = \mathbf{X}^{\mathrm{T}}\mathbf{D}\xi^{(c)}$ **/*** compute $\mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$ via Algorithm 3 (CG) ***/**

    $c = c+1$

  $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}\left(\hat{\boldsymbol{\beta}}\right)$ **/*** compute the unbiased $\hat{\boldsymbol{\beta}}$ ***/**

  $\tilde{p}_i = \frac{1}{1+e^{\ln\frac{\delta_1}{\delta_0}-\mathbf{x}_i\boldsymbol{\beta}}}$ **/*** compute the optimal probabilities ***/**

Lampiran 3. Algoritma Menghitung $\hat{\boldsymbol{\beta}}$ menggunakan CG linier

Algoritma 2 (Maalouf dan Siddiqi, 2014)

Menghitung $\hat{\boldsymbol{\beta}}$ menggunakan Linear CG, dimana $\mathbf{A} = \mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}$, dan $\mathbf{b} = \mathbf{X}^T\mathbf{D}\mathbf{z}$

Data: $\mathbf{A}, \mathbf{b}, \hat{\beta}^{(0)}$

Result: $\hat{\boldsymbol{\beta}}$ such that $\mathbf{A}\hat{\boldsymbol{\beta}} = \mathbf{b}$

begin

$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\hat{\beta}^{(0)}$ /* initialize the residual */

$c = 0$

while $\left\|\mathbf{r}^{(c+1)}\right\|^2 > \varepsilon_2$ **and** $c \leq$ *Max CG Iterations* **do**

  if $c = 0$ **then**

    $\zeta^{(c)} = 0$

  else

    $\zeta^{(c)} = \dfrac{\mathbf{r}^{\mathrm{T}(c+1)}\mathbf{r}^{(c+1)}}{\mathbf{r}^{\mathrm{T}(c)}\mathbf{r}^{(c)}}$ /* update A-Conjugacy enforcer */

  $\mathbf{d}^{(c+1)} = \mathbf{r}^{(c+1)} + \zeta^{(c)}\mathbf{d}^{(c)}$ /* update the serch direction */

  $s^{(c)} = \dfrac{\mathbf{r}^{\mathrm{T}(c)}\mathbf{r}^{(c)}}{\mathbf{d}^{\mathrm{T}(c)}\mathbf{A}\mathbf{d}^{(c)}}$ /* compute the optimal step length */

  $\hat{\beta}^{(c+1)} = \hat{\beta}^{(c)} + \zeta^{(c)}\mathbf{d}^{(c+1)}$ /* obtain an approximate solution */

  $\mathbf{r}^{(c+1)} = \mathbf{r}^{(c)} - s^{(c)}\mathbf{A}\mathbf{d}^{(c+1)}$ /* Update the residual */

  $c = c + 1$

Lampiran 4. Algoritma Menghitung B($\hat{\beta}$) menggunakan CG linier

Algoritma 3 (Maalouf dan Siddiqi, 2014)

Menghitung $\mathbf{B}(\hat{\boldsymbol{\beta}})$ menggunakan Linear CG, dimana $\mathbf{A} = \mathbf{X}^T\mathbf{D}\mathbf{X} + \lambda\mathbf{I}$, dan $\mathbf{b} = \mathbf{X}^T\mathbf{D}\xi$

Data: $\mathbf{A}, \mathbf{b}, \hat{\beta}^{(0)}$

Result: $\mathbf{B}(\hat{\boldsymbol{\beta}})$ such that $\mathbf{AB}(\hat{\boldsymbol{\beta}}) = \mathbf{b}$

begin

$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{AB}(\hat{\beta})^{(0)}$ /* initialize the residual */

$c = 0$

while $\left\|\mathbf{r}^{(c+1)}\right\|^2 > \varepsilon_3$ **and** $c \leq$ *Max CG Iterations* **do**

  if $c = 0$ **then**

    $\zeta^{(c)} = 0$

  else

    $\zeta^{(c)} = \dfrac{\mathbf{r}^{\mathrm{T}(c+1)}\mathbf{r}^{(c+1)}}{\mathbf{r}^{\mathrm{T}(c)}\mathbf{r}^{(c)}}$ /* update A-Conjugacy enforcer */

  $\mathbf{d}^{(c+1)} = \mathbf{r}^{(c+1)} + \zeta^{(c)}\mathbf{d}^{(c)}$ /* update the serch direction */

  $s^{(c)} = \dfrac{\mathbf{r}^{\mathrm{T}(c)}\mathbf{r}^{(c)}}{\mathbf{d}^{\mathrm{T}(c)}\mathbf{A}\mathbf{d}^{(c)}}$ /* compute the optimal step length */

  $\mathbf{B}(\hat{\beta})^{(c+1)} = \mathbf{B}(\hat{\beta})^{(c)} + \zeta^{(c)}\mathbf{d}^{(c+1)}$ /* obtain an approximate solution */

  $\mathbf{r}^{(c+1)} = \mathbf{r}^{(c)} - s^{(c)}\mathbf{A}\mathbf{d}^{(c+1)}$ /* Update the residual */

  $c = c + 1$

Lampiran 5. Data yang digunakan

No	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
1	0	0	2	1	2	2	1	2	0	2	1	1	0	0	0
2	0	0	3	3	4	2	1	1	0	1	1	1	0	0	0
3	0	0	1	3	1	2	1	2	0	1	3	0	1	0	0
4	0	0	2	2	3	2	1	1	0	1	3	1	1	1	0
5	0	0	3	3	4	2	1	1	0	1	1	1	0	0	0
6	0	0	3	2	4	2	1	1	0	1	1	1	1	0	0
7	1	0	2	4	2	2	1	2	1	2	1	1	1	0	0
8	0	0	2	3	2	2	1	2	0	2	3	1	1	0	0
9	0	0	2	1	3	2	1	2	0	2	3	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.631	0	0	3	4	1	2	1	1	0	1	1	0	0	0	0
5.632	0	0	3	4	4	2	1	1	0	1	3	1	0	0	0
5.633	0	0	2	3	4	2	2	1	1	1	3	1	1	0	0
5.634	0	0	2	3	4	2	2	2	0	1	1	1	0	0	0
5.635	0	0	3	1	4	1	1	1	0	1	3	1	0	0	0
5.636	0	0	1	3	2	2	1	1	1	1	3	1	1	0	0
5.637	0	0	1	2	3	2	1	1	1	1	1	1	1	0	0
5.638	0	0	3	3	2	2	1	1	0	1	3	1	1	0	0
5.639	0	0	3	2	1	2	1	1	0	1	1	0	0	0	0

Lampiran 6. Syntax R untuk model RE-WLR

contoh partisi data 1:5 dan lambda =1

```
{
  for(lambda in 1:1){
    data1 = training5
    data2 = testing5.1
    maxIRLS = 30
    maxCG1 = 200
    maxCG2 = 200
    mf <- model.frame(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+ X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8
+ X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14
,data=data1)
    y <- model.response(mf)
    X <- model.matrix(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+ X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8
+ X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14
,data=data1)

    mft <- model.frame(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+ X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8
+ X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14
,data=data2)
    yt <- model.response(mft)
    Xt <- model.matrix(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+ X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8
+ X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14
,data=data2)
    beta = rep(0, ncol(X))
    Beta = matrix(beta)
    deltadev = 1
    c = 0
    I = diag(ncol(X))
    I[1,1]=0
    tau = 0.052846
    ybar = mean(y)
    w1 = ybar/tau
    w0 = (1 - ybar)/(1 - tau)
    ts=750
    cat("Hasil model klasifikasi RE-WLR dengan lambda ",lambda,
        "sbb:\n")
    while (abs(deltadev) > 0.001 & c <= maxIRLS){
      if (c > 0)Beta = uBeta
      p = matrix(nrow=nrow(X), ncol=1)
      V = matrix(data=0, nrow=nrow(X), ncol=nrow(X))
      W = matrix(data=0, nrow=nrow(X), ncol=nrow(X))
      Z = matrix(nrow=nrow(X), ncol=1)
      D = matrix(data=0, nrow=nrow(X), ncol=nrow(X))
      for (i in 1:nrow(X)){
        p[i] = (exp (X[i,] %*% Beta))/(1 + exp (X[i,] %*% Beta))
        V[i,i] = p[i] * (1-p[i])
        W[i,i] = (w1 * y[i]) + (w0 * (1 - y[i]))
        Z[i] = X[i,] %*% Beta + (y[i] - p[i])/V[i,i]
        D[i,i] = V[i,i] * W[i,i]
      }
      if (c == 0){
        logB = 0
        for (i in 1:nrow(X)){
          logB1 = W[i,i]*(y[i]*log(p[i])+(1-y[i])*log(1-p[i])) + logB
          logB = logB1
        }
        dev = (-2) * (logB - lambda/2 * sum(Beta^2))
      }
      Q = X %*% solve((t(X) %*% D %*% X) + lambda * I )%*% t(X)
      q = diag(Q)
      E = matrix(nrow=nrow(X), ncol=1)
```

```
    for (i in 1:nrow(X)){
      E[i] = 0.5 * q[i] * ((1+w1) * p[i] - w1)
    }
    A = (t(X) %*% D %*% X) + lambda * I
    b = t(X) %*% D %*% Z
    C = t(X) %*% D %*% E
    r = b - A %*% Beta
    t = 0
    sumr2 = sum(r^2)
    d = r
    while (sumr2 > 0.005 & t <= maxCG1){
      Ad = A %*% d
      s = as.numeric((t(r) %*% r)/(t(d) %*% Ad))
      Beta = Beta + s * d
      r1 = r - s * Ad
      a = as.numeric((t(r1) %*% r1)/(t(r) %*% r))
      d1 = r1 + a * d
      sumr2 = sum(r1^2)
      t = t + 1
      r = r1
      d = d1
    }
    if (t>maxCG1) print("Belum konvergen, tambah nilai max iterasi CG1.")
    bbeta = rep(0, length(C))
    BBeta = matrix (bbeta)
    rb = C - A %*% BBeta
    h = 0
    sumrb2 = sum(rb^2)
    db = rb
    while (sumrb2 > 0.005 & h <= maxCG2){
      Adb = A %*% db
      sb = as.numeric((t(rb) %*% rb)/(t(db) %*% Adb))
      BBeta = BBeta + sb * db
      rb1 = rb - sb * Adb
      ab = as.numeric((t(rb1) %*% rb1)/(t(rb) %*% rb))
      db1 = rb1 + ab * db
      sumrb2 = sum(rb1^2)
      h = h + 1
      rb = rb1
      db = db1
    }
    if (h>maxCG2) print("Belum konvergen, tambah nilai max iterasi CG2.")
    uBeta = Beta - BBeta
    logB = 0
    for (i in 1:nrow(X)){
      p[i]=(exp (X[i,] %*% uBeta))/(1 + exp (X[i,] %*% uBeta))
      logB1 = W[i,i]*(y[i]*log(p[i])+(1 - y[i]) * log(1-p[i])) + logB
      logB = logB1
    }
    devold = dev
    dev = (-2) * (logB - lambda/2 * sum(uBeta^2))
    deltadev = (dev - devold)/dev
    c = c + 1
    cat("Iterasi ",c,", dev = ",dev,", deltadev = ",deltadev,"\n")
  }
  if (c>maxIRLS) print("Belum konvergen, tambah nilai max iterasi IRLS.")
  phat = matrix(nrow=nrow(X), ncol=1)
  yhat = matrix(nrow=nrow(X), ncol=1)
  fitt = matrix(nrow=nrow(X), ncol=1)
  for (k in 1:nrow(X)){
    phat[k]=(exp (X[k,] %*% uBeta))/(1 + exp (X[k,] %*% uBeta))
    if (phat[k] < 0.5)yhat[k]=0
    else yhat[k] = 1
    if (yhat[k] == y[k]) fitt[k] = 1
    else fitt[k] = 0
  }
  tepat1 = mean(fitt)
  pthat = matrix(nrow=nrow(Xt), ncol=1)
  ythat = matrix(nrow=nrow(Xt), ncol=1)
  fit = matrix(nrow=nrow(Xt), ncol=1)
```

```
    for (k in 1:nrow(Xt)){
      pthat[k]=(exp (Xt[k,] %*% uBeta))/(1 + exp (Xt[k,] %*% uBeta))
      if (pthat[k] < 0.5)ythat[k]=0
      else ythat[k] = 1
      if (ythat[k] == yt[k]) fit[k] = 1
      else fit[k] = 0
    }
    VarUbeta=(n/(n+ncol(X)))^2*solve(t(X)%*%D%*%X)
    SEbeta=sqrt(diag(VarUbeta))
    SEBeta=matrix(SEbeta)
    t.hit=(uBeta/SEBeta)^2
    tepat = mean(fit)
    cat("Hasil estimasi parameter Beta:\n")
    p.value=c()
    for (d in 1:ncol(Xt)){
    p.value[d]=pchisq(t.hit[d],1,lower.tail = F)}
    hasil=data.frame(beta=uBeta,se=
SEbeta,t.hit=round(t.hit,4),p.value=round(p.value,4),expBeta=round(exp(
uBeta),3))
    list(hasil=hasil)
    print(hasil)
    cat("Jumlah iterasi = ",c,"\n")
    cat("Confusion matrix Data Training:\n")
    library(crossval)
    cm1 = confusionMatrix(y, yhat, negative=0)
    print(cm1)
    de1 = diagnosticErrors(cm1)
    print(de1)
    cat("Ketepatan klasifikasi RE-WLR Data Training dengan lambda ",
lambda, "yaitu ", tepat1,"\n")
    library(verification)
    rocArea1= roc.area(y, yhat)
    cat("ROC Area Data Training sbb:\n")
    print(rocArea1)
    cat("Confusion matrix Data Testing:\n")
    cm = confusionMatrix(yt, ythat, negative=0)
    print(cm)
    de = diagnosticErrors(cm)
    print(de)
    cat("Ketepatan klasifikasi RE-WLR Data Testing dengan lambda ",
lambda, "yaitu ", tepat,"\n")
    rocArea= roc.area(yt, ythat)
    cat("ROC Area Data Testing sbb:\n")
    print(rocArea)
    cat("-----------------------------------------------------\n")
    lambda = lambda + 1
  }
}
```

Lampiran 7. Syntax R untuk model TR-PC

contoh partisi data 1:5 dan lambda =0.01

```
{
  for(lambda in 0.01:1){
    data1 = training5
    data2 = testing5.1
    maxIRLS = 30
    maxCG1 = 200
    maxCG2 = 200
    mf <- model.frame(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+ X4.1 + X4.2 + X4.3 + X5.1+X5.2+ X6.1 + X6.2 + X7.1 + X7.2 + X8 + X9.1
+ X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14,data=data1)
    y <- model.response(mf)
    X <- model.matrix(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+  X4.1 + X4.2 + X4.3 +X5.1+X5.2+ X6.1 + X6.2 + X7.1 + X7.2 + X8 + 9.1
+ X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14 ,data=data1)

    mft <- model.frame(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+  X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8 +
X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14,data=data2)
    yt <- model.response(mft)
    Xt <- model.matrix(Y ~ X1 + X2.1 + X2.2 + X2.3 + X3.1 + X3.2 + X3.3
+  X4.1 + X4.2 + X4.3 + X5.1 + X5.2 + X6.1 + X6.2 + X7.1 + X7.2 + X8 +
X9.1 + X9.2 + X10.1 + X10.2 + X10.3 + X11 + X12 + X13 + X14,data=data2)

    beta = rep(0, ncol(X))
    Beta = matrix(beta)
    deltadev = 1
    c = 0
    I = diag(ncol(X))
    I[1,1]=0
    tau = 0.052846
    ybar = mean(y)
    delta1 = ybar/tau
    delta0 = (1 - ybar)/(1 - tau)
    ts=750
    cat("Hasil model klasifikasi TR-PC dengan lambda", lambda,"sbb:\n")
    while (abs(deltadev) > 0.001 & c <= maxIRLS){
      if (c > 0)Beta = uBeta
      p = matrix(nrow=nrow(X), ncol=1)
      V = matrix(data=0, nrow=nrow(X), ncol=nrow(X))
      Z = matrix(nrow=nrow(X), ncol=1)
      D = matrix(data=0, nrow=nrow(X), ncol=nrow(X))
      for (i in 1:nrow(X)){
        p[i] = 1/(1 + exp (log(delta1/delta0)-X[i,] %*% Beta))
        V[i,i] = p[i] * (1-p[i])
        Z[i] = X[i,] %*% Beta + (y[i] - p[i])/V[i,i]
        D[i,i] = V[i,i]
      }
      if (c == 0){
        logB = 0
        for (i in 1:nrow(X)){
          logB1 = (y[i] * log(p[i]) + (1 - y[i]) * log(1-p[i])) + logB
          logB = logB1
        }
        dev = (-2) * (logB - lambda/2 * sum(Beta^2))
      }
      Q = X %*% solve((t(X) %*% D %*% X) + lambda * I )%*% t(X)
      q = diag(Q)
      E = matrix(nrow=nrow(X), ncol=1)
      for (i in 1:nrow(X)){
        E[i] = q[i] * (p[i] - 0.5)
      }
      A = (t(X) %*% D %*% X) + lambda * I
      b = t(X) %*% D %*% Z
      C = t(X) %*% D %*% E
      r = b - A %*% Beta
      t = 0
```

```
    sumr2 = sum(r^2)
    d = r
    while (sumr2 > 0.0005 & t <= maxCG1){
      Ad = A %*% d
      s = as.numeric((t(r) %*% r)/(t(d) %*% Ad))
      Beta = Beta + s * d
      r1 = r - s * Ad
      a = as.numeric((t(r1) %*% r1)/(t(r) %*% r))
      d1 = r1 + a * d
      sumr2 = sum(r1^2)
      t = t + 1
      r = r1
      d = d1
    }
    if (t>maxCG1) print("Belum konvergen, tambah nilai max iterasi CG1.")
    bbeta = rep(0, length(C))
    BBeta = matrix (bbeta)
    rb = C - A %*% BBeta
    h = 0
    sumrb2 = sum(rb^2)
    db = rb
    while (sumrb2 > 0.0005 & h <= maxCG2){
      Adb = A %*% db
      sb = as.numeric((t(rb) %*% rb)/(t(db) %*% Adb))
      BBeta = BBeta + sb * db
      rb1 = rb - sb * Adb
      ab = as.numeric((t(rb1) %*% rb1)/(t(rb) %*% rb))
      db1 = rb1 + ab * db
      sumrb2 = sum(rb1^2)
      h = h + 1
      rb = rb1
      db = db1
    }
    if (h>maxCG2) print("Belum konvergen, tambah nilai max iterasi CG2.")
    uBeta = Beta - BBeta
    logB = 0
    for (i in 1:nrow(X)){
      p[i]=1/(1 + exp (log(delta1/delta0)-X[i,] %*% uBeta))
      logB1 = (y[i] * log(p[i]) + (1 - y[i]) * log(1-p[i])) + logB
      logB = logB1
    }
    devold = dev
    dev = (-2) * (logB - lambda/2 * sum(uBeta^2))
    deltadev = (dev - devold)/dev
    c = c + 1
    cat("Iterasi ",c,", dev = ",dev,", deltadev = ",deltadev,"\n")
  }
  if (c>maxIRLS) print("Belum konvergen, tambah nilai max iterasi IRLS.")
  VarUbeta=(n/(n+ncol(X)))^2*solve(t(X)%*%D%*%X)
  phat = matrix(nrow=nrow(X), ncol=1)
  yhat = matrix(nrow=nrow(X), ncol=1)
  fitt = matrix(nrow=nrow(X), ncol=1)
  for (k in 1:nrow(X)){
    phat[k]=(1)/(1 + exp (log(delta1/delta0)- X[k,] %*% uBeta))
    if (phat[k] < 0.5)yhat[k]=0
    else yhat[k] = 1
    if (yhat[k] == y[k]) fitt[k] = 1
    else fitt[k] = 0
  }

  tepat1 = mean(fitt)
  pthat = matrix(nrow=nrow(Xt), ncol=1)
  fit = matrix(nrow=nrow(Xt), ncol=1)
  pthatC = matrix(nrow=nrow(Xt), ncol=1)
```

```
    for (k in 1:nrow(Xt)){
      pthat[k]=1/(1 + exp (log(delta1/delta0)-Xt[k,] %*% uBeta))
      if (pthat[k] < 0.5)ythat[k]=0
      else ythat[k] = 1
      if (ythat[k] == yt[k]) fit[k] = 1
      else fit[k] = 0
    }

    SEbeta=sqrt(diag(VarUbeta))
    SEBeta=matrix(SEbeta)
    t.hit=(uBeta/SEBeta)^2
    tepat = mean(fit)
    cat("Hasil estimasi parameter Beta:\n")
    p.value=c()
    for (d in 1:ncol(Xt)){
    p.value[d]=pchisq(t.hit[d],1,lower.tail = F)}
    hasil=data.frame(beta=uBeta,se= SEbeta,t.hit=round(t.hit,4),
    p.value=round(p.value,4), expBeta=round(exp(uBeta),3))
    list(hasil=hasil)
    print(hasil)
    cat("Jumlah iterasi = ",c,"\n")
    cat("Confusion matrix Data Training:\n")
    library(crossval)
    cm1 = confusionMatrix(y, yhat, negative=0)
    print(cm1)
    de1 = diagnosticErrors(cm1)
    print(de1)
    cat("Ketepatan klasifikasi TR-PC dengan lambda ", lambda,
        "yaitu ", tepat1,"\n")
    library(verification)
    rocArea1= roc.area(y, yhat)
    cat("ROC Area Data Training sbb:\n")
    print(rocArea1)
    cat("Confusion matrix Data Testing:\n")
    cm = confusionMatrix(yt, ythat, negative=0)
    print(cm)
    de = diagnosticErrors(cm)
    print(de)
    cat("Ketepatan klasifikasi TR-PC Data Testing dengan lambda ",
        lambda,    "yaitu ", tepat,"\n")
    rocArea= roc.area(yt, ythat)
    cat("ROC Area Data Testing sbb:\n")
    print(rocArea)
    cat("------------------------------------------------------------\n")
    lambda = lambda + 1
  }
}
```

Lampiran 8. Contoh Output R untuk model RE-WLR

contoh partisi data 1:5 dan lambda =1

```
Hasil model klasifikasi RE-WLR data partisi  150 : 750  dengan lambda
1 sbb:
Iterasi  1 , dev =  952.3592 , deltadev =  -0.6491616
Iterasi  2 , dev =  879.6271 , deltadev =  -0.08268518
Iterasi  3 , dev =  870.7937 , deltadev =  -0.01014409
Iterasi  4 , dev =  871.0766 , deltadev =  0.0003248273
Hasil estimasi parameter Beta:
                   beta        se   t.hit p.value expBeta
(Intercept) -4.85759599 0.6570054 54.6644  0.0000   0.008
X1           0.47253509 0.2819504  2.8088  0.0937   1.604
X2.1         0.19275905 0.2339307  0.6790  0.4099   1.213
X2.2        -0.36963359 0.2710163  1.8602  0.1726   0.691
X2.3         0.10505643 0.2968534  0.1252  0.7234   1.111
X3.1         0.27029216 0.2844587  0.9029  0.3420   1.310
X3.2         0.81590337 0.2795464  8.5186  0.0035   2.261
X3.3         1.82407109 0.2757956 43.7430  0.0000   6.197
X4.1         0.83109711 0.5646097  2.1667  0.1410   2.296
X4.2         0.76920602 0.5903323  1.6978  0.1926   2.158
X4.3         0.45589011 0.5870486  0.6031  0.4374   1.578
X5.1        -1.11200527 0.2182370 25.9631  0.0000   0.329
X5.2         0.87868091 0.2761584 10.1239  0.0015   2.408
X6.1         0.02442493 0.1879280  0.0169  0.8966   1.025
X6.2         1.29128709 0.7857236  2.7009  0.1003   3.637
X7.1         0.40949365 0.1784064  5.2683  0.0217   1.506
X7.2         1.11015336 0.4349232  6.5154  0.0107   3.035
X8           1.75470857 0.1840112 90.9330  0.0000   5.782
X9.1         0.17589830 0.1965186  0.8012  0.3707   1.192
X9.2         0.69060695 0.2443186  7.9900  0.0047   1.995
X10.1        0.01408284 0.2428964  0.0034  0.9538   1.014
X10.2        0.06079031 0.2285796  0.0707  0.7903   1.063
X10.3        0.27559285 0.3181860  0.7502  0.3864   1.317
X11          1.52302809 0.2995384 25.8530  0.0000   4.586
X12          1.28806948 0.2507980 26.3773  0.0000   3.626
X13          1.51184953 0.2524142 35.8748  0.0000   4.535
X14         -0.95042882 0.1991788 22.7695  0.0000   0.387
Jumlah iterasi =  4
Confusion matrix Data Training:
 FP  TP  TN  FN
107 115 643  35
attr(,"negative")
[1] 0
      acc      sens      spec       ppv       npv       lor
0.8422222 0.7666667 0.8573333 0.5180180 0.9483776 2.9829000
attr(,"negative")
[1] 0
Ketepatan klasifikasi RE-WLR Data Training dengan lambda  1 yaitu
0.8422222
ROC Area Data Training sbb:
$A
[1] 0.812

$n.total
[1] 900

$n.events
[1] 150

$n.noevents
[1] 750

$p.value
[1] 3.787437e-59
```

```
Confusion matrix Data Testing:
  FP   TP   TN   FN
 223   56 1186   18
attr(,"negative")
[1] 0
      acc      sens      spec       ppv       npv       lor
0.8374916 0.7567568 0.8417317 0.2007168 0.9850498 2.8061497
attr(,"negative")
[1] 0
Ketepatan klasifikasi RE-WLR Data Testing dengan lambda  1 yaitu
0.8374916
ROC Area Data Testing sbb:
$A
[1] 0.7992442

$n.total
[1] 1483

$n.events
[1] 74

$n.noevents
[1] 1409

$p.value
[1] 5.154669e-38

------------------------------------------------------------------------
```

Lampiran 9. Contoh Output R untuk model TR-PC

contoh partisi data 1:5 dan lambda =0.01

```
Hasil model klasifikasi TR-PC data partisi  150 : 750  dengan lambda
0.01 sbb:
Iterasi  1 , dev =  537.5448 , deltadev =  -0.5364212
Iterasi  2 , dev =  480.8383 , deltadev =  -0.1179326
Iterasi  3 , dev =  466.6416 , deltadev =  -0.03042314
Iterasi  4 , dev =  463.5364 , deltadev =  -0.006698875
Iterasi  5 , dev =  462.9189 , deltadev =  -0.001334016
Iterasi  6 , dev =  462.8109 , deltadev =  -0.0002334259
Hasil estimasi parameter Beta:
                  beta        se   t.hit p.value expBeta
(Intercept) -5.85195987 1.3479292 18.8481  0.0000   0.003
X1           0.56529767 0.3814291  2.1965  0.1383   1.760
X2.1         0.08057826 0.3249507  0.0615  0.8042   1.084
X2.2        -0.39217765 0.3887406  1.0178  0.3131   0.676
X2.3         0.06279136 0.4053059  0.0240  0.8769   1.065
X3.1         0.38678897 0.4228683  0.8366  0.3604   1.472
X3.2         1.02250081 0.4059799  6.3433  0.0118   2.780
X3.3         2.09108692 0.4000705 27.3194  0.0000   8.094
X4.1         1.77317232 1.2615603  1.9755  0.1599   5.890
X4.2         1.74329310 1.2830856  1.8460  0.1743   5.716
X4.3         1.25425046 1.2774307  0.9640  0.3262   3.505
X5.1        -1.07909463 0.2964228 13.2524  0.0003   0.340
X5.2         1.09769049 0.3794315  8.3694  0.0038   2.997
X6.1         0.02615322 0.2577498  0.0103  0.9192   1.026
X6.2         2.63462181 1.2865610  4.1935  0.0406  13.938
X7.1         0.39506774 0.2555913  2.3892  0.1222   1.484
X7.2         1.35913372 0.5724510  5.6370  0.0176   3.893
X8           1.65799294 0.2555412 42.0963  0.0000   5.249
X9.1         0.17395098 0.2787484  0.3894  0.5326   1.190
X9.2         0.87522197 0.3232786  7.3296  0.0068   2.399
X10.1       -0.04598019 0.3444090  0.0178  0.8938   0.955
X10.2       -0.02634752 0.3191575  0.0068  0.9342   0.974
X10.3        0.18799790 0.4285816  0.1924  0.6609   1.207
X11          1.46151883 0.4863367  9.0310  0.0027   4.313
X12          1.33011487 0.3954323 11.3145  0.0008   3.781
X13          1.51892340 0.3341540 20.6623  0.0000   4.567
X14         -0.92491293 0.2729598 11.4817  0.0007   0.397
Jumlah iterasi =  6
Confusion matrix Data Training:
 FP  TP  TN  FN
 34  80 716  70
attr(,"negative")
[1] 0
      acc      sens      spec       ppv       npv       lor
0.8844444 0.5333333 0.9546667 0.7017544 0.9109415 3.1808510
attr(,"negative")
[1] 0
Ketepatan klasifikasi TR-PC Data Training ( 150 : 750 ) dan lambda
0.01 yaitu  0.8844444
ROC Area Data Training sbb:
$A
[1] 0.744

$n.total
[1] 900

$n.events
[1] 150

$n.noevents
[1] 750

$p.value
[1] 1.042794e-60
```

```
Confusion matrix Data Testing:
  FP   TP   TN   FN
  60   35 1349   39
attr(,"negative")
[1] 0
      acc      sens      spec       ppv       npv       lor
0.9332434 0.4729730 0.9574166 0.3684211 0.9719020 3.0045607
attr(,"negative")
[1] 0
Ketepatan klasifikasi TR-PC Data Testing dengan lambda  0.01 yaitu
0.9332434
ROC Area Data Testing sbb:
$A
[1] 0.7151948

$n.total
[1] 1483

$n.events
[1] 74

$n.noevents
[1] 1409

$p.value
[1] 1.978897e-49

------------------------------------------------------------------------
```

Lampiran 10. Hasil Klasifikasi Model RE-WLR dan TR-PC Data Training

lambda	partisi data	RE-WLR					TR-PC				
		akurasi	sens	spe	gmean	auc	akurasi	sens	spe	gmean	auc
0.01	1:3	74.00%	96.00%	66.67%	80.00%	0.813	85.33%	62.00%	93.11%	75.98%	0.776
	1:4	79.20%	86.67%	77.33%	81.87%	0.820	86.13%	53.33%	94.33%	70.93%	0.738
	1:5	83.67%	76.00%	85.20%	80.47%	0.806	88.44%	53.33%	95.47%	71.36%	0.744
	1:6	86.48%	71.33%	89.00%	79.68%	0.802	89.71%	46.67%	96.89%	67.24%	0.718
	1:10	92.30%	52.67%	96.27%	71.20%	0.745	93.09%	38.00%	98.60%	61.21%	0.683
	rata2	83.13%	76.53%	82.89%	78.64%	0.797	88.54%	50.67%	95.68%	69.34%	0.732
0.1	1:3	74.00%	96.00%	66.67%	80.00%	0.813	85.00%	60.67%	93.11%	75.16%	0.769
	1:4	79.20%	86.67%	77.33%	81.87%	0.820	86.13%	52.67%	94.50%	70.55%	0.736
	1:5	83.67%	76.00%	85.20%	80.47%	0.806	88.44%	53.33%	95.47%	71.36%	0.744
	1:6	86.48%	71.33%	89.00%	79.68%	0.802	89.71%	46.67%	96.89%	67.24%	0.718
	1:10	92.24%	52.00%	96.27%	70.75%	0.741	93.03%	37.33%	98.60%	60.67%	0.680
	rata2	83.12%	76.40%	82.89%	78.55%	0.796	88.46%	50.13%	95.71%	68.99%	0.729
1	1:3	74.00%	96.00%	66.67%	80.00%	0.813	84.50%	54.67%	94.44%	71.85%	0.746
	1:4	79.47%	86.00%	77.83%	81.81%	0.819	86.40%	51.33%	95.17%	69.89%	0.733
	1:5	84.22%	76.67%	85.73%	81.07%	0.812	88.89%	47.33%	97.20%	67.83%	0.723
	1:6	86.57%	70.67%	89.22%	79.40%	0.799	89.52%	44.00%	97.11%	65.37%	0.706
	1:10	92.42%	48.67%	96.80%	68.64%	0.727	92.79%	33.33%	98.73%	57.37%	0.660
	rata2	83.34%	75.60%	83.25%	78.19%	0.794	88.42%	46.13%	96.53%	66.46%	0.713
2	1:3	73.00%	95.33%	65.56%	79.05%	0.804	83.83%	51.33%	94.67%	69.71%	0.730
	1:4	79.07%	86.00%	77.33%	81.55%	0.817	86.00%	46.67%	95.83%	66.87%	0.713
	1:5	84.00%	77.33%	85.33%	81.23%	0.813	89.00%	45.33%	97.73%	66.56%	0.715
	1:6	86.67%	70.67%	89.33%	79.45%	0.800	89.43%	42.67%	97.22%	64.41%	0.699
	1:10	92.30%	46.00%	96.93%	66.78%	0.715	92.67%	30.67%	98.87%	55.06%	0.648
	rata2	83.01%	75.07%	82.90%	77.61%	0.790	88.19%	43.33%	96.86%	64.52%	0.701
3	1:3	72.00%	95.33%	64.22%	78.25%	0.798	84.50%	52.00%	95.33%	70.41%	0.737
	1:4	79.07%	86.00%	77.33%	81.55%	0.817	85.07%	42.00%	95.83%	63.44%	0.689
	1:5	83.89%	77.33%	85.20%	81.17%	0.813	93.80%	35.14%	96.88%	58.34%	0.660
	1:6	86.86%	70.67%	89.56%	79.55%	0.801	89.52%	41.33%	97.56%	63.50%	0.694
	1:10	92.42%	44.00%	97.27%	65.42%	0.706	92.48%	28.00%	98.93%	52.63%	0.635
	rata2	82.85%	74.67%	82.72%	77.19%	0.787	89.07%	39.69%	96.91%	61.67%	0.683
4	1:3	71.50%	95.33%	63.56%	77.84%	0.794	84.83%	52.00%	95.78%	70.57%	0.739
	1:4	79.07%	86.67%	77.17%	81.78%	0.819	85.33%	41.33%	96.33%	63.10%	0.688
	1:5	84.11%	77.33%	85.47%	81.30%	0.814	94.00%	32.43%	97.23%	56.16%	0.648
	1:6	87.05%	70.67%	89.78%	79.65%	0.802	89.81%	40.00%	98.11%	62.65%	0.691
	1:10	92.55%	42.67%	97.53%	64.51%	0.701	92.42%	26.00%	99.07%	50.75%	0.625
	rata2	82.85%	74.53%	82.70%	77.02%	0.786	89.28%	38.35%	97.30%	60.65%	0.678
5	1:3	71.67%	95.33%	63.78%	77.98%	0.796	84.67%	51.33%	95.78%	70.12%	0.736
	1:4	78.80%	87.33%	76.67%	81.83%	0.820	85.20%	40.67%	96.33%	62.59%	0.685
	1:5	83.78%	77.33%	85.07%	81.11%	0.812	88.56%	40.00%	98.27%	62.70%	0.691
	1:6	87.05%	70.00%	89.89%	79.32%	0.799	89.81%	39.33%	98.22%	62.16%	0.688
	1:10	92.67%	42.67%	97.67%	64.55%	0.702	92.42%	25.33%	99.13%	50.11%	0.622
	rata2	82.79%	74.53%	82.61%	76.96%	0.786	88.13%	39.33%	97.55%	61.53%	0.684

Lampiran 11. Rata-Rata Hasil Klasifikasi Model RE-WLR dan TR-PC Pada Data Testing

lambda	partisi data	RE-WLR					TR-PC				
		akurasi	sens	spe	gmean	auc	akurasi	sens	spe	gmean	auc
0.01	1:3	70.40%	87.70%	69.49%	78.06%	0.786	89.92%	54.86%	91.76%	70.95%	0.733
	1:4	79.24%	80.54%	79.18%	80.13%	0.799	92.43%	52.84%	94.51%	70.63%	0.737
	1:5	83.90%	82.43%	83.97%	83.12%	0.832	93.34%	57.03%	95.25%	73.65%	0.761
	1:6	87.19%	69.32%	88.13%	78.15%	0.787	93.33%	43.11%	95.97%	64.27%	0.695
	1:10	92.57%	49.59%	94.83%	68.54%	0.722	94.64%	39.46%	97.54%	61.98%	0.685
0.1	1:3	70.37%	87.43%	69.47%	77.93%	0.785	89.94%	54.46%	91.80%	70.70%	0.731
	1:4	79.48%	80.00%	79.45%	79.98%	0.797	92.57%	51.76%	94.71%	69.97%	0.732
	1:5	84.00%	82.43%	84.08%	83.17%	0.833	93.41%	56.62%	95.34%	73.42%	0.760
	1:6	87.17%	70.95%	88.02%	79.00%	0.795	93.36%	42.30%	96.04%	63.69%	0.692
	1:10	92.60%	49.59%	94.85%	68.55%	0.722	94.61%	38.65%	97.55%	61.36%	0.681
1	1:3	69.93%	88.78%	68.94%	78.23%	0.789	90.76%	48.38%	92.98%	67.05%	0.707
	1:4	79.74%	80.41%	79.71%	80.05%	0.801	92.77%	45.14%	95.27%	65.47%	0.702
	1:5	84.05%	82.84%	84.12%	83.46%	0.835	93.63%	50.41%	95.90%	69.45%	0.732
	1:6	87.44%	68.92%	88.41%	78.03%	0.787	93.85%	40.00%	96.68%	62.13%	0.683
	1:10	92.93%	47.97%	95.29%	67.58%	0.716	94.79%	36.49%	97.85%	59.72%	0.672
2	1:3	69.49%	90.54%	68.39%	78.68%	0.795	91.18%	44.86%	93.61%	64.79%	0.692
	1:4	79.69%	83.24%	79.50%	81.34%	0.814	92.80%	40.81%	95.53%	62.31%	0.682
	1:5	83.91%	81.35%	84.05%	82.68%	0.827	94.17%	50.14%	96.49%	69.47%	0.733
	1:6	87.48%	67.43%	88.54%	77.25%	0.780	94.15%	40.00%	97.00%	62.23%	0.685
	1:10	93.08%	44.73%	95.62%	65.37%	0.702	94.96%	34.19%	98.15%	57.90%	0.662
3	1:3	69.18%	91.08%	68.03%	78.71%	0.796	91.22%	43.65%	93.72%	63.93%	0.687
	1:4	79.55%	85.41%	79.24%	82.26%	0.823	93.23%	41.35%	95.95%	62.85%	0.687
	1:5	83.99%	80.68%	84.16%	82.39%	0.824	94.46%	47.30%	96.94%	67.61%	0.721
	1:6	87.39%	66.22%	88.50%	76.53%	0.774	94.36%	40.00%	97.21%	62.30%	0.686
	1:10	93.22%	43.38%	95.84%	64.45%	0.696	94.96%	32.16%	98.26%	56.17%	0.652
4	1:3	68.42%	92.57%	67.15%	78.83%	0.799	91.50%	42.84%	94.05%	63.45%	0.684
	1:4	79.39%	85.41%	79.08%	82.17%	0.822	93.45%	40.54%	96.22%	62.31%	0.684
	1:5	84.03%	79.73%	84.26%	81.95%	0.820	94.59%	45.54%	97.16%	66.40%	0.714
	1:6	87.42%	64.86%	88.61%	75.78%	0.759	94.64%	39.05%	97.56%	61.67%	0.683
	1:10	93.36%	42.57%	96.03%	63.91%	0.692	95.11%	27.97%	98.64%	52.46%	0.633
5	1:3	68.11%	92.97%	66.81%	78.81%	0.799	91.60%	40.81%	94.27%	62.00%	0.675
	1:4	79.16%	85.41%	78.84%	82.05%	0.821	93.66%	38.78%	96.54%	61.03%	0.677
	1:5	84.11%	78.78%	84.39%	81.53%	0.816	94.70%	43.78%	97.37%	65.18%	0.706
	1:6	87.67%	64.32%	88.89%	75.58%	0.766	94.73%	37.97%	97.71%	60.87%	0.678
	1:10	93.56%	42.57%	96.24%	63.98%	0.694	95.17%	27.30%	98.74%	51.85%	0.630

halaman ini sengaja dikosongkan

BIOGRAFI PENULIS

Penulis dilahirkan pada tanggal 25 Januari 1988 di Lumajang, Jawa Timur, sebagai anak ketiga dari empat bersaudara dari pasangan Bapak Eko Bayu Kusnanto dan Ibu Sulasih. Penulis menempuh pendidikan formal di SDN Ranuyoso 1 (1994-2000), SLTPN 1 Sukodono (2000-2003) dan SMAN 2 Lumajang (2003-2006). Selanjutnya penulis melanjutkan pendidikan di Sekolah Tinggi Ilmu Statistik (STIS) Jakarta (2006-2010) Jurusan Statistik Sosial Kependudukan. Setelah menyelesaikan pendidikan DIV di STIS, penulis ditugaskan bekerja di BPS Kabupaten Tabanan dan BPS Kabupaten Badung. Pada tahun 2016 penulis memperoleh kesempatan untuk melanjutkan pendidikan Pasca Sarjana Departemen Statistika, Fakultas Matematika, Komputasi dan Sains Data (FMKSD), Institut Teknologi Sepuluh Nopember (ITS) Surabaya. Pembaca yang ingin memberikan kritik, saran dan pertanyaan mengenai penelitian ini dapat menghubungi melalui email sony.trias@gmail.com .