TUGAS AKHIR – KS184822

ANALISIS *IMBALANCED MULTICLASS* PADA STATUS KEPEMILIKAN ASURANSI DENGAN METODE *MULTINOMIAL LOGISTIC REGRESSION*

ROSIKHU ILMI HIDAYATI
NRP 062115 4000 0077

Dosen Pembimbing
Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.
Santi Puteri Rahayu, M.Si., Ph.D

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

TUGAS AKHIR – KS184822

ANALISIS *IMBALANCED MULTICLASS* PADA STATUS KEPEMILIKAN ASURANSI DENGAN METODE *MULTINOMIAL LOGISTIC REGRESSION*

ROSIKHU ILMI HIDAYATI
NRP 062115 4000 0077

Dosen Pembimbing
Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.
Santi Puteri Rahayu, M.Si., Ph.D

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

(Halaman ini sengaja dikosongkan)

FINAL PROJECT – KS184822

ANALYSIS IMBALANCED MULTICLASS ON INSURANCE STATUS USING MULTINOMIAL LOGISTIC REGRESSION METHOD

ROSIKHU ILMI HIDAYATI
NRP 062115 4000 0077

Supervisors
Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.
Santi Puteri Rahayu, M.Si., Ph.D

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

(Halaman ini sengaja dikosongkan)

LEMBAR PENGESAHAN

ANALISIS *IMBALANCED MULTICLASS* PADA STATUS KEPEMILIKAN ASURANSI DENGAN METODE *MULTINOMIAL LOGISTIC REGRESSION*

TUGAS AKHIR
Diajukan untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Statistika
pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :
Rosikhu Ilmi Hidayati
NRP. 062115 4000 0077

Disetujui oleh Pembimbing:
Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.
NIP. 19820326 200312 1 004
Santi Puteri Rahayu, M.Si., Ph.D.
NIP. 19750115 199903 2 003

Mengetahui,
Kepala Departemen Statistika

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2019

(Halaman ini sengaja dikosongkan)

ANALISIS *IMBALANCED MULTICLASS* PADA STATUS KEPEMILIKAN ASURANSI DENGAN METODE *MULTINOMIAL LOGISTIC REGRESSION*

Nama Mahasiswa : Rosikhu Ilmi Hidayati
NRP : 062115 4000 0077
Departemen : Statistika, FMKSD-ITS
Pembimbing : Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.
Santi Puteri Rahayu, M.Si., Ph.D

Abstrak

Masalah kesehatan di Indonesia sangat bervariasi dan untuk mengatasinya pemerintah menggalakkan program asuransi. Saat ini asuransi berkembang pesat akan tetapi perkembangan itu tidak diiringi oleh pemahaman masyarakat akan pentingnya asuransi. Hal itu tampak masih minimnya jumlah nasabah yang terdaftar sebagai peserta asuransi. Tidak seperti negara maju di barat dimana asuransi menjadi kewajiban bagi seluruh warganya. Mengingat asuransi cukup penting maka dilakukan analisis status kepemilikan asuransi dengan metode regresi logistik multinomial. Untuk mengatasi kasus imbalanced data dimana status kepemilikan asuransi cenderung pada satu kategori, maka dilakukan balancing data dengan SMOTE, random undersampling dan combine sampling. Hasil analisis menunjukkan bahwa analisis dengan balancing undersampling lebih baik dibandingkan metode yang lain karena model yang dihasilkan fit dengan nilai AUC sebesar 54,78%, dimana status kepemilikan asuransi dipengaruhi oleh lokasi tempat tinggal dan frekuensi rawat jalan. Diharapkan hasil analisis ini dapat memberikan informasi bagi pemerintah dalam upaya meningkatkan SDM dibidang kesehatan.

Kata kunci : Asuransi, AUC, Imbalanced Data, Regresi Logistik Multinomial.

(Halaman ini sengaja dikosongkan)

ANALYSIS IMBALANCED MULTICLASS ON INSURANCE STATUS USING MULTINOMIAL LOGISTIC REGRESSION METHOD

Name	**: Rosikhu Ilmi Hidayati**
Student Number	**: 062115 4000 0077**
Department	**: Statistics, FMCDS-ITS**
Supervisors	**: Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.**
	Santi Puteri Rahayu, M.Si., Ph.D

Abstract

Health problems in Indonesia are very diverse and to overcome them the government promotes insurance programs. Currently, insurance is growing rapidly but the development is not accompanied by a public understanding of the importance of insurance. It seems that there is still a minimum number of customers registered as insurance participants. Unlike developed countries in the west where insurance is an obligation for all citizens. Given that insurance is important, an analysis of insurance ownership status is carried out with multinomial logistic regression method. To overcome imbalanced data where insurance ownership status tends to be in one category, preprocessing is done with SMOTE, undersampling and combine sampling. The analysis shows that balancing data with undersampling is better than the other methods because the model is fitted with AUC 54,78%, where the ownership status of insurance is influenced by the location of residence and frequency of outpatient care. It is expected that the results of this analysis can provide information for the government in an effort to improve human resources in the health sector.

Keywords : AUC, Imbalanced Data, Insurance, Multinomial Logistic Regression.

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Puji syukur penulis panjatkan atas rahmat dan hidayah yang diberikan Allah SWT sehingga penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul "Analisis *Imbalanced Multiclass* pada Status Kepemilikan Asuransi dengan Metode *Multinomial Logistic Regression*" dengan lancar.

Penulis menyadari bahwa Tugas Akhir ini dapat terselesaikan tidak terlepas dari bantuan dan dukungan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Kedua orang tua, atas segala do'a, nasehat, kasih sayang, dan dukungan yang diberikan kepada penulis demi kesuksesan dan kebahagiaan penulis.
2. Dr. Suhartono selaku Kepala Departemen Statistika dan Dr. Santi Wulan Purnami, S.Si., M.Si. selaku Ketua Program Studi Sarjana yang telah memberikan fasilitas, sarana, dan prasarana.
3. Dr. Sutikno, M.Si selaku dosen-dosen yang menjadi dosen wali selama masa studi yang telah banyak memberikan saran dan arahan dalam proses belajar di Departemen Statistika.
4. Dr.rer.pol. Heri Kuswanto, S.Si., M.Si. selaku dosen pembimbing dan Santi Puteri Rahayu, M.Si., Ph.D selaku co pembimbing yang telah meluangkan waktu dan dengan sangat sabar memberikan bimbingan,saran,dukungan serta motivasi selama penyusunan Tugas Akhir.
5. Dr. Purhadi dan Dr. Moh. Atok, M.Si selaku dosen penguji yang selalu sabar dalam mengomentari serta memberikan masukan dan saran dalam penyelesaian Tugas Akhir.
6. Seluruh dosen Statistika ITS yang telah memberikan ilmu dan pengetahuan yang tak ternilai harganya, serta segenap karyawaan Departemen Statistika ITS.
7. Teman-teman Statistika ITS ∑26 angkatan 2015, yang selalu memberikan dukungan kepada penulis selama ini.

8. Semua teman, relasi dan berbagai pihak yang tidak bisa penulis sebutkan namanya satu persatu yang telah membantu dalam penulisan laporan ini.

Besar harapan penulis untuk mendapatkan kritik dan saran yang membangun sehingga Tugas Akhir ini dapat memberikan manfaat bagi semua pihak yang terkait.

Surabaya, Juli 2019

Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

(Halaman ini sengaja dikosongkan)

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kesehatan merupakan hal yang sangat penting dimana kesehatan menjadi tanggung jawab dari setiap orang, masyarakat, pemerintah dan swasta. Kualitas sumber daya manusia suatu negara dapat dilihat melalui Indeks Pembangunan Manusia (IPM) yang dikembangkan oleh United Nations Development Program (UNDP) sejak tahun 1980. Pembangunan manusia ialah proses memperbanyak pilihan masyarakat, terutama pilihan untuk menjalani umur panjang dan sehat, memperoleh pendidikan, menikmati standar hidup yang layak, serta memperoleh pekerjaan. Pada tahun 2014 IPM Indonesia berada di peringkat ke-110 dari 188 negara di dunia. Akan tetapi pada tahun 2017 peringkat Indonesia turun menjadi peringkat ke-113. Pada tahun 2017 tersebut, terjadi peningkatan IPM di Indonesia sebesar 0.004 menjadi 0.698. Akan tetapi peningkatan tersebut tidak memeberikan pengaruh signifikan karena nilai IPM Indonesia kurang dari IPM rata-rata dunia. Setelah ditelisik lebih lanjut, Peningkatan IPM dari tahun ke tahun di Indonesia tidak diiringi dengan peningkatan dibidang kesehatan. Hal ini menunjukkan bahwa perlu adanya peningkatan kualitas sumber daya manusia dibidang kesehatan.

Masalah kesehatan di Indonesia sangat bervariasi yang dipengaruhi oleh berbagai faktor yang sangat kompleks dan saling terkait seperti keadaan sosial ekonomi, budaya, dan kependudukan. Untuk mengatasi masalah kesehatan di Indonesia, pemerintah menggalakkan program asuransi. Dalam pasal 246 KUHD disebutkan bahwa asuransi adalah suatu perjanjian dimana seorang penanggung mengaitkan diri kepada seorang tertanggung dengan menerima suatu premi, untuk memberikan penggantian kepadanya karena suatu kerugian, kerusakan atau kehilangan keuntungan

yang mungkin akan dideritanya. Tidak hanya asuransi kesehatan, asuransi lain yang berkembang di Indonesia yaitu asuransi jiwa, asuransi pendidikan, asuransi perjalanan serta asuransi kerugian. Pada intinya asuransi memberikan banyak manfaat dan keuntungan yaitu melindungi risiko investasi, dapat mengurangi kekhawatiran, mendorong usaha pencegahan kerusakan, membantu pemeliharaan kesehatan, dan lain-lain (Darmawi, 2004). Saat ini asuransi berkembang pesat di Indonesia. Tidak hanya asuransi milik pemerintah atau biasa dikenal dengan asuransi Badan Usaha Milik Negara (BUMN), asuransi milik swasta juga berkembang cukup pesat. Sampai saat ini, terdapat 15 asuransi milik pemerintah atau BUMN meliputi PT Asuransi Kesehatan Indonesia, PT Asuransi Jasa Raharja, PT Asuransi Jiwasraya, PT Taspen, PT Asuransi Jasa Indonesia (JASINDO) dan lain-lain. Untuk asuransi swasta yang berdiri di Indonesia meliputi AIA Finansial, Allianz, Avirst AXA Mandiri, CIGNA, Prudential dan Asuransi Sinar Mas. Berkembang pesatnya asuransi di Indonesia tidak didukung dengan pemahaman masyarakat akan pentingnya asuransi. Hal itu tampak pada masih minimnya jumlah nasabah yang terdaftar sebagai peserta asuransi. Pada 2017, Otoritas Jasa Keuangan (OJK) memberikan informasi bahwa warga Indonesia yang menjadi peserta asuransi tidak lebih dari 10%. Tidak seperti negara-negara maju di barat dimana asuransi menjadi kewajiban bagi seluruh warganya.

Indonesia pernah melakukan survei mengenai status kepemilikan asuransi pada *Indonesia Family Life Survey* (IFLS). IFLS biasa dikenal dengan Survei Aspek Kehidupan Rumah Tangga (SAKERTI). Survei ini diadakan atas kerjasama antara organisasi penelitian Amerika Serikat (RAND Corporation), Lembaga Demografi Universitas Indonesia, Pusat Studi Kependudukan dan Kebijakan Universitas Gajah Mada serta lembaga penelitian SurveyMETER. IFLS terdiri dari dua survei yaitu survei rumah tangga (*household survey*) serta survei komunitas dan fasilitas (*Community Facility Survey*). Survei ini tidak hanya mengenai asuransi tetapi survei diberbagai bidang

meliputi sosial-demografi, kesehatan, pendidikan, ekonomi dan lain-lain. IFLS telah dilakukan selama 5 gelombang yaitu pada tahun 1993, 1997, 2000, 2007 dan 2014 di 13 provinsi di Indonesia. Jumlah responden pada IFLS-1 sebanyak 7224 rumah tangga dengan jumlah individu sebanyak 22000 orang. IFLS selanjutnya dilakukan dengan menggunakan responden yang sama dengan persentase *re-contact* IFLS 2, 3, 4 dan 5 berturut-turut sebesar 94.4%, 95.3%, 93.6% dan 92%. IFLS-5 dilakuakan dengan sistem yang berbeda dari sebelumnya dimana survei kali ini menggunakan sistem *Computer-Assisted Personal Interview* (CAPI) dan tidak lagi menggunakan kuesioner kertas. Program CAPI telah dipersiapkan dan diuji coba selama kurang lebih 18 bulan. Selain itu, pengambilan data pada IFLS-5 juga telah menggunakan alat perekam suara sehingga kualitas data dapat terkontrol dengan baik (Strauss, Witoelar, & Sikoki, 2016). Data survei IFLS ini telah digunakan dalam beberapa penelitian. Pada tahun 2011, Ewijk melakukan analisis efek jangka panjang puasa ramadhan terhadap ibu hamil. Vaezghasemi et al. (2016) juga melakukan analisis data IFLS mengenai variasi *Body Mass Index* (BMI) di rumah tangga dan kabupaten selama dua dekade. Selanjutnya pada tahun 2018, Peltzer & Pengpid melakukan pemodelan terhadap faktor penyebab hipertensi dengan studi kasus pada IFLS-5. Pada tahun yang sama, Isaura et al. menggunakan data IFLS-4 untuk studi longitudinal terhadap skor konsumsi makanan, indeks massa tubuh dan hipertensi pada orang dewasa. Terbukanya data IFLS untuk publik membuka kesempatan peneliti untuk berkarya diberbagai bidang.

Penelitian sebelumnya mengenai asuransi pada data *Indonesia Family Life Survey* (IFLS) dilakukan oleh Hidayat et. al (2004). Hasil penelitian menunjukkan bahwa asuransi wajib untuk pegawai negeri sipil (ASKES) memiliki dampak positif terhadap akses ke fasilitas rawat jalan umum, sementara asuransi wajib untuk pegawai swasta (Jamsostek) memiliki dampak positif terhadap akses ke fasilitas rawat jalan umum dan swasta. Zerria (2016) melakukan penelitian mengenai faktor permintaan asuransi

di Mena dan menunjukkan bahwa capaian pendidikan serta pendapatan berpengaruh signifikan terhadap permintaan asuransi. Selanjutnya tahun 2017, Bernal juga melakukan penelitian efek dari kepemilikan asuransi kesehatan dengan studi kasus masyarakat Peru, Amerika Selatan. Dari hasil penelitian tersebut, menunjukkan bahwa peningkatan jumlah kepemilikan asuransi membawa dampak positif terhadap peningkatan kesehatan, dimana asuransi tidak hanya digunakan untuk masyarakat golongan menengah keatas. Pada tahun 2019, Liu et al. melakukan penelitian hubungan faktor kesehatan dengan asuransi. Dan pada tahun yang sama, Erlangga et al. meneliti dampak asuransi Jaminan Kesehatan Nasional (JKN) terhadap permintaan layanan kesehatan. Dari penelitian-penelitian sebelumnya, dapat diambil kesimpulan bahwa kepemilikan asuransi tidak hanya disebabkan oleh faktor kesehatan, namun ada faktor lain seperti pendidikan, pendapatan dan lain-lain.

Berdasarkan uraian tersebut, perlu dilakukan penelitian lebih lanjut mengenai asuransi. Hal ini dilakukan untuk mengetahui faktor-faktor apa saja yang berpengaruh terhadap kepemilikan asuransi di Indonesia. Penelitian ini tidak hanya dibatasi dengan responden yang memiliki asuransi, tetapi juga dilakukan terhadap responden yang tidak memiliki asuransi. Asuransi ini meliputi asuransi milik pemerintah atau BUMN dan asuransi milik swasta. Variabel respon pada penelitian ini merupakan variabel kategorik yang bersifat *polichotomus*, sehingga metode regresi logistik multinomial merupakan metode yang tepat. Mudhu (2013) melakukan analisis regresi logistik multinomial pada kasus kanker payudara di Karnataka Selatan. Dari penelitian tersebut, menunjukkan bahwa model logistik multinomial tepat digunakan dan memberikan nilai residual yang kecil. Selanjutnya pada tahun 2017, Asampana juga menggunakan regresi logistik multinomial mengenai faktor penentu kinerja akademik mahasiswa pada pelajaran matematika. Hasil penelitian memberikan informasi bahwa regresi logistik multinomial dapat

memprediksi faktor penentu kinerja akademik siswa dengan mendapatkan 9 variabel prediktor yang berpengaruh signifikan.

Kasus kepemilikan asuransi di Indonesia merupakan *imbalanced data,* dimana kondisi data tidak berimbang antara kelas data satu dengan kelas data yang lain. Kelas data yang banyak merupakan kelas data mayoritas, sedangkan kelas data sedikit merupakan kelas data minoritas. Kondisi *imbalanced* data menjadi masalah karena mesin *classifier learning* akan condong memprediksi ke kelas data yang banyak (mayoritas) dibandingkan dengan kelas minoritas (Japkowicz & Stephen, 2002). *Sampling-based approaches* merupakan pendekatan sampling yang meodifikasi distribusi data training sehingga kedua kelas (mayoritas dan minoritas) dipersentasikan dengan baik. Pendekatan *sampling* dibedakan menjadi dua yaitu *undersampling* dan *oversampling*. Metode *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan kelas minor. Masalah yang muncul dari metode *oversampling* adalah masalah *overfitting* yang menyebabkan aturan klasifikasi menjadi semakin spesifik meskipun akurasi untuk data training semakin baik. Sedangkan metode *undersampling* dilakukan dengan mengurangi jumlah data kelas mayor agar data menjadi seimbang. Kekurangan metode *undersampling* adalah semakin berkurangnya informasi dari data karena banyak data yang dihilangkan, yang banyak informasinya sehingga efektivitas klasifikasi menurun, sedangkan penghapusan data yang tidak relevan, berlebihan ataupun *noise* mengakibatkan efektivitas klasifikasi meningkat (Chawla, et al., 2002).

Salah satu metode *oversampling* adalah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) yang diperkenalkan oleh Chawla, et al. (2002). SMOTE digunakan untuk menambah jumlah data kelas minoritas dengan cara mereplikasi data secara acak agar seimbang dengan jumlah kelas mayoritas. Sedangkan, salah satu metode *undersampling* adalah

random under sampling. Metode ini bekerja dengan menghitung selisih antara kelas mayoritas dan minoritas kemudian dilakukan perulangan selisih hasil perhitungan, selama perulangan data kelas mayoritas dihapus secara acak, sehingga jumlah kelas mayoritas sama dengan minoritas. Gaudio *et al.* (2013) memperoleh kesimpulan bahwa metode gabungan *combine sampling* efektif untuk menangani *higher imbalanced* data. Sehingga pada penelitian ini akan dilakukan dengan metode regresi logistik multinomial pada *imbalanced data,* dan *balanced data* dengan *prepocessing* SMOTE, *random undersampling* serta *combine sampling.* Hasil penelitian ini diharapkan dapat dijadikan referensi untuk mengetahui hal-hal apa saja yang mempengaruhi produktivitas industri asuransi serta keberhasilannya dalam meningkatkan mutu layanan kesehatan di Indonesia.

1.2 Rumusan Masalah

Masalah kesehatan di Indonesia sangat bervariasi dan untuk mengatasinya pemerintah menggalakkan program asuransi. Saat ini asuransi berkembang pesat di Indonesia akan tetapi penggunaannya tidak meratanya, dimana asuransi pemerintah lebih mendominasi dibandingkan dengan asuransi swasta dan penduduk yang tidak memiliki asuransi. Mengingat pentingnya asuransi, maka pada penelitian ini dilakukan analisis untuk mengetahui karakteristik status kepemilikan asuransi di Indonesia dengan metode regresi logistik multinomial. Untuk mengatasi kasus imbalanced data dilakukan dengan *prepocessing* SMOTE, *random undersampling* dan *combine sampling*.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dibahas sebelumnya, tujuan yang ingin dicapai dalam penelitian ini sebagai berikut.

1. Memperoleh karakteristik data berdasarkan status kepemilikan asuransi pada *Indonesia Family Life Survey* (IFLS) serta faktor-faktor yang mempengaruhinya.
2. Mengetahui ketepatan klasifikasi dan faktor signifikan data *imbalanced* data *balanced* dengan *prepocessing* SMOTE, *undersampling* dan *combine sampling* pada status kepemilikan asuransi dengan metode *multinomial logistic regression.*
3. Mendapatkan perbandingan performansi dari *imbalanced data* dan *balanced data* dengan *prepocessing* SMOTE, *undersampling* dan *combine sampling* pada status kepemilikan asuransi.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak sebagai berikut.

1. Memberikan wawasan keilmuan dalam penerapan metode analisis data *imbalanced multiclass* dengan metode *multinomial logistic regression.*
2. Bagi pemerintah, hasil penelitian ini diharapkan dapat memberikan informasi yang berguna sebagai upaya untuk meningkatkan kualitas sumber daya manusia dibidang kesehatan.
3. Dapat dijadikan sebagai referensi untuk mengetahui hal-hal apa saja yang mempengaruhi produktivitas industri asuransi serta keberhasilannya dalam meningkatkan mutu layanan kesehatan di Indonesia.

1.5 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini adalah data pada *Indonesia Family Life Survey* (IFLS) gelombang ke 5, dimana survei tersebut dilakukan pada september 2014 hingga maret 2015. Topik asuransi yang digunakan yaitu asuransi kesehatan dengan unit penelitian individu usia kerja. Menurut undang-undang no 13 tahun 2003 menyebutkan bahwa penduduk tergolong tenaga kerja jika memasuki usia kerja, dimana usia kerja

yang berlaku di Indonesia adalah umur 15 tahun keatas. Metode yang digunakan untuk mengatasi *imbalanced multiclass* dalam penelitian ini adalah SMOTE, *Random Undersampling* dan *Combine Sampling.* Sedangkan metode analisisnya yaitu *Multinomial Logistic Regression* dengan menggunakan full set data untuk *training* dan *testing.*

BAB II
TINJAUAN PUSTAKA

Bab ini membahas mengenai landasan teori yang digunakan dalam penelitian meliputi statistika deskriptif, uji independensi, asumsi multikolinieritas, *imbalanced data* (SMOTE dan *random under sampling*) dan regresi logistik multinomial (estimasi parameter, signifikansi parameter, uji kesesuaian model, *odds ratio* serta evaluasi ketepatan klasifikasi).

2.1 Statistika Deskriptif

Statistika Deskriptif merupakan suatu metode statistika yang menyajikan data dalam bentuk tabel, diagram, grafik, dan ukuran penyimpangan, tetapi tidak menghasilkan penarikan kesimpulan yang berlaku secara generalisasi. Maka statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian data sehingga memberikan informasi yang berguna (Walpole, 2012). Statistika deskriptif dilakukan dengan memeriksa ukuran pemusatan data dan ukuran penyebaran data. Terdapat berbagai macam cara penyajian dalam statistika deskriptif melalui diagram seperti *pie chart* dan *bar chart*. *Pie chart* yaitu diagram yang digunakan untuk menyajikan data dengan skala pengukuran nominal atau ordinal. Sedangkan *bar chart* menunjukkan keterangan dengan batang tegak atau mendatar dan sama lebar dengan batang-batang terpisah (Aczel & Sounderpandian, 2008). Selain menggunakan diagram, cara peyajian statistika deskriptif juga dapat dilakukan dengan menggunakan tabel, grafik, serta *cross tabulation*.

Cross tabulation merupakan salah satu metode statistika yang menggambarkan dua atau lebih variabel secara bersama-sama dan hasilnya berupa tabel kontingensi. Tabel kontingensi dapat menunjukkan hubungan antar variabel kategorikal. Sebuah tabel dibuat dengan *I* baris untuk kategori *X* dan *J* kolom untuk kategori *Y*, maka sel dari tabel tersebut akan menunjukkan *IJ* hasil yang

mungkin (Agresti, 2013). Tabel 2.1 menunjukkan *cross tabulation* berukuran *I* x *J*.

Tabel 2.1 *Ilustrasi Cross Tabulation I x J*

Variabel X	**Variabel Y**				**Total**
	1	2	…	J	
1	n_{11}	n_{12}	…	n_{1J}	n_{1}
2	n_{21}	n_{22}	…	n_{2J}	n_{2}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
I	n_{I1}	n_{I2}	…	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	…	$n_{.J}$	$n_{..}$

2.2 Asumsi Multikolinieritas

Multikolinieritas merupakan adanya hubungan linier antara beberapa atau semua variabel prediktor pada analisis regresi (Gujarati & Porter, 2009). Asumsi ini merupakan satu-satunya asumsi yang harus dipenuhi dalam menggunakan metode regresi logistik. Pengecekan asumsi multikolinieritas dapat dilakukan dengan melihat nilai *Variance Inflation Factor* (VIF) dari setiap variabel prediktor. Jika nilai VIF lebih besar dari 10, mengindikasikan bahwa terjadi adanya kasus multikolinieritas.

Nilai VIF diperoleh dengan melakukan regresi masing-masing variabel prediktor dengan variabel prediktor lainnya dan melihat nilai R^2. Rumus untuk mendapatkan nilai VIF dari variabel prediktor ke-j dapat dijelaskan dalam persamaan (2.1) sebagai berikut.

$$VIF_j = \frac{1}{1-R_j^2}, \quad j = 1, 2, ..., p \tag{2.1}$$

R_j^2 merupakan koefisien determinasi antara X_j dengan variabel prediktor lainnya. R^2 dapat dinyatakan dengan persamaan (2.2) sebagai berikut.

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.2}$$

dimana :

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{2.3}$$

$$SST = \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 \tag{2.4}$$

Keterangan :

SSE = *Sum Square Error*

SST = *Sum Square Total*

y_i = Nilai variabel respon ke-i

$\hat{y}_i$ = Nilai dugaan variabel respon ke-i

$\bar{y}_i$ = Rata-rata variabel respon

Selain dari nilai VIF, indikasi adanya multikolinieritas dapat dilihat dengan korelasi pearson (r_{xy}) dengan rumus (2.5) sebegai berikut.

$$r_{xy} = \frac{\sum_{j=1}^{n} \left(x_j - \bar{x} \right) \left(y_j - \bar{y} \right)}{\sqrt{\sum_{j=1}^{n} \left(x_j - \bar{x} \right)^2} \sqrt{\sum_{j=1}^{n} \left(y_j - \bar{y} \right)^2}} \tag{2.5}$$

Jika nilai $r_{xy} > 0.95$ maka menunjukkan adanya hubungan multikolinierias antar variabel.

2.3 Uji Independensi

Pengujian independensi bertujuan untuk memeriksa ada tidaknya hubungan antara dua variabel yang sedang diamati (Agresti, 2013). Adapun hipotesis yang digunakan adalah sebagai berikut.

H_0 : $\boldsymbol{\rho = I}$ (tidak terdapat hubungan antar kedua variabel)

H_1 : $\boldsymbol{\rho \neq I}$ (terdapat hubungan antar kedua variabel)

Statistik uji yang digunakan yaitu *pearson Chi-Squared test* dengan rumus pada persamaan (2.6).

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}\right] \tag{2.6}$$

dimana

$$e_{ij} = \frac{n_{i.}\ x\ n_{.j}}{n..} \tag{2.7}$$

keterangan :

n_{ij} = Total frekuensi observasi untuk baris ke-i dan kolom ke-j dengan i=1, 2, …, I dan j=1, 2, …, J

$n_{i.}$ = Total frekuensi observasi untuk baris ke-i dengan i=1,2,…, I

$n_{.j}$ = Total frekuensi observasi untuk baris ke-j dengan j=1,2,…, J

e_{ij} = Nilai harapan pada baris ke-i dan kolom ke-j dengan i=1,2, …,I dan j=1,2,…,J

Daerah penolakan yaitu tolak H_0 jika nilai $\chi^2 > \chi^2_{\alpha;(I-1)(J-1)}$ yang menunjukkan bahwa terdapat hubungan antar kedua variabel.

2.4 Data *Imbalanced*

Data *imbalanced* merupakan kondisi data yang tidak berimbang dengan jumlah data suatu kelas melebih jumlah data kelas yang lain, kelas data yang banyak merupakan kelas mayoritas atau kelas negatif sedangkan kelas data yang sedikit merupakan kelas minoritas atau kelas positif. Pendekatan pada level data untuk menangani masalah *imbalanced data* adalah dengan menggunakan *Sampling-based approaches*. Dengan adanya penerapan *sampling* pada data yang *imbalanced*, tingkat *imbalanced data* semakin kecil dan klasifikasi dapat dilakukan dengan tepat (Solberg & Solberg, 1996). *Sampling-based approaches* yaitu memodifikasi distribusidari data *training* sehingga kedua kelas data (negatif maupun positif) dipresentasikan dengan baik di dalam data *training*. *Sampling* sendiri dibedakan menjadi dua yaitu *undersampling* dan *oversampling*. Metode *oversampling* dilakukan

untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan jumlah data kelas minor dan metode *undersampling* dilakukan dengan cara mengurangi jumlah data kelas mayor. Pada penelitian ini menggunakan *Combine Undersampling* yaitu mengkombinasikan metode *Synthetic Minority Oversampling Technique* (SMOTE) dan *Random Undersampling*, sebagai berikut.

2.4.1 *Synthetic Minority Oversampling Technique* (SMOTE)

Untuk mengatasi masalah *imbalanced data* dapat menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE) yang merupakan salah satu metode *oversampling* yang ditemukan oleh Chawla (2013) berupa teknik penambahan jumlah sampel pada kelas negatif dengan melakukan replikasi data pada kelas negatif secara acak sehingga menghasilkan jumlah data yang sama dengan data padaa kelas positif atau data mayor. Data yang direplikasi merupakan data yang berasal dari kelas minor. Metode yang digunakan pada algoritma SMOTE adalah *k-nearest neighbors* (ketetanggaan data) yang termasuk dalam kelompok metode statistik nonparametrik. Metode ini bekerja dengan cara mengelompkkan data terdekat yang dipilih berdasarkan jarak *euclidean* antara kedua data. Penentuan jumlah replikasi yang dilakukan disesuaikan dengan jumlah anggota pada kelas mayor. Jumlah replikasi harus sesuai dengan jumlah *k* pada *nearest neighbor*, jika jumlah replikasi sebanyak *n* maka jumlah *k* sebanyak *n-1*.

Misalkan terdapat dua struktur data dengan p dimensi yaitu $\mathbf{x}^T = \left[x_1, x_2, ..., x_p \right]$ dan $\mathbf{y}^T = \left[y_1, y_2, ..., y_p \right]$ maka jarak *euclidean* $d(\mathbf{x}, \mathbf{y})$ antara kedua vector data adalah sebagai berikut.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \qquad (2.8)$$

Synthetic atau replikasi data dapat dilakukan dengan menggunakan persamaan sebagai berikut.

$$\mathbf{x}_{syn} = \mathbf{x}_i + \left(\mathbf{x}_{knn} - \mathbf{x}_i\right)\tau \tag{2.9}$$

dengan,

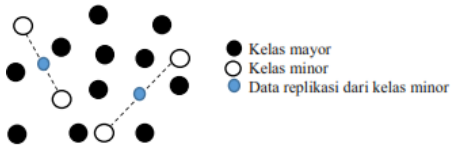
$\mathbf{x}_{syn}$: Data hasil replikasi

$\mathbf{x}_i$: Data yang akan direplikasi

$\mathbf{x}_{knn}$: Data memiliki jarak terdekat dari data yg akan direplikasi

τ : Bilangan random antara 0 sampai 1

Ilustrasi mengenai *Synthetic Minority Oversampling Technique* (SMOTE) digambarkan pada Gambar 2.1 sebagai berikut.



Gambar 2.1 Ilustrasi SMOTE

Sebagai contoh akan dijelaskan ilustrasi mengena metode SMOTE dalam menangani *imbalanced data*. Berikut ini akan digunakan 16 data dengan jumlah variabel prediktor sebanyak 2 yaitu X_1 dan X_2 dengan 3 kelas pada variabel respon (Y) yakni kelas 1 sebanyak 9 data, kelas 2 sebanyak 4 data, dan kelas 3 sebanyak 3 data. Sehingga kelas mayor dengan jumlah data terbanyak dimiliki oleh variabel respon dengan kelas 1. Sedangkan variabel respon kelas 2 dan kelas 3 merupakan kelas minor. Data simulasi SMOTE akan ditampilkan pada Tabel 2.2 sebagai berikut.

Tabel 2.2 Data Simulasi untuk SMOTE

No	X_1	X_2	Y	No	X_1	X_2	Y
1	2	3	1	9	2	4	1
2	1	2	1	10	4	6	2
3	2	4	1	11	5	1	2
4	2	6	1	12	1	3	2

Tabel 2.2 Data Simulasi SMOTE (Lanjutan)

No	X_1	X_2	Y	No	X_1	X_2	Y
5	3	4	1	13	2	5	2
6	4	3	1	14	3	6	3
7	5	5	1	15	4	1	3
8	1	2	1	16	4	7	3

Berdasarkan tabel 2.2 maka persentase banyaknya data masing-masing kelas pada variabel respon tidak seimbang dibuktikan dari nilai *persentase* yang dihasilkan, sehingga kelas mayor adalah data yang terdapat pada kelas 1 dibuktikan dari jumlah *persentase* terbesar yakni sebesar 53, sedangkan *persentase* jumlah anggota untuk kelas 2 sebesar 25% dan kelas 3 sebesar 19%.

Sehingga diperlukan penanganan untuk kasus *imbalanced data* menggunakan metode SMOTE dengan langkah-langkah sebagai berikut.

1. Setiap data pada kelas minor akan direplikasikan dengan mencari tetangga terdekat (x_{knn}) dengan menggunakan jarak *euclidean* pada persamaan 2.8.
2. Menghitung *synthetic* data pada kelas minor tersebut dengan menggunakan persamaan 2.9.

Distribusi data yang dihasilkan pada kelas mayor dan minor dengan menggunakan SMOTE adalah sebagai berikut.

Tabel 2.3 Distribusi Data SMOTE

Sebelum Replikasi		Jumlah Replikasi	Setelah Replikasi	
Mayor	**Minor**		**Mayor**	**Minor**
9 (56%)	4 (25%)	1 kali	9 (34%)	8 (32%)
	3 (19%)	2 kali		9 (34%)

Berdasarkan Tabel 2.2, dapat diketahui jumlah data setelah menggunakan SMOTE jumlahnya bertambah yang semula berjumlah 16 data menjadi 26 data. Sedangkan jika dilihat dari

persentase pada masing-masing kelas besarnya seimbang. Jumlah anggota pada kelas mayor yakni kelas 1 yaitu sebanyak 9 data, sedangkan jumlah anggota pada kelas 2 sebanyak 4 data sehingga perlu dilakukan replikasi sebanyak 1 kali sehingga jumlah tetangga terdekatnya sama dengan nol. Sedangkan jumlah data pada kelas 3 sebanyak 3 data sehingga perlu dilakukan replikasi sebanyak 2 kali dan jumlah tetangga terdekatnya sama dengan satu untuk masing-masing data. Sehingga data baru setelah dilakukan replikasi menggunakan SMOTE ditunjukkan pada Tabel 2.4 sebagai berikut.

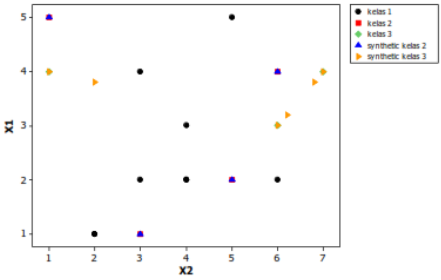
Tabel 2.4 Data Simulasi setelah Menggunakan SMOTE

No	X_1	X_2	Y	No	X_1	X_2	Y
1	2	3	1	14	3	6	3
2	1	2	1	15	4	1	3
3	2	4	1	16	4	7	3
4	2	6	1	17*	4	6	2
5	3	4	1	18*	5	1	2
6	4	3	1	19*	1	3	2
7	5	5	1	20*	2	5	2
8	1	2	1	21*	3	6	3
9	2	4	1	22*	3	6	3
10	4	6	2	23*	4	1	3
11	5	1	2	24*	3	2	3
12	1	3	2	25*	4	7	3
13	2	5	2	26*	3	6	3

* data hasil replikasi SMOTE

Setelah dilakukan replikasi terlihat jumlah anggota kelas minor telah seimbang dengan jumlah anggota pada kelas mayor. Berdasarkan Gambar 2.3 dapat diketahui, sebaran data hasil replikasi kelas 2 yang ditunjukkan dengan pola yang berwarna biru, sedangkan data replikasi kelas 3 ditunjukkan dengan pola yang berwarna kuning. Titik koordinat anggota masing-masing kelas

setelah dilakukan replikasi akan ditunjukkan pada Gambar 2.2 sebagai berikut.



Gambar 2.2 Plot Data setelah Replikasi dengan SMOTE

2.4.2 Random Undersampling

Random Undersampling (RUS) menghitung selisih antara kelas mayoritas dan minoritas kemudian dilakukan perulangan selisih hasil perhitungan, selama perulangan data kelas mayoritas dihapus secara acak, sehingga jumlah kelas mayoritas sama dengan minoritas. Langkah pertama pada *Random Undersampling* adalah pemilihan dataset kemudian dihitung selisih antara kelas mayoritas dan minoritas, jika masih terdapat selisih antara jumlah kelas maka dataset kelas mayoritas akan dihapus secara acak sampai jumlah kelas mayoritas sama dengan kelas minoritas.

2.5 Regresi Logistik Multinomial

Regresi logistik multinomial merupakan regresi logistik yang digunakan saat variabel dependen mempunyai skala yang bersifat *polichotomus* atau multinomial. Skala multinomial adalah suatu pengukuran yang dikategorikan menjadi lebih dari dua kategori (Agresti, 2013). Metode yang digunakan dalam penelitian ini adalah regresi logistik dengan skala nominal dengan tiga kategori.

Mengacu pada regresi logistik *trichotomus* (Hosmer & Lemeshow, 2000) untuk model regresi dengan variabel dependen berskala nominal tiga kategori digunakan kategori varibel hasil Y dengan koding 0, 1, dan 2. Variabel Y terparameterisasi menjadi

dua fungsi logit. Sebelumnya perlu ditentukan kategori hasil mana yang digunakan untuk membandingkan. Pada umumnya digunakan Y=0 untuk pembanding. Untuk membentuk fungsi logit, akan dibandingkan Y=1 dan Y=2 terhadap Y=0. Bentuk model regresi logistik dengan p variabel prediktor dengan menggunakan transformasi logit akan didapatkan dua fungsi logit pada persamaan (2.10) dan persamaan (2.11).

$$g_1(\mathbf{x}) = \ln\left[\frac{P(Y=1\mid\mathbf{x})}{P(Y=0\mid\mathbf{x})}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1k}x_k = \mathbf{x}^T\boldsymbol{\beta}_1 \tag{2.10}$$

dan

$$g_2(\mathbf{x}) = \ln\left[\frac{P(Y=2\mid\mathbf{x})}{P(Y=0\mid\mathbf{x})}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2k}x_k = \mathbf{x}^T\boldsymbol{\beta}_2 \tag{2.11}$$

dengan

$\mathbf{x} = \begin{bmatrix}1 & x_1 & x_2 & \ldots & x_k\end{bmatrix}^T$

$\boldsymbol{\beta}_1 = \begin{bmatrix}1 & \beta_{11} & \beta_{12} & \ldots & \beta_{1k}\end{bmatrix}^T$

$\boldsymbol{\beta}_2 = \begin{bmatrix}1 & \beta_{21} & \beta_{22} & \ldots & \beta_{2k}\end{bmatrix}^T$

Berdasarkan kedua fungsi logit pada persamaan (2.10) dan (2.11) didapatkan fungsi peluang regresi logistik multinomial untuk setiap kategori variabel respon pada persamaan (2.12), persamaan (2.13) dan persamaan (2.14).

$$\pi_0(\mathbf{x}) = P(Y=0\mid\mathbf{x}) = \frac{1}{1+\exp g_1(\mathbf{x}) + \exp g_2(\mathbf{x})} \tag{2.12}$$

$$\pi_1(\mathbf{x}) = P(Y=1\mid\mathbf{x}) = \frac{\exp g_1(\mathbf{x})}{1+\exp g_1(\mathbf{x}) + \exp g_2(\mathbf{x})} \tag{2.13}$$

$$\pi_2(\mathbf{x}) = P(Y=2\mid\mathbf{x}) = \frac{\exp g_2(\mathbf{x})}{1+\exp g_1(\mathbf{x}) + \exp g_2(\mathbf{x})} \tag{2.14}$$

Secara umum *conditional probability* dari ketiga kategori variabel dependen dapat dituliskan pada persamaan (2.15).

$$\pi_j(\mathbf{x}) = P(Y = j \mid \mathbf{x}) = \frac{\exp g_j(\mathbf{x}_i)}{\sum_{j=0}^{2} \exp g_j(\mathbf{x}_i)} \tag{2.15}$$

dimana

$$\exp g_0(\mathbf{x}_i) = 1 \tag{2.16}$$

dengan j kategori dari variabel dependen atau respon (Hosmer & Lemeshow, 2000).

2.5.1 Estimasi Parameter Model Regresi Logistik Multinomial

Estimasi parameter dalam model regresi logistik dilakukan dengan metode *Maximum Likelihood Estimation (MLE).* Metode *Maximum Likelihood* mengestimasi parameter **β** dengan cara memaksimalkan fungsi *likelihood* (Agresti, 2013). Penduga parameter maksimum merupakan penduga yang konsisten dan efisien untuk ukuran sampel yang besar. Hosmer & Lemeshow (2000) menyatakan apabila suatu variabel dependen atau respon mempunyai tiga kategori, maka akan terbentuk kemungkinan tiga *outcome* sehingga didapatkan fungsi *likelihood* seperti pada persamaan (2.17).

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_0(\mathbf{x}_i)^{y_{0i}} \, \pi_1(\mathbf{x}_i)^{y_{1i}} \, \pi_2(\mathbf{x}_i)^{y_{2i}} \right] \tag{2.17}$$

dengan i = 1, 2, …, n. Persamaan (2.17) digunakan untuk menyusun *ln likelihood* seperti pada persamaan (2.18).

$$\begin{aligned} L(\boldsymbol{\beta}) &= \ln\left(l(\boldsymbol{\beta})\right) \\ &= \sum_{i=1}^{n} y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - \ln\left(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}\right) \end{aligned} \tag{2.18}$$

dengan melakukan differensial terhadap persamaan (2.18), maka dapat dihitung parameter-parameter regresi logistik multinomial dengan persamaan (2.19) dan persamaan (2.20).

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} = \sum_{i=1}^{n} x_i^T (y_{1i} - \pi_{1i}) \tag{2.19}$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} = \sum_{i=1}^{n} x_i^T (y_{2i} - \pi_{2i}) \tag{2.20}$$

dengan $\boldsymbol{\pi}_i$ adalah penyederhanaan dari $\boldsymbol{\pi}_1\left(\mathbf{x}_i\right)$ dan $\boldsymbol{\pi}_{2i}$ adalah penyederhanaan dari $\boldsymbol{\pi}_2\left(\mathbf{x}_i\right)$. Persamaan (2.19) dan persamaan (2.20) merupakan persamaan non-linear sehingga diperlukan metode iterasi numerik untuk memperoleh $\boldsymbol{\beta}$ yang konvergen. Metode iterasi *Newton Raphson* digunakan untuk menyelesaikan persamaan non-linear, dengan rumus iterasi seperti pada persamaan (2.21).

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}\left(\boldsymbol{\beta}^{(t)}\right) g\left(\boldsymbol{\beta}^{(t)}\right) \tag{2.21}$$

dimana

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^T & \beta_2^T \end{bmatrix}^T$$

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 1 & \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \end{bmatrix}^T$$

$$\boldsymbol{\beta}_2 = \begin{bmatrix} 1 & \beta_{21} & \beta_{22} & \cdots & \beta_{2k} \end{bmatrix}^T$$

$$\mathbf{g}\left(\boldsymbol{\beta}\right) = \begin{bmatrix} \dfrac{\partial \ln L(\cdot)}{\partial \boldsymbol{\beta}_1^T} \\ \dfrac{\partial \ln L(\cdot)}{\partial \boldsymbol{\beta}_2^T} \end{bmatrix}$$

$$\mathbf{H}\left(\boldsymbol{\beta}\right) = \begin{bmatrix} \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} & \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^T} \\ \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1^T} & \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^T} \end{bmatrix}$$

Langkah-langkah iterasi *Newton Raphson* adalah sebagai berikut.

1. Menentukan nilai awal estimasi parameter $\hat{\boldsymbol{\beta}}^{(0)}$

   dimana $\boldsymbol{\beta}_{(0)} = \begin{pmatrix} 0 & 0 & \cdots & 0 \end{pmatrix}^T$ atau $\boldsymbol{\beta}_{(0)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$
2. Membentuk vektor gradient $\mathbf{g}$ dan matriks Hessian $\mathbf{H}$

3. Memasukkan nilai $\hat{\boldsymbol{\beta}}^{(0)}$ pada elemen $\mathbf{g}$ dan $\mathbf{H}$ sehingga diperoleh $\mathbf{g}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$ dan $\mathbf{H}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$
4. Iterasi dimulai t=0 menggunakan persamaan (2.21). Nilai $\hat{\boldsymbol{\beta}}^{(t)}$ adalah sekumpulan penaksir parameter yang konvergen pada iterasi ke-t.
5. Apabila belum diperoleh estimasi parameter yang konvergen, maka langkah (3) diulang kembali hingga nilai $\left\|\hat{\boldsymbol{\beta}}^{(t+1)}-\hat{\boldsymbol{\beta}}^{(t)}\right\| \leq \varepsilon$, dengan ε merupakan bilangan yang sangat kecil. Hasil estimasi yang diperoleh adalah $\hat{\boldsymbol{\beta}}^{(t+1)}$ pada iterasi terakhir.

2.5.2 Pengujian Signifikansi Parameter Model Regresi Logistik Multinomial

Pengujian signifikansi parameter dilakukan untuk mengetahui apakah taksiran parameter berpengaruh signifikan terhadap model atau tidak, selain itu juga untuk mengetahui seberapa besar pengaruh dari masing-masing parameter tersebut. Pengujian signifikansi parameter secara serentak pada model Regresi Logistik Multinomial dengan 3 kategori respon dilakukan menggunakan *Likelihood Ratio Test,* dengan hipotesis sebagai berikut.

H_0 : $\beta_{11} = \beta_{21} = ... = \beta_{2k} = 0$

H_1 : Minimal terdapat $\beta_{ij} \neq 0$ dengan i = 1, 2 dan j = 1,2, … , k

Statistik uji *Likelihood Ratio Test* dijelaskan pada persamaan (2.22) berikut.

$$G^2 = 2\left[\ln l\left(\hat{\Omega}\right) - \ln l\left(\hat{\omega}\right)\right] \tag{2.22}$$

dengan penjelasan pada persamaan (2.23) dan (2.24)

$$\ln l\left(\hat{\Omega}\right) = \sum_{j=1}^{k}\left[y_{1j}x_j^T\hat{\boldsymbol{\beta}}_1 + y_{2j}x_j^T\hat{\boldsymbol{\beta}}_2 - \ln\left(1+e^{x_j^T\hat{\boldsymbol{\beta}}_1}+e^{x_j^T\hat{\boldsymbol{\beta}}_2}\right)\right] \tag{2.23}$$

$$\ln l(\hat{\omega}) = \sum_{j=1}^{k}\left[y_j \hat{\beta}_{01} + y_j \hat{\beta}_{02} - \ln\left(1 + e^{\hat{\beta}_{01}} + e^{\hat{\beta}_{02}}\right)\right] \tag{2.24}$$

cara mendapatkan parameter β tanpa melibatkan variabel bebas yaitu melalui fungsi *likelihood* dimana fungsi log *likelihood* dijelaskan pada persamaan (2.25).

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_{1i} g_1(\mathbf{x}_i) + y_{2i} g_2(\mathbf{x}_i) - \ln\left(1 + e^{g_1(\mathbf{x}_i)} + e^{g_2(\mathbf{x}_i)}\right) \\ &= \sum_{i=1}^{n} y_{1i}\left(\beta_{10} + \sum_{k=1}^{p}\beta_{1k}x_{ik}\right) + y_{2i}\left(\beta_{20} + \sum_{k=1}^{p}\beta_{2k}x_{ik}\right) \\ &\quad - \ln\left\{1 + \exp\left(\beta_{10} + \sum_{k=1}^{p}\beta_{1k}x_{ik}\right) + \exp\left(\beta_{20} + \sum_{k=1}^{p}\beta_{2k}x_{ik}\right)\right\} \end{aligned} \tag{2.25}$$

Estimasi parameter β_{10} yang merupakan intersep atau konstanta dari $g_1(\mathbf{x}_i)$ adalah sebagai berikut.

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{10}} &= \sum_{i=1}^{n} y_{1i} - \frac{\exp\left(\beta_{10} + \sum_{k=1}^{p}\beta_{1k}x_{ik}\right)}{1 + \exp\left(\beta_{10} + \sum_{k=1}^{p}\beta_{1k}x_{ik}\right) + \exp\left(\beta_{20} + \sum_{k=1}^{p}\beta_{2k}x_{ik}\right)} \\ &= \sum_{i=1}^{n} y_{1i} - \pi_1(x_i) \end{aligned} \tag{2.26}$$

sedangkan untuk parameter β_{11} yang merupakan koefisien variabel x_i dari $g_1(\mathbf{x}_i)$ pada persamaan (2.27).

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{11}} &= \sum_{i=1}^{n} y_{1i}x_{i1} - \frac{\exp\left(\beta_{10} + \beta_{11}x_{i1} + \sum_{k=2}^{p}\beta_{1k}x_{ik}\right)x_{i1}}{1 + \exp\left(\beta_{10} + \sum_{k=1}^{p}\beta_{1k}x_{ik}\right) + \exp\left(\beta_{20} + \sum_{k=1}^{p}\beta_{2k}x_{ik}\right)} \\ &= \sum_{i=1}^{n} y_{1i}x_{i1} - \left(\pi_1(x_i)\right)x_{i1} \\ &= \sum_{i=1}^{n} x_{i1}\left(y_{1i} - \pi_1(x_i)\right) \end{aligned} \tag{2.27}$$

Bentuk secara umum turunan parsial pertama fungsi log *likelihood* terhadap parameter yang akan diestimasi β_{jk} adalah sebagai berikut.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ik}\left(y_{ji} - \pi_{ji}\right) \tag{2.28}$$

dimana j = 1,2 dan k = 0, 1, 2, …, p dengan x_{i0} bernilai 1.

Nilai G pada statistik uji *Likelihood Ratio Test* mengikuti distribusi *Chi-Square* dengan derajat bebas 2k. Jika $G \geq \chi^2_{2k,\alpha}$ maka dapat diputuskan untuk menolak H_0 yang artinya minimal terdapat satu variabel independen atau prediktor yang berpengaruh signifikan terhadap variabel dependen atau respon. Selanjutnya dapat dilakukan uji signifikansi parameter secara parsial untuk mengetahui variabel-variabel independen mana yang berpengaruh signifikan terhadap variabel dependen. Pengujian signifikansi parameter secara parsial dapat dilakukan dengan *Wald Test* dengan hipotesis sebagai berikut.

$H_0 : \beta_{ij} = 0$

$H_1 : \beta_{ij} \neq 0$, dengan i = 1, 2 dan j = 1,2, … , k.

Statistik uji *Wald Test* dapat dihitung dengan rumus pada persamaan (2.29).

$$W^2 = \left[\frac{\hat{\beta}_{ij}}{se\left(\hat{\beta}_{ij}\right)}\right] \tag{2.29}$$

dimana

$$se\left(\hat{\beta}_{ij}\right) = \sqrt{\text{var}(\hat{\beta}_j)}$$

$$\mathbf{H}(\boldsymbol{\beta}) = \begin{bmatrix} \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} & \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^T} \\ \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1^T} & \dfrac{\partial^2 \ln L(\cdot)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^T} \end{bmatrix} = \begin{bmatrix} \text{var}\left(\hat{\beta}_1^T\right) & \text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ \text{cov}\left(\hat{\beta}_2, \hat{\beta}_1\right) & \text{var}\left(\hat{\beta}_2^T\right) \end{bmatrix}$$

Daerah penolakan H_0 adalah $W^2 > \chi^2_{1,\alpha}$ atau nilai P-value < α, sehingga dapat disimpulkan bahwa variabel prediktor secara parsial berpengaruh signifikan terhadap variabel respon (Hosmer & Lemeshow, 2000).

2.5.3 Uji Kesesuaian Model

Setelah estimasi model regresi logistik diperoleh, selanjutnya menguji seberapa besar kesesuaian model dalam menjelaskan variabel respon (Hosmer & Lemeshow, 2000) yang disebut dengan uji *goodness of fit* (uji kesesuaian model). Uji ini dilakukan dengan tujuan untuk mengetahui apakah tidak ada perbedaan antara hasil observasi dengan kemungkinan hasil prediksi model. Uji kesesuaian model memiliki hipotesis pengujian sebagai berikut.

H_0 : Model sesuai (tidak terdapat berbedaan antara hasil observasi dengan kemungkinan prediksi model)

H_1 : Model tidak sesuai (terdapat berbedaan antara hasil observasi dengan kemungkinan prediksi model)

Statistik uji yang digunakan yaitu pada persamaan (2.30) sebagai berikut.

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(O_k - n_k' \bar{\pi}_k\right)^2}{n_k' \bar{\pi}_k \left(1 - \bar{\pi}_k\right)} \tag{2.30}$$

dengan $\bar{\pi}_k = \left(\sum_{j=1}^{C_k} \frac{m_j \hat{\pi}_j}{n_k'} \right)$

Keterangan :

O_k : Observasi grup ke-k

$\bar{\pi}_k$: Estimasi probabilitas rata-rata

g : Jumlah grup (kombinasi kategori dalam model serentak)

n_k' : Banyak observasi pada grup ke-k

Pengambilan keputusan didasarkan pada tolak H_0 jika $\chi^2_{hitung} \geq \chi^2_{(db,\alpha)}$ dengan $db = g - 2$

2.5.4 *Odds Ratio*

Odds ratio merupakan suatu ukuran untuk mengetahui tingkat resiko (kecenderungan) yaitu perbandingan antara peluang dua variabel prediktor X_j, antara kejadian-kejadian yang masuk kategori sukses dan gagal (Hosmer & Lemeshow, 2000). Untuk menyederhanakan estimasi dan interpretasi *odds ratio* pada bentuk respon multinomial diperlukan generalisasi notasi yang digunakan pada kasus dengan respon biner termasuk respon yang dibandingkan seperti halnya nilai kovariat.

Diasumsikan bahwa *outcome* dengan label Y=0 merupakan *outcome* permbanding (*reference*). Indeks pada *odds ratio* mengindikasikan perbandingan terhadap *outcome*. *Odds ratio* untuk Y = j dengan Y = 0 pada nilai kovariat x = a dengan x = b yaitu pada persamaan (2.31).

$$OR_j(a,b) = \frac{P(Y = j \mid x = a)/P(Y = 0 \mid x = a)}{P(Y = j \mid x = b)/P(Y = 0 \mid x = b)} \tag{2.31}$$

Berdasarkan model logit, misalnya ingin melihat *odds ratio* yang membandingkan dua level dari F, $F = f_1$ terhadap $F = f_0$, pada $X = x$. Maka *odds ratio* diperoleh melalui tahapan berikut.

$$g(f_1, x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 (f_1 x) \tag{2.32}$$

dan

$$g(f_0, x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 (f_0 x) \tag{2.33}$$

Selanjutnya dihitung selisihnya untuk mendapatkan log *odds ratio* pada persamaan (2.34) berikut.

$$\begin{aligned} &\ln\left[OR(F = f_1, F = f_0, X = x)\right] \\ &\quad = g(f_1, x) - g(f_0, x) = \beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0) \end{aligned} \tag{2.34}$$

Kemudian nilai *odds ratio* bisa didapatkan dengan cara meng-eksponensialkan selisih tersebut sehingga hasilnya pada persamaan (2.35)

$$OR = \exp\left[\beta_1\left(f_1 - f_0\right) + \beta_3 x\left(f_1 - f_0\right)\right] \tag{2.35}$$

2.5.5 Evaluasi Ketepatan Klasifikasi

Evaluasi ketepatan klasifikasi adalah suatu evaluasi yang melihat peluang kesalahan yang dilakukan oleh suatu fungsi klasifikasi (Johnson & Winchern, 1992). Nilai APER (*Apparent Error Rate*) menyatakan nilai proporsi sampel yang salah diklasifikasikan oleh fungsi klasifikasi. Penentuan ketetepatan pengklasifikasian dapat diketahui melalui tabel 2.5 sebagai berikut.

Tabel 2.5 Perhitungan Ketepatan Pengklasifikasian

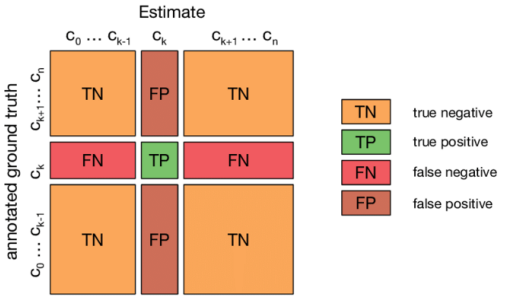
Actual Membership	***Predicted Membership***			***Total***
	$y = 1$	$y = 2$	$y = 3$	
$y = 1$	n_{11}	n_{12}	n_{13}	$N_{1.}$
$y = 2$	n_{21}	n_{22}	n_{23}	$N_{2.}$
$y = 3$	n_{31}	n_{32}	n_{33}	$N_{3.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	N_{total}

Keterangan :

n_{11} : Jumlah Y_i dari y = 1 tepat diklasifikasikan sebagai y = 1
n_{12} : Jumlah Y_i dari y = 1 tepat diklasifikasikan sebagai y = 2
n_{13} : Jumlah Y_i dari y = 1 tepat diklasifikasikan sebagai y = 3
n_{21} : Jumlah Y_i dari y = 2 tepat diklasifikasikan sebagai y = 1
n_{22} : Jumlah Y_i dari y = 2 tepat diklasifikasikan sebagai y = 2
n_{23} : Jumlah Y_i dari y = 2 tepat diklasifikasikan sebagai y = 3
n_{31} : Jumlah Y_i dari y = 3 tepat diklasifikasikan sebagai y = 1
n_{32} : Jumlah Y_i dari y = 3 tepat diklasifikasikan sebagai y = 2
n_{33} : Jumlah Y_i dari y = 3 tepat diklasifikasikan sebagai y = 3

Untuk mendapatkan klasifikasi yang lebih optimal dan lebih spesifik maka dapat dilihat nilai *sensitivity* dan *specificity*. *Sensitivity* adalah tingkat positif benar atau ukuran performansi

untuk mengukur kelas positif (minor) sedangkan *specificity* adalah tingkat negatif benar atau ukuran performansi untuk mengukur kelas negatif (mayor). Jika *confusion matrix* multiclass, maka perlu dilakukan pembentukan *confusion matrix biner* dengan pembagian pada Gambar 2.3.



Gambar 2.3 *Komponen Confusion Matrix*

Selanjutnya didapatkan nilai *sensitivity* dan *specificity* yang terdapat pada persamaan (2.36) dan (2.37) sebagai berikut.

$$Sensitivity_{class} = \frac{TP_{class}}{(TP_{class} + FN_{class})} \tag{2.36}$$

$$Specificity_{class} = \frac{TN_{class}}{(TN_{class} + FP_{class})} \tag{2.37}$$

Akurasi dapat dihitung dengan membagi jumlah pengamatan yang diklasifikasikan benar oleh total pengamatan. Berikut merupakan rumus perhitungan AUC pada klasifikasi biner dan AUC pada klasifikasi *multiclass* (Hand & Till, 2001).

$$\hat{A}(c_i|c_j) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\Psi(\rho_i > \rho_j) \tag{2.38}$$

$$\Psi(\rho_i > \rho_j) = \begin{cases} 1, \rho_i > \rho_j \\ \frac{1}{2}, \rho_i = \rho_j \\ 0, \rho_i < \rho_j \end{cases} \tag{2.39}$$

$$AUC(c_i, c_j) = \frac{\hat{A}(c_i|c_j) + \hat{A}(c_j|c_i)}{2} \tag{2.40}$$

$$AUC_{total} = \frac{2}{C(C-1)} \sum_{i<j} AUC(c_i, c_j) \quad (2.41)$$

dimana :

i = 1, 2, ... , m

j = 1, 2, … , n

Keterangan :

ρ_i = Peluang suatu observasi dengan kelas positif ke-*k* diklasifikasikan ke kelas positif

ρ_j = Peluang suatu observasi dengan kelas negatif ke-*l* diklasifikasikan ke kelas positif

m = Jumlah observasi kelas positif

n = Jumlah observasi kelas negatif

C = Jumlah kelas pada klasifikasi *multiclass*

2.6 *Indonesia Family Life Survey* (IFLS)

Indonesia Family Life Survey (IFLS) atau Survei Aspek Kehidupan Rumah Tangga (SAKERTI) merupakan survei longitudinal yang paling komprehensif yang pernah dilakukan di Indonesia. Survei ini diadakan atas kerja sama antara organisasi penelitian Amerika Serikat RAND, Lembaga Demografi Universitas Indonesia, Pusat Studi Kependudukan dan Kebijakan Universitas Gajah Mada serta lembaga penelitian SurveyMETER. IFLS terdiri dari dua survei yaitu survei rumah tangga (*household survey*) dan survei komunitas dan fasilitas (*Community Facility Survey*). Setiap rumah tangga dan individu disurvei dengan menggunakan kuisioner *Household* (HH). Pada *Community Facility Survey* (CFS) sampel diambil dari fasilitas-fasilitas atau kondisi sekitar rumah tangga dengan target responden yaitu pemimpin desa atau lurah.

Pertanyaan-pertanyaan dalam kuisioner IFLS dikategorikan menjadi beberapa buku yang terbagi dalam beberapa modul. Untuk survei rumah tangga terdapat tiga buku utama yaitu buku K, buku 1 dan buku 2. Secara umum ketiga buku tersebut menggambarkan karakteristik sosiao-demografi dan ekonomi rumah tangga. Masing masing buku dibagi menjadi beberapa seksi. Untuk tingkat individu, terdapat empat buku yaitu buku dengan responden

anggota rumah tangga dewasa (buku 3A dan 3B), perempuan yang menikah atau pernah menikah (buku 4), dan anak-anak usia 15 tahun ke bawah (buku 5). Sedangkan untuk *Community Facility Survey*, kuisioner terbagi menjadi tiga kategori yaitu kuisioner komunitas, fasilitas kesehatan dan fasilitas pendidikan. Kuisioner komunitas dibagi menjadi enam buku, kuisioner fasilitas kesehatan menjadi tiga buku, dan kuisioner fasilitas pendidikan tiga buku.

IFLS telah dilakukan selama 5 gelombang yaitu pada tahun 1993, 1997, 2000, 2007 dan 2014. Gelombang pertama atau IFLS 1 dilakukan pada agustus 1993 hingga januari 1994. Survei ini mewakili sekitar 83% dari populasi Indonesia. Pemilihan sampel dilakukan secara acak mengikuti Survei Sosial Ekonomi Nasional (SUSENAS) 1993 dan terpilih sebanyak 321 daerah yang tersebar di 13 provinsi di Indonesia. Ke 13 provinsi itu meliputi 4 provinsi di Sumatera (Sumatera Utara, Sumatera Barat, Jambi, Lampung), 5 Provinsi di Jawa (DKI Jakarta, Jawa Barat, Jawa Tengah, Yogyakarta, Jawa Timur), Bali, Nusa Tenggara Barat, Kalimantan Selatan dan Sulawesi Selatan.

Dalam IFLS-1 dilakukan survei ke 7224 rumah tangga dengan jumlah individu sebanyak 22000 orang. IFLS-2 ditindaklanjuti dengan sampel yang sama pada tahun 1997 dengan *persantase re-contact* sebesar 94.4%. Pada januari 1998, dilakukan survei kembali untuk mengetahui dampak krisis ekonomi Asia yang melanda Indonesia. Sampel yang digunakan hanya 25% dari sampel IFLS-2 karena keterbatasan waktu dan Sumber Daya Manusia (SDM). Gelombang 3 dilakukan pada juni hingga november 2000. Responden sebanyak 10574 rumah tangga dengan persentase sebesar 95.3%. IFLS-4 dilakukan pada november 2007 hingga mei 2008 dan IFLS 5 pada september 2014 hingga maret 2015. Persentase *re-contact* responden IFLS 4 dan 5 secara berturut-turut sebesar 93.6% dan 92%. IFLS-5 mememiliki kelebihan dibandingkan survei sebelumnya yaitu IFLS-5 telah menggunakan sistem *Computer-Assisted Personal Interview* (CAPI) dan tidak lagi menggunakan kuesioner kertas. Program CAPI telah dpersiapkan dan diuji coba selama kurang lebih 18

bulan. Selain itu, pengambilan data pada IFLS-5 juga telah menggunakan alat perekam suara sehingga kualitas data dapat terkontrol dengan baik (Strauss, Witoelar, & Sikoki, 2016).

2.7 Asuransi

Asuransi adalah mekanisme proteksi atau perlindungan dari risiko kerugian keuangan dengan cara mengalihkan risiko kepada pihak lain. Menurut Undang-Undang No 40 tahun 2014 tentang perasuransian memberikan definisi asuransi yaitu perjanjian antara dua pihak, yaitu perusahaan asuransi dan pemegang polis, yang menjadi dasar bagi penerimaan premi oleh perusahaan asuransi sebagai imbalan untuk beberapa hal sebagai berikut.

a. Memberikan pergantian kepada tertanggung atau pemegang polis karena kerugian, kerusakan, biaya yang timbul, kehilangan keuntungan, atau tanggung jawab hukum kepada pihak ketiga yang mungkin diderita tertanggung atau pemegang polis karena terjadinya suatu peristiwa yang tidak pasti
b. Memberikan pembayaran yang didasarkan pada meningggalnya tertanggung atau pembayaran yang didasarkan pada hidupnya tertanggung dengan manfaat yang besarnya telah ditetapkan dan/atau didasarkan pada hasil pengelolaan dana

Dari segi kepemilikannya, asuransi kepemilikan yaitu pemegang polis terhadap tuntutan orang lain akan barang miliknya. Adapun bentuk kepemilikan perusahaan asuransi menurut Salim (2006) dapat dilihat dari dua faktor yaitu sebagai berikut.

1. Perusahaan Swasta
   a. *Stock company* yaitu perusahaan didirikan dengan menjual stock atau saham di pasar bursa
   b. *Mutual company* yaitu perusahaan didirikan untuk dan dari *policy holder*, jadi sama dengan bentuk koperasi
2. Perusahaan Negara atau Pemerintah

Di Indonesia sebagian besar perusahaan asuransi dimiliki oleh pemerintah. Dalam hal ini pemerintah yang memegang semua kebijaksanaan serta pelaksanaan operasional perusahaan-perusahaan milik negara tersebut. Tujuan perusahaan disini selain untuk menaikkan kesejahteraan sosial masyarakat, juga sebagai lembaga penabungan, untuk menghimpun modal yang bisa digunakan sebagai sumber-sumber pembelanjaan di dalam pembangunan ekonomi Indonesia.

Secara garis besar, perbedaan antara asuransi milik pemerintah dan asuransi milik publik yaitu pada Tabel 2.6.

Tabel 2.6 Perbedaan Asuransi Milik Pemerintah dan Swasta

Perbedaan	Asuransi Pemerintah	Asuransi Swasta
Saham	Asuransi yang sahamnya dimiliki sebagian besar atau bahkan 100% oleh pemerintah Indonesia	Asuransi yg kepemilikan sahamnya sepenuhnya dimiliki oleh swasta
Biaya	Biaya lebih murah karena adanya bantuan subsidi dari pemerintah	Asuransi swasta harga preminya *pure* dibayar oleh nasabah sendiri
Layanan	Layanan harus melalui rujukan dari puskesmas terlebih dahulu,	Layanan yang lebih cepat karena asuransi swasta memiliki premi yang lebih mahal

Faktor-faktor yang berhubungan dengan kepemilikan asuransi kesehatan adalah sebagai berikut.

1. Usia

Usia akan berpengaruh terhadap resiko kesehatan serta *demand* terhadap asuransi kesehatan. Seseorang yang berusia tua akan lebih sering sakit dibandingkan dengan yang muda, sehingga risiko sakitnya akan berbeda. Seseorang yang berusia lebih tua akan lebih sering sakit dibandingkan dengan yang muda, sehingga risiko sakitnya akan berbeda dan akan berpengaruh pada *demand* akan asuransi kesehatan (HIAA, 2000).

2. Jenis Kelamin

Jenis kelamin juga berpengaruh terhadap resiko kesehatan serta *demand* terhadap asuransi. Hal ini berkaitan dengan perbedaan kekebalan tubuh antara laki-laki dan perempuan. Selain itu, angka harapan hidup juga berpengaruh dimana angka harapan hidup perempuan lebih tinggi daripada laki-laki (HIAA, 2000).

3. Pendidikan

Tingkat pendidikan mempengaruhi seseorang dalam berasuransi karena semakin tinggi tingkat pendidikan maka pengetahuan akan keuangan menjadi semakin luas. Sehingga kesadaran, niat dan kepemilikan asuransi jiwa juga semakin besar. Pendidikan memang mempengaruhi tingkat kesadaran akan asuransi, hal ini sejalan dengan penelitian yang dilakukan oleh Kumar et al. (2011) yang menyatakan pendidikan berpengaruh signifikan terhadap *corp insurance* di India. Pengetahuan dan pemahaman seseorang tentu dipengaruhi oleh pendidikannya.

4. Pekerjaan

Pekerjaan akan berpengaruh pada faktor risiko kesehatan dimana setiap pekerjaan akan mempunyai faktor resiko yang berbeda. Pekerjaan akan berpengaruh pada premi yang akan dikenakan pada pembeli sehingga akan mempengaruhi demand terhadap asuransi. Variabel ini dapat mengukur kesanggupan seseorang atau sekelompok orang untuk memperoleh pelayanan kesehatan (Feldstein, 1988).

5. Pendapatan

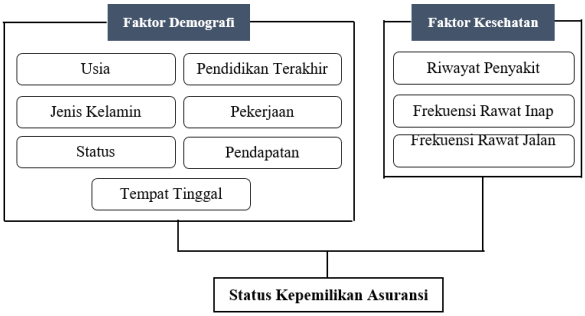
Pendapatan mempengaruhi keputusan kepemilikan asuransi jiwa, asuransi jiwa adalah salah satu investasi proteksi dalam keluarganya. Besarnya pendapatan atau penghasilan seseorag dapat mempengaruhi *demand* terhadap asuransi dimana semakin meningkatnya pendapatan seseorang maka kemampuan membayar premi akan semakin besar (Feldstein, 1988).

6. Kondisi Kesehatan

Resiko sakit yang timbul dari masalah kesehatan adalah ketidaknyamanan fisik dan mental, pengeluaran biaya kesehatan dan hilangnya produktivitas atau pendapatan karena tidak bisa bekerja. Kondisi kesehatan dan riwayat penyakit akan menentukan

premi asuransi yang akan dikenakan pada calon peserta. Sesorang akan membutuhkan asuransi kesehatan kalau orang tersebut menyadari bahwa ia mempunya risiko untuk jatuh sakit dan akan mengalami kerugian finansial akibat sakit tersebut.

Pada dasarnya yang mempengaruhi kepemilikan asuransi kesehatan yaitu didasarkan pada dua faktor yaitu faktor demografi dan faktor kesehatan. Faktor demografi meliputi usia, jenis kelamin, pendidikan, pekerjaan, pendapatan, status, tempat tinggal dan region. Faktor kesehatan meliputi adakah riwayat penyakit yang diderita serta frekuensi rawat inap dan rawat jalan.



Gambar 2.4 Gambaran Faktor yang Mempengaruhi Status Kepemilikan Asuransi

(Halaman ini sengaja dikosongkan)

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder tentang kepemilikan asuransi kesehatan dan faktor-faktor yang diduga mempengaruhinya. Data tersebut diperoleh dari *Indonesia Family Life Survey* (IFLS) oleh RAND Corporation yang berasal dari https://www.rand.org. Pada penelitian ini hanya menggunakan IFLS gelombang 5 dengan responden sebanyak 89372 individu yang terbagi dalam 15921 rumah tangga. Unit penelitian yang digunakan yaitu individu yang telah memasuki usia kerja yaitu berumur 15 tahun keatas.

Data pada IFLS yang digunakan yaitu pada buku K dan buku 3B, dimana pada setiap seksi data tersebut terdapat kode rumah tangga dan kode individu. Penjelasan lebih lanjut mengenai kode tersebut yaitu sebagai berikut.

1. Kode rumah tangga (HHid)

   Kode untuk rumah tangga terdiri dari 7 digit yaitu (XXX + HH + SS) dimana XXX merupakan tiga digit angka yang menunjukkan wilayah atau area numerasi. HH menunjukkan kode rumah tangga, dan SS merupakan kode rumah tangga pecahan atau panel dengan kode sebagai berikut.

   00 : Rumah tangga panel
   11 : IFLS-2
   21 : IFLS-2+
   31 : IFLS-3
   41 : IFLS-4

2. Kode individu (*Pidlink)*

   Kode ini biasa disebut dengan *personal idlink* yang merupakan gabungan dari dua kode yaitu HHid + AR01. HHid yang merupakan kode rumah tangga dan AR01 kode tiap responden.

3.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini terdiri dari 11 variabel prediktor dan sebuah variabel respon yang dijelaskan sebagai berikut.

1. Variabel respon (Y) adalah variabel terikat atau variabel yang berubah bergantung pada perubahan variabel bebas. Pada penelitian ini variabel respon yang digunakan yaitu kepemilikan asuransi yang selanjutnya dikategorikan menjadi tiga yaitu :
   0 : jika responden tidak memiliki asuransi
   1 : jika responden menggunakan asuransi milik pemerintah
   2 : jika responden menggunakan asuransi milik swasta
2. Variabel prediktor (X) terdiri dari variabel-variabel yang diduga berpengaruh terhadap variabel respon, yaitu :
   a. Usia (X_1), menyatakan usia responden *Indonesia Family Life Survey* (IFLS) gelombang ke 5. Variabel usia memiliki skala rasio.
   b. Jenis kelamin (X_2), menyatakan jenis kelamin dari responden yang memiliki skala nominal.
      0 : jika responden berjenis kelamin laki-laki
      1 : jika responden berjenis kelamin perempuan
   c. Pendidikan Terakhir (X_3), menyatakan pendidikan terakhir responden pada saat survey ini dilakukan. Pendidikan terakhir memiliki skala ordinal.
      0 : SD sederajat
      1 : SMP sederajat
      2 : SMA sederajat
      3 : Perguruan tinggi
   d. Pekerjaan (X_4), menyatakan jenis pekerjaan yang dimiliki responden pada saat dilakukan survey IFLS. Variabel tipe pekerjaan memiliki skala nominal.
      0 : Sekolah
      1 : Bekerja
      2 : Mengurus rumah tangga
      3 : Pensiun

4 : Menganggur

e. Pendapatan (X_5), variabel ini menyatakan jumlah pendapatan responden dengan skala ordinal.
   0 : pendapatan < Rp 1.000.000,00
   1 : pendapatan Rp 1.000.000,00 - Rp 2.000.000,00
   2 : pendapatan Rp 2.000.000,00 - Rp 3.000.000,00
   3 : pendapatan > Rp 3.000.000,00
f. Status (X_6), menyatakan status responden yang memiliki skala nominal. Status ini terdiri dari beberapa kategori sebagai berikut.
   0 : Belum Kawin
   1 : Kawin
   2 : Cerai hidup/mati
g. Tempat tinggal (X_7), variabel ini memiliki skala nominal. Pada *Indonesia Family Life Survey* (IFLS) dilakukan survei dengan perbandingan tempat tinggal sebesar 2 :3.
   0 : Pedesaan
   1 : Perkotaan
h. Riwayat penyakit (X_8), menyatakan apakah terdapat riwayat penyakit pada responden. Variabel ini memiliki skala nominal.
   0 : jika memiliki riwayat penyakit
   1 : jika tidak memiliki riwaya penyakit
i. Frekuensi kunjungan rawat inap (X_9), menyatakan jumlah kunjungan rawat inap yang pernah dilakukan responden selama 12 bulan terakhir. Variabel ini memiliki skala rasio.
j. Frekuensi kunjungan rawat jalan (X_{10}), menyatakan jumlah kunjungan rawat jalan dengan skala rasio.

3.3 Struktur Data

Struktur data untuk penelitian ini disajikan pada Tabel 3.1, dimana variabel respon merupakan kepemilikan asuransi kesehatan di Indonesia.

Tabel 3.1 Struktur Data

Y	**X_1**	**X_2**	**…**	**X_{10}**
Y_1	X_{11}	X_{21}	…	X_{101}
Y_2	X_{12}	X_{22}	…	X_{102}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Y_i	X_{1i}	X_{2i}	…	X_{10i}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Y_n	X_{1n}	X_{2n}	…	X_{10n}

Data *Indonesia Family Life Survey* (IFLS) terdiri dari beberapa buku dan seksi, dimana dalam setiap seksi terdapat kode rumah tangga (HHid) dan kode individu atau *personal idlink* (Pidlink). Dalam penelitian ini menggunakan buku dan seksi pada Tabel 3.2.

Tabel 3.2 Kode Variabel Penelitian

Variabel	**Buku**	**Seksi**	**Kode Pertanyaan**
Y	3B	AK	AK01
X_1	K	AR	AR09
X_2			AR07
X_3			AR16
X_4			AR15C
X_5			AR15B
X_6			AR13
X_7		SC	SC05
X_9	3B	CD	CD01
X_{10}		RN	RN02
X_{11}		RJ	RJ02

3.4 Langkah Analisis

Langkah-langkah analisis yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Mengidentifikasi karakteristik data status kepemilikan asuransi pada *Indonesia Family Life Survey* (IFLS) serta faktor-faktor yang mempengaruhinya.
2. Menganalisis status kepemilikan asuransi dengan metode *multinomial logistic regression* pada *imbalanced data* dan *balanced data.*
   a. *Imbalanced data*
      i. Melakukan pengecekan asumsi meliputi deteksi multikolinieritas dan independensi terhadap masing-masing variabel (Subbab 2.2 dan 2.3)
      ii. Melakukan pengujian signifikansi parameter secara serentak untuk mengetahui paling tidak terdapat satu variabel prediktor yang berpengaruh terhadap variabel respon. Selanjutnya dilakukan pengujian signifikansi parameter secara parsial (Subbab 2.5.2).
      iii. Pemodelan regresi logistik multinomial terhadap variabel yang berpengaruh signifikan, dan dilanjutkan dengan interpretasi *odds ratio* (Subbab 2.5.4).
      iv. Uji kesesuaian model untuk melihat model yang dibentuk sudah baik atau belum (Subbab 2.5.3).
      v. Menghitung ketepatan klasifikasi dari model untuk mengetahui seberapa besar observasi secara tepat diklasifikasikan (Subbab 2.5.5).
   b. *Balanced data* dengan SMOTE
      i. Melakukan *balancing* data untuk menyeimbangkan kelas mayor dan minor dengan SMOTE (Subbab 2.4.1).
      ii. Melakukan pengecekan asumsi meliputi deteksi multikolinieritas dan independensi terhadap masing-masing variabel (Subbab 2.2 dan 2.3).

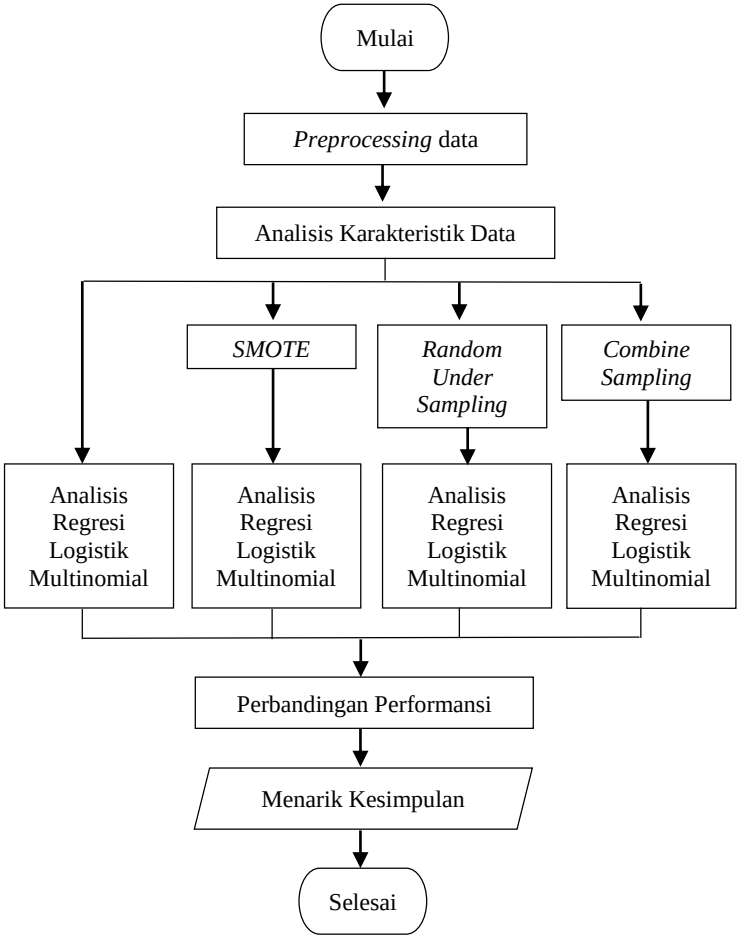
iii. Melakukan pengujian signifikansi parameter secara serentak untuk mengetahui paling tidak terdapat satu variabel prediktor yang berpengaruh terhadap variabel respon. Selanjutnya dilakukan pengujian signifikansi parameter secara parsial (Subbab 2.5.2).
iv. Pemodelan regresi logistik multinomial terhadap variabel yang berpengaruh signifikan, dan dilanjutkan dengan interpretasi *odds ratio* (Subbab 2.5.4).
v. Uji kesesuaian model untuk melihat model yang dibentuk sudah baik atau belum (Subbab 2.5.3).
vi. Menghitung ketepatan klasifikasi dari model untuk mengetahui seberapa besar observasi secara tepat diklasifikasikan (Subbab 2.5.5).

c. *Balanced data* dengan *undersampling*
i. Melakukan *balancing* data terlebih dahulu untuk menyeimbangkan antara kelas mayor dan minor dengan *random undersampling* (Subbab 2.4.2).
ii. Melakukan pengecekan asumsi meliputi deteksi multikolinieritas dan independensi terhadap masing-masing variabel (Subbab 2.2 dan 2.3).
iii. Melakukan pengujian signifikansi parameter secara serentak untuk mengetahui paling tidak terdapat satu variabel prediktor yang berpengaruh terhadap variabel respon. Selanjutnya dilakukan pengujian signifikansi parameter secara parsial (Subbab 2.5.2).
iv. Pemodelan regresi logistik multinomial terhadap variabel yang berpengaruh signifikan, dan dilanjutkan dengan interpretasi *odds ratio* (Subbab 2.5.4).
v. Uji kesesuaian model untuk melihat model yang dibentuk sudah baik atau belum (Subbab 2.5.3).
vi. Menghitung ketepatan klasifikasi dari model untuk mengetahui seberapa besar observasi secara tepat diklasifikasikan (Subbab 2.5.5).

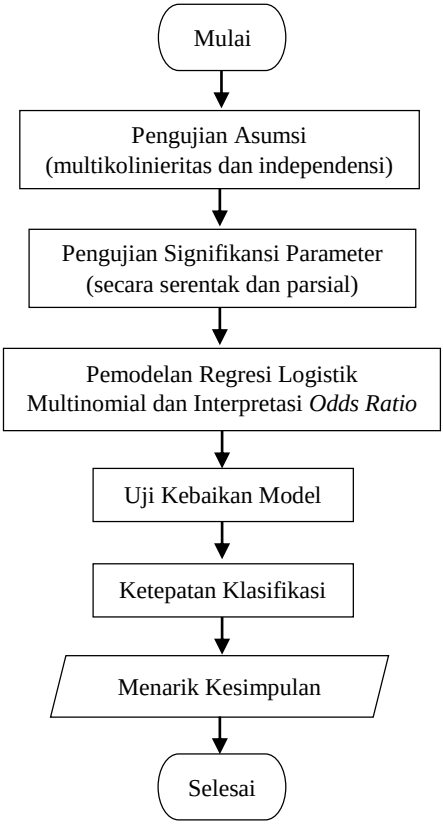
d. *Balanced data* dengan *combine sampling*
   i. Melakukan *balancing* data terlebih dahulu untuk menyeimbangkan antara kelas mayor dan minor dengan *combine sampling* (Subbab 2.4.1 dan 2.4.2).
   ii. Melakukan pengecekan asumsi meliputi deteksi multikolinieritas dan independensi terhadap masing-masing variabel (Subbab 2.2 dan 2.3).
   iii. Melakukan pengujian signifikansi parameter secara serentak untuk mengetahui paling tidak terdapat satu variabel prediktor yang berpengaruh terhadap variabel respon. Selanjutnya dilakukan pengujian signifikansi parameter secara parsial (Subbab 2.5.2).
   iv. Pemodelan regresi logistik multinomial terhadap variabel yang berpengaruh signifikan, dan dilanjutkan dengan interpretasi *odds ratio* (Subbab 2.5.4).
   v. Uji kesesuaian model untuk melihat model yang dibentuk sudah baik atau belum (Subbab 2.5.3).
   vi. Menghitung ketepatan klasifikasi dari model untuk mengetahui seberapa besar observasi secara tepat diklasifikasikan (Subbab 2.5.5).

3. Membandingkan performansi dari hasil analisis regresi logistik multinomial *imbalanced data* dan *balanced data* dengan *prepocessing* SMOTE, *undersampling* dan *combine sampling.*

Berdasarkan langkah-langkah penelitian yang akan dilakukan dan penjelasan sebelumnya, pada penelitian ini akan menggunakan dua diagram alir. Diagram alir untuk analisis ini secara general ditamplilkan pada Gambar 3.1 dan diagram alir untuk tahapan regresi logistik multinomial dapat dilihat pada Gambar 3.2.



Gambar 3.1 Diagram Alir Penelitian



Gambar 3.2 Diagram Alir Regresi Logistik Multinomial

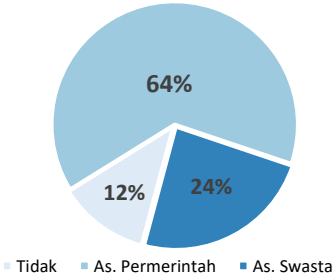
(Halaman ini sengaja dikosongkan)

BAB IV
ANALISIS DAN PEMBAHASAN

Pada penelitian ini dilakukan analisis mengenai status kepemilikan asuransi dengan metode *multinomial logistic regression*. Pada analisis tersebut *prepocessing* data dilakukan untuk mengatasi *imbalance multiclass* pada variabel respon menggunakan metode *Synthetic Minority Oversampling Technique* (SMOTE), *Random Undersampling* (RUS) dan *combine sampling*. Selanjutnya dilakukan perbandingan kebaikan hasil klasifikasi dari setiap metode menggunakan nilai *area under curve* (AUC).

4.1 Karakteristik Data Status Kepemilikan Asuransi

Status kepemilikan asuransi di Indonesia yang diamati dalam penelitian ini meliputi asuransi pemerintah, asuransi swasta dan tidak memiliki asuransi. Status kepemilikan asuransi di Indonesia disajikan pada Gambar 4.1.

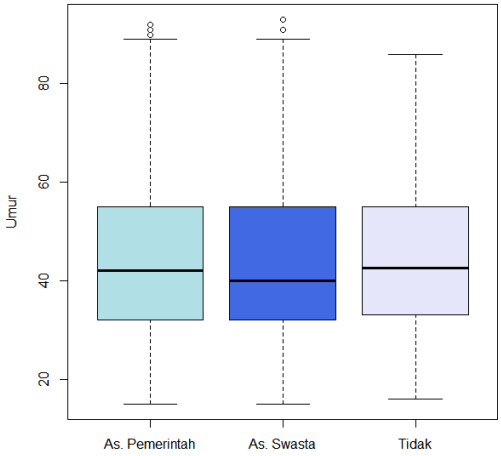


Gambar 4.1 Karakteristik Status Kepemilikan Asuransi

Gambar 4.1 menunjukkan informdasi bahwa dari 6855 responden yang disurvei, sebanyak 12% atau 682 responden tidak memiliki asuransi dan sisanya telah memiliki asuransi baik asuransi pemerintah maupun asuransi swasta. Asuransi pemerintah lebih mendominasi dikalangan masyarakat Indonesia dengan persentase sebesar 64% dibandingkan dengan asuransi swasta yang hanya 24% atau 1600 responden. Perbedaan kedua asuransi tersebut yaitu dari segi harga dan pelayanan. Asuransi swasta

memiliki pelayanan yang lebih cepat daripada asuransi pemerintah karena premi dibayarkan oleh nasabah sendiri. Sedangkan asuransi pemerintah terkenal dengan biaya yang murah karena adanya bantuan subsidi dari pemerintah. Sebanyak 64% masyarakat Indonesia lebih memilih menggunakan asuransi pemerintah dengan alasan harga lebih murah meskipun pelayanan yang diberikan tidak sebaik asuransi swasta.

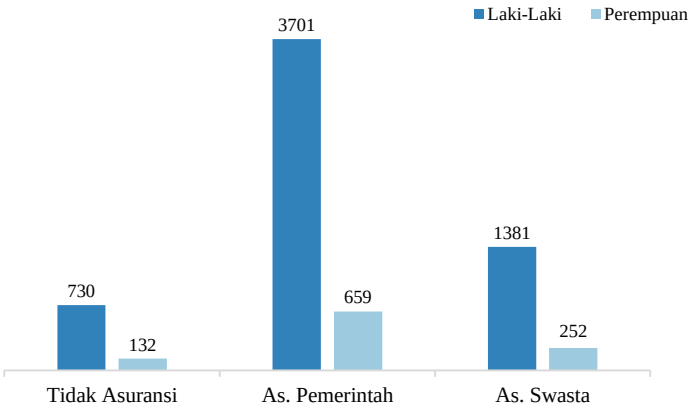
Selanjutnya akan dilakukan karakteristik data terhadap faktor-faktor yang mempengaruhi status kepemilikan asuransi meliputi umur, jenis kelamin, pendidikan, pekerjaan, status, tempat tinggal, dan frekuensi rawat inap serta rawat jalan. Gambar 4.2 merupakan hasil visualisasi data faktor umur terhadap status kepemilikan asuransi di Indonesia.



Gambar 4.2 *Boxplot* Umur Berdasarkan Status Kepemilikan Asuransi

Gambar 4.2 menunjukkan bahwa persebaran umur berdasarkan status kepemilikan asuransi tidak terlihat. *Boxplot* tersebut memberikan informasi bahwa asuransi swasta memiliki nilai median lebih rendah dibandingkan dengan asuransi pemerintah dan tidak memiliki asuransi. Berdasarkan persebaran umurnya, status kepemilikan asuransi memusat diantara 30 hingga 55 tahun, serta terdapat beberapa nilai *outlier* pada asuransi

pemerintah dan swasta. Dari segi jenis kelamin, beberapa penelitian dibidang kesehatan menunjukkan bahwa perempuan memiliki sistem kekebalan tubuh lebih kuat dibandingkan dengan laki-laki. Hal ini selaras dengan kepemilikan asuransi di Indonesia dimana asuransi lebih banyak dimiliki oleh laki-laki dibandingkan perempuan. Asuransi pemerintah dimiliki 659 perempuan dan 3701 laki-laki atau dengan kata lain perbandingan perempuan dan laki-laki menggunakan asuransi pemerintah sebesar 1:5. Tidak hanya asuransi pemerintah, asuransi swasta juga didominasi laki-laki sebanyak 1381 orang. Untuk lebih jelasnya, visualisasi variabel jenis kelamin berdasarkan status kepemilikan asuransi ditampilkan pada Gambar 4.3 sebagai berikut.



Gambar 4.3 *Bar Chart* Jenis Kelamin terhadap Status Kepemilikan Asuransi

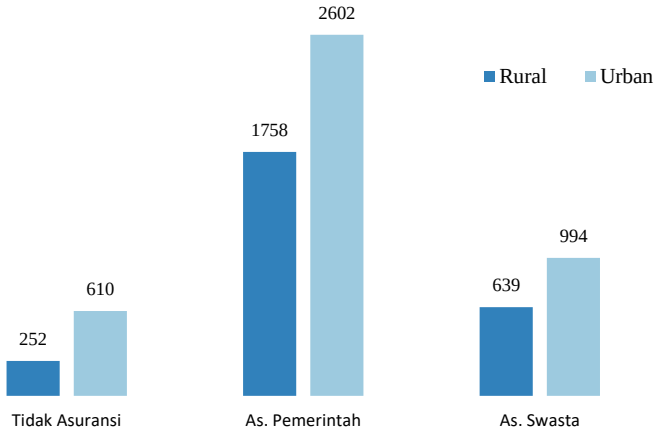
Selain umur dan jenis kelamin, terdapat faktor lain yang mempengaruhi status kepemilikan asuransi seperti pendidikan terakhir, pekerjaan, status, pendapatan, tempat tinggal, riwayat penyakit, frekuensi rawat jalan serta rawat inap. Tabel 4.1 menunjukkan *cross tabulation* antara status kepemilikan asuransi dengan pendidikan terakhir serta pekerjaan.

Tabel 4.1 *Cross Tabulation* Status Kepemilikan Asuransi dengan Pendidikan dan Pekerjaan

Faktor	Status Kepemilikan		
	Tidak	**As. Pemerintah**	**As. Swasta**
SD Sederajat	4,8%	23,8%	8,5%
SMP Sederajat	2,1%	10,9%	4,1%
SMA Sederajat	3,5%	19,3%	7,4%
Perguruan Tinggi	2,2%	9,6%	3,7%
Sekolah	0,6%	2,9%	1,1%
Bekerja	9,3%	48,9%	18,1%
Ibu RT	1%	4,5%	1,9%
Pensiun	0,7%	3,3%	1,1%
Menganggur	0,9%	4%	1,6%

Berdasarkan pendidikan, responden dengan pendidikan terakhir SD Sederajat paling banyak disetiap asuransi dengan persentase asuransi pemerintah dan swasta secara berturut-turut sebesar 23,8% dan 8,5%. Asuransi pemerintah merupakan salah satu asuransi dengan persentase terbanyak disetiap jenjang pendidikan mengingat harganya yang relatif lebih rendah jika dibandingkan dengan asuransi swasta. Responden yang tidak memiliki asuransi paling sedikit yaitu responden degan pendidikan terakhir perguruan tinggi sebesar 2,2%. Hal ini menunjukkan bahwa tingginya pendidikan berpengaruh terhadap kepedulian akan pentingnya kesehatan. Berdasarkan pekerjaannya, responden yang bekerja lebih banyak menggunakan asuransi pemerintah dengan persentase sebesar 48,9%. Responden yang masih duduk dibangku sekolah juga lebih banyak yang menggunakan asuransi dengan persentase asuransi pemerintah dan swasta sebesar 2,9% dan 1,1%, sedangkan responden yang tidak memiliki asuransi sebesar 0,6% atau sekitar 40 orang. Responden yang lain seperti ibu rumah tangga, pensiunan dan lain-lain juga lebih banyak yang memiliki asuransi dibandingkan dengan responden tidak memilikinya. Faktor lain yang berpengaruh dalam kepemilikan

asuransi yaitu tempat tinggal yang ditunjukkan pada hasil visualisasi Gambar 4.4 sebagai berikut.



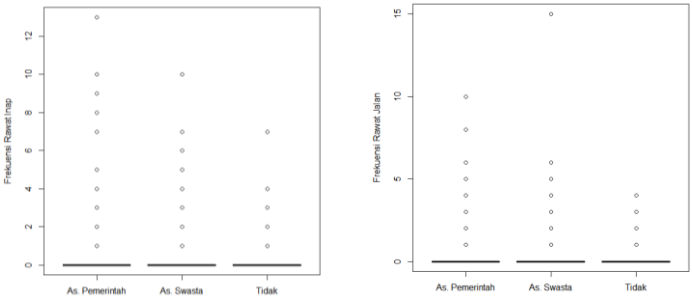
Gambar 4.4 *Bar Chart* Tempat Tinggal Berdasarkan Status Kepemilikan Asuransi

Gambar 4.4 menunjukkan bahwa responden terbanyak pengguna asuransi baik asuransi pemerintah maupun swasta yaitu masyarakat yang tinggal di daerah urban atau perkotaan. Pengguna asuransi pemerintah di daerah urban sebanyak 2602 orang lebih banyak dibandingkan dengan masyarakat rural yang hanya sebesar 1758 orang. Perbandingan pengguna asuransi swasta antara masyarakat rural dan urban yaitu 2:3 dimana masyarakat rural sebanyak 639 orang dan urban sebanyak 994 orang. Faktor lain yang mempengaruhi kepemilikan asuransi di Indonesia yaitu frekuensi kunjungan terhadap fasilitas kesehatan yang meliputi rawat inap dan rawat jalan. Karakteristik data dari faktor tersebut ditampilkan dalam Tabel 4.2 sebagai berikut.

Tabel 4.2 Statistika Deskriptif Frekuensi Kunjungan Rawat Inap dan Rawat Jalan

Faktor	Min	Max	Modus
Frekuensi Rawat Inap	0	13	0
Frekuensi Rawat Jalan	0	15	0

Responden pengguna asuransi kebanyakan tidak melakukan kunjungan rawat inap dan rawat jalan. Hal ini ditunjukkan dengan nilai modus dari masing-masing variabel yang bernilai 0. Selama 12 bulan terakhir, responden melakukan kunjungan rawat inap terbanyak yaitu 13 kali sedangkan untuk frekuensi rawat jalan terbanyak yaitu 15 kali. Gambar 4.5 menunjukkan frekuensi kunjungan ke fasilitas kesehatan disetiap asuransi.



Gambar 4.5 *Boxplot* Frekuensi Kunjungan Rawat Inap dan Rawat Jalan terhadap Status Kepemilikan Asuransi

Responden yang tidak memiliki asuransi, cenderung lebih sedikit melakukan rawat jalan dan rawat inap dibandingkan dengan responden yang memiliki asuransi. Frekuensi rawat inap terbanyak yaitu responden asuransi pemerintah sedangkan untuk rawat jalan paling banyak dilakukan oleh responden pada asuransi swasta. Secara berturut turut, frekuensi rawat inap terbanyak yang dilakukan responden asuransi pemerintah dan swasta yaitu 13 dan 10 kali. Sedangkan frekuensi rawat jalan terbanyak yaitu 10 kali untuk asuransi pemerintah dan 15 kali untuk asuransi swasta.

4.2 Analisis Regresi Logistik Multinomial pada Status Kepemilikan Asuransi

Penelitian ini menggunakan metode *multinomial logistic regression.* Untuk mengatasi *imbalance multiclass* pada variabel respon metode yang digunakan yaitu *Synthetic Minority Oversampling Technique* (SMOTE), *Random Undersampling* (RUS) dan *combine sampling*. Berikut hasil analisisnya.

4.2.1 Analisis Regresi Logistik Multinomial Status Kepemilikan Asuransi pada *Imbalanced Data*

Pada penelitian ini dilakukan dengan metode regresi logistik multinomial untuk mengetahui faktor-faktor apa saja yang berpengaruh terhadap status kepemilikan asuransi di Indonesia. Selain itu, juga dapat mengetahui ukuran ketepatan klasifikasinya.

A. Deteksi Multikolinieritas dan Independensi

Sebelum dilakukan analisis lebih lanjut, terlebih dahulu dilakukan pengujian asumsi untuk mengetahui ada tidaknya hubungan antar variabel. Pengujian asumsi ini meliputi uji multikolinieritas dan uji independensi.

Pengujian multikolinieritas dilakukan dengan melihat nilai *pearson correlation* antar variabel prediktor numerik. Hasil uji *pearson correlation* ditampilkan pada Tabel 4.3 sebagai berikut.

Tabel 4.3 *Pearson Correlation Imbalanced Data*

	X_1	X_9
X_9	0,002	
X_{10}	-0,028	-0,020

Berdasarkan Tabel 4.3 diatas menunjukkan bahwa nilai *pearson correlation* yang dihasilkan yaitu 0,002, -0,0028 dan -0,020. Hal ini menunjukkan bahwa tidak terdapat multikolinieritas karena nilai *pearson correlation* cukup kecil.

Pengujian asumsi selanjutnya yaitu uji independensi. Pengujian ini dilakukan dengan menggunakan *chi-square test* dengan hipotesis sebagai berikut.

Hipotesis

H_0 : Tidak ada hubungan antara variabel prediktor dengan variabel respon

H_1 : Ada hubungan antara variabel prediktor dengan variabel respon

Taraf signifikan : $\alpha = 0,05$

Hasil pengujian independensi pada *imbalanced data* ditampilkan pada Tabel 4.4 sebagai berikut.

Tabel 4.4 Hasil Uji Independensi *Imbalanced Data*

Variabel	*Chi-Square*	*P-value*
X_2	0,100	0,951
X_3	5,596	0,470
X_4	6,081	0,638
X_5	4,001	0,675
X_6	2,664	0,616
X_7	37,525	0,000
X_8	2,078	0,354

Adanya hubungan antara variabel respon dan prediktor dapat dilihat dengan menggunakan nilai *P-value*, jika nilai *P-value* kurang dari taraf signifikansi atau tingkat kepercayaan maka dikatakan ada hubungan antara variabel respon dan prediktor. Dalam penelitian ini menggunakan taraf signifikansi 5% atau 0,05, Berdasarkan Tabel 4.4 menunjukkan bahwa terjadi tolak H_0 pada variabel X_7 yang berarti terdapat hubungan yang signifikan antara tempat tinggal dan status kepemilikan asuransi di Indonesia.

Dari hasil pengujian asumsi berupa uji multikolinieritas dan uji independensi menunjukkan bahwa variabel yang memenuhi asumsi yaitu variabel X_1, X_7, X_9 dan X_{10}. Sehingga dapat dilakukan analisis regresi logistik multinomial pada keempat variabel tersebut yang meliputi umur, tempat tinggal, frekuensi rawat inap dan rawat jalan.

B. Pengujian Signifikansi Parameter

Pengujian signifikansi parameter ini dilakukan secara serentak dan parsial. Uji serentak bertujuan untuk melihat ada tidaknya pengaruh antara variabel prediktor terdapat status kepemilikan asuransi di Indonesia. Hipotesis yang digunakan yaitu sebagai berikut.

Hipotesis

H_0 : $\beta_1 = \beta_7 = \beta_9 = \beta_{10} = 0$.

H_1 : Minimal terdapat satu $\beta_i \neq 0$, dimana i=1,7,9,10

Taraf signifikan : $\alpha = 0,05$

Output dari pengujian serentak analisis regresi logistik multinomial pada *imbalanced data* disajikan pada Tabel 4.5.

Tabel 4.5 *Output* Uji Serentak pada *Imbalanced Data*

	Value	**df**	***P-value***
Chi-Square	46,228	8	0,000

Berdasarkan Tabel 4.5 dengan derajat bebas 8 didapatkan nilai *value* sebesar 46,228 dan *P-value* sebesar 0,000. Sehingga dapat disimpulkan dalam analisis regresi logistik multinomial pada *imbalanced data* ini minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap status kepemilikan asuransi, dan bisa dilanjutkan untuk pengujian parameter secara parsial.

Uji parsial dilakukan untuk melihat pengaruh dari masing-masing variabel prediktor terhadap status kepemilikan asuransi di Indoneesia. Berikut merupakan hipotesis untuk status kepemilikan asuransi berupa asuransi pemerintah.

Hipotesis

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$, dimana i=1,7,9,10

Taraf signifikan : $\alpha = 0,05$

Output pengujian parsial untuk status kepemilikan asuransi yang merupakan asuransi pemerintah disajikan pada Tabel 4.6

Tabel 4.6 *Output* Pengujian Parsial *Imbalanced Data* (Asuransi Pemerintah)

Variabel	***Estimate Parameter***	***Std. Error***	***Wald***	**df**	***P-value***
Intercept	1,511	0,117	167,574	1	0,000
X_1	-0,002	0,002	0,631	1	0,427

Tabel 4.6 *Output* Pengujian Parsial *Imbalanced Data* (Asuransi Pemerintah) (Lanjutan)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
$[X_7 = 0]$	0,497	0,081	37,346	1	0,000
X_9	0,021	0,064	0,107	1	0,744
X_{10}	0,092	0,066	1,913	1	0,167

Berdasarkan Tabel 4.6 dapat dilihat bahwa terdapat satu variabel yang berpengaruh signifikan terhadap asuransi pemerintah. Hal ini bisa dilihat dari nilai *P-value* yang kurang dari 5% hanya ada pada variabel X_7. Maka dapat disimpulkan bahwa tempat tinggal (X_7) berpengaruh signifikan terhadap status kepemilikan asuransi pemerintah. Selanjutnya yaitu pengujian parsial terhadap status kepemilikan asuransi yang merupakan asuransi swasta. Berikut hipotesis yang digunakan.

Hipotesis

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$, dimana i=1,7,9,10

Taraf signifikan : $\alpha = 0,05$

Output pengujian parsial untuk status skepemilikan asuransi yang merupakan asuransi swasta disajikan pada Tabel 4.7.

Tabel 4.7 *Output* Pengujian Parsial *Imbalanced Data* (Asuransi Swasta)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
Intercept	0,666	0,132	25,571	1	0,000
X_1	-0,005	0,003	3,142	1	0,076
$[X_7 = 0]$	0,454	0,091	25,065	1	0,000
X_9	0,073	0,070	1,089	1	0,297
X_{10}	0,072	0,074	0,965	1	0,326

Tabel 4.7 menunjukkan bahwa terdapat satu variabel yang berpengaruh signifikan terhadap status kepemilikan asuransi swasta yaitu X_7 atau tempat tinggal. Bisa dilihat dari nilai *P-value* kurang dari 5% untuk variabel X_7. Sehingga dapat diputuskan tolak

H_0. Sedangkan untuk variabel lain tidak berpengaruh signifikan karena *P-value* lebih dari 5%. Maka dapat disimpulkan bahwa terdapat satu variabel yang berpengaruh signifikan terhadap status kepemilikan asuransi swasta.

C. Pemodelan Regresi Logistik Multinomial

Setelah dilakukan pengujian signifikansi parameter secara serentak dan parsial, selanjutnya adalah pembentukan model regresi logistik multinomial. Hasil uji signifikansi parameter menunjukkan bahwa status kepemilikan asuransi pemerintah dan swasta hanya dipengaruhi oleh satu variabel yaitu tempat tinggal. Berikut merupakan nilai estimasi parameter *imbalanced data* untuk status kepemilikan asuransi pemerintah dan swasta.

Tabel 4.8 Estimasi Parameter pada *Imbalanced Data*

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Pemerintah	Intercept	1,511	
	[$X_7 = 0$]	0,497	1,643
As. Swasta	Intercept	0,666	
	[$X_7 = 0$]	0,454	1,575

Berdasarkan Tabel 4.8 didapatkan dua model logit, model logit 1 untuk asuransi pemerintah dan model logit 2 untuk asuransi swasta. Model logit yang didapatkan yaitu sebagai berikut.

$$g_1(x) = 1{,}511 + 0{,}497\, x_7(0) \quad (4.1)$$

$$g_2(x) = 0{,}666 + 0{,}454\, x_7(0) \quad (4.2)$$

Nilai Exp (B) pada Tabel 4.8 merupakan nilai *odds ratio.* Variabel X_7 pada logit pertama dan kedua menunjukkan nilai koefisien yang positif yang berarti perbandingan antara penduduk rural dan urban yang memilih menggunakan asuransi pemerintah sebesar 1,643 kali lebih besar dibandingkan dengan tidak menggunakan asuransi, dan yang memilih asuransi swasta sebesar 1,575 kali.

D. Uji Kesesuaian Model

Untuk melihat model yang dibentuk sudah baik atau belum maka dilihat dari nilai *goodness of fit*nya. Hipotesisnya adalah sebagai berikut.

Hipotesis

H_0 : Model fit (tidak ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

H_1 : Model belum fit (ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

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

Output goodness of fit model pada *imbalanced data* ditampilkan pada Tabel 4.9.

Tabel 4.9 *Goodness of Fit* Model *Imbalanced Data*

	Value	**Df**	***P-value***
Pearson Chi-Square	1628,796	1594	0,266

Berdasarkan Tabel 4.9 bisa dilihat bahwa model yang dibentuk sudah fit dengan nilai *P-value* sebesar 0,266. Nilai *P-value* ini lebih dari 5%, sehingga dapat disimpulkan gagal tolak H_0 yang berarti model yang dibentuk fit atau tidak memiliki perbedaan yang nyata antara observasi dengan hasil prediksi.

E. Ketepatan Klasifikasi

Ketepatan klasifikasi bertujuan untuk mengetahui apakah data yang digunakan telah diklasifikasikan dengan benar. Hasil ketepatan klasifikasi berupa *confusion matrix* ditampilkan pada Tabel 4.10 sebagai berikut.

Tabel 4.10 *Confusion Matrix* pada *Imbalanced Data*

Observed	***Predicted***		
	Tidak	**As.Pemerintah**	**As.Swasta**
Tidak	0	862	0
As.Pemerintah	0	4360	0
As. Swasta	0	1633	0

Berdasarkan Tabel 4.10 *confusion matrix* menunjukkan bahwa klasifikasi untuk tiap kategori cenderung mengarah pada satu kelas yaitu asuransi pemerintah. Hal tersebut terjadi karena kasus *imbalanced data* dimana kategori asuransi pemerintah memiliki persentase yang lebih tinggi dibandingkan dengan kategori yang lain. Untuk mengatasi hal tersebut, perlu adanya *preprocessing* mengenai *imbalanced data.*

Nilai *sensitivity* berdasarkan *confusion matrix multiclass* diatas untuk kelas tidak asuransi, asuransi pemerintah dan asuransi swasta yaitu 0, 1 dan 0. Sedangkan untuk *specificity* tiap kelasnya sebesar 1, 0 dan 1. Nilai AUC yang dihasilkan yaitu sebesar 54,03%. Karena data yang digunakan *imbalanced*, sehingga hasil klasifikasi mengarah pada kelas mayor. Maka dari itu dilakukan *balancing data* terlebih dahulu sebelum memulai analisis.

4.2.2 Analisis Regresi Logistik Multinomial Status Kepemilikan Asuransi dengan Metode Balancing SMOTE

Dalam analisis regresi logistik multinomial ini, dilakukan *preprocessing* data terlebih dahulu dengan mrnggunakan *Synthetic Minority Oversampling Technique* atau biasa dikenal dengan SMOTE. Selanjutnya dilakukan pengujian asumsi serta analisis regresi logistik multinomial yang meliputi pengujian serentak, pengujian parsial, *odd ratio*, uji kebaikan model dan ketepatan klasifikasi. Hasil analisis dapat dilihat sebagai berikut.

A. *Preprocessing* Data dengan Metode *Balancing* SMOTE

Jumlah data kepemilkan asuransi yaitu 6855 responden. Variabel respon pada asuransi dikategorikan berdasarkan kepemilikannya yang meliputi tidak memiliki asuransi, memiliki asuransi pemerintah dan asuransi swasta. Persentase jumlah anggota pada masing-masing variabel respon tidak seimbang yaitu 64% responden menggunakan asuransi pemerintah, 24%

responden menggunakan asuransi swasta dan sisanya sebanyak 12% responden memilih untuk tidak menggunakan asuransi. Berdasarkan persentasenya dapat diketahui bahwa kelas mayor atau kelas dengan data terbanyak yaitu asuransi pemerintah. Maka perlu dilakukan *preprocessing* data *imbalanced* untuk kelas minor pada kategori asuransi swasta dan tidak asuransi.

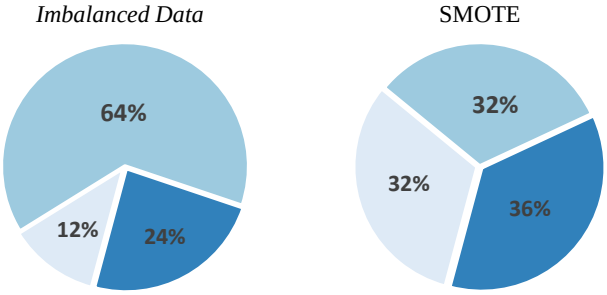
Untuk kelompok minor tidak asuransi dengan jumlah anggota 862 data maka perlu dilakukan replikasi sebanyak 4 kali untuk menyeimbangkan jumlah anggota dengan kelas mayor yaitu asuransi pemerintah. Sedangkan untuk kelompok minor asuransi swasta dengan jumlah anggota sebanyak 1633 data, replikasi yang perlu dilakukan adalah sebanyak 2 kali. Distribusi data setelah dilakukan replikasi dengan metode SMOTE ditunjukkan pada Tabel 4.11 sebagai berikut.

Tabel 4.11 Distribusi Data pada Masing-Masing Kategori dengan Metode SMOTE

Sebelum Replikasi		**Jumlah**	**Setelah Replikasi**	
Mayor	**Minor**	**Replikasi**	**Mayor**	**Minor**
As.Pemerin 4360 (64%)	Tidak As. 862 (12%)	4 kali	As.Pemerin 4360 (32%)	Tidak As. 4310 (32%)
	As.Swasta 1633 (24%)	2 kali		As.Swasta 4899 (36%)

Sehingga jumlah data yang semula 6855 data akan bertambah sampai dapat dikatakan jumlah anggota pada masing-masing kelas seimbang. Selain data pada variabel respon yang akan bertambah, jumlah data pada masing-masing variabel prediktor juga akan bertambah mengikuti jumlah data pada variabel respon. Dengan menyeimbangkan data pada masing-masing kategori pada variabel respon diharapkan tidak terjadi kasus *underfitting* atau *overfitting* sehingga menghasilkan nilai AUC yang lebih baik. Ilustrasi

komposisis data dengan metode SMOTE dapat dilihat pada Gambar 4.8 sebagai berikut.



Gambar 4.6 Perbandingan Komposisi Data Sebelum dan Sesudah *Resampling* dengan SMOTE

Berdasarkan Gambar 4.8 dapat diketahui bahwa persentase banyaknya anggota pada masing-masing kategori sudah seimbang. Semula jumlah data penelitian yang digunakan sebanyak 6855 data, saat ini bertambah menjadi 13569 data. Data tersebut berasal dari *synthetic* data yang dihasilkan berdasarkan replikasi menggunakan SMOTE. Persentase setiap kategori pada *imbalanced data* yang semula 12%, 24% dan 64% berubah menjadi 32%, 32% dan 36%. Setelah didapatkan jumlah data dengan proporsi kategori yang seimbang, dilanjutkan dengan tahapan analisis dengan metode regresi logistik multinomial.

B. Deteksi Multikolinieritas dan Independensi

Selanjutnya dilakukan pengujian asumsi pada data *balanced* hasil SMOTE yang terdiri dari 13569 data. Pengujian ini meliputi uji multikolinieritas dan uji independensi. Uji multikolinieritas merupakan uji asumsi untuk melihat apakah ada hubungan antar variabel dalam penelitian dengan menggunakan nilai *pearson correlation. Output pearson correlation* ditampilkan pada Tabel 4.12 sebagai berikut.

Tabel 4.12 *Pearson Correlation* Data Balanced SMOTE

	X_1	X_9
X_9	0,004	
X_{10}	-0,029	-0,021

Nilai *pearson Correlation* pada data *balanced* metode SMOTE yaitu 0,004, -0,029 dan -0,021. Nilai korelasi ketiga variabel tersebut cukup kecil sehingga dapat diambil kesimpulan bahwa tidak ada multikolinieritas antar variabel dan dilanjutkan pada uji independensi.

Uji Independensi merupakan uji asumsi yang digunakan untuk memeriksa ada tidaknya hubungan antara dua variabel yang diamati dengan menggunakan *chi-square test.* Hipotesis yang digunakan yaitu sebagai berikut.

Hipotesis

H_0 : Tidak ada hubungan antara variabel prediktor dengan variabel respon

H_1 : Ada hubungan antara variabel prediktor dengan variabel respon

Taraf signifikan : $\alpha = 0,05$

Output uji independensi ditampilkan pada Tabel 4.16 sebagai berikut.

Tabel 4.13 *Output* Uji Independensi Data *Balanced* SMOTE

Variabel	*Chi-Square*	*P-value*
X_2	130,445	0,000
X_3	111,495	0,000
X_4	49,151	0,000
X_5	84,311	0,000
X_6	76,257	0,000
X_7	138,127	0,000
X_8	6,562	0,038

Berdasarkan Tabel 4.13 menunjukkan bahwa nilai *P-value* yang dihasilkan cukup kecil dimana variabel X_2, X_3, X_4, X_5, X_6 dan X_7 memiliki nilai *P-value* yang sama yaitu 0,000, dan variabel X_8 dengan *P-value* sebesar 0,038. Nilai *P-value* yang dihasilkan kurang dari taraf signifikansi atau tingkat kepercayaan sebesar 5%. Jadi dapat diambil kesimpulan tolak H_0 yang berarti terdapat hubungan yang signifikan antara variabel.

Hasil pengujian asumsi multikolinieritas dan uji independensi menunjukkan bahwa data *balanced* SMOTE memenuhi kedua pengujian tersebut. Jadi dapat disimpulkan analisis regresi logistik multinomial dengan data *balanced* SMOTE dapat dilakukan dengan seluruh variabel.

C. Pengujian Signifikansi Parameter

Uji signifikansi parameter dalam analisis regresi logistik multinomial dilakukan dengan dua tahap yaitu pengujian secara serentak dan pengujian secara parsial. Pengujian serentak digunakan untuk melihat pengaruh variabel prediktor terdapat status kepemilikan asuransi. Hipotesis dalam pengujian serentak yaitu sebagai berikut.

Hipotesis

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_i = 0$.

H_1 : Minimal terdapat satu $\beta_i \neq 0$, dimana i=1,2,3,…,10

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

Hasil pengujian serentak ditampilkan pada Tabel 4.17 berikut.

Tabel 4.14 Output Uji Serentak pada Data Balanced SMOTE

	Value	**df**	***P-value***
Chi-Square	1334,386	36	0,000

Output dari uji serentak pada data *balanced* dengan derajat bebas 36 menghasilkan *value* sebesar 1334,386 dan *P-value* sebesar 0,000. Jadi dapat diputuskan tolak H_0 ang berarti minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap status kepemilikan asuransi, sehingga dapat dilanjutkan pengujian secara

parsial. Pengujian parsial dilakukan pada status kepemilikan asuransi pemerintah dan swasta. Hipotesis yang digunakan untuk status kepemilikan asuransi pemerintah yaitu sebagai berikut.

Hipotesis

H_0 : $\beta_i=0$

H_1 : $\beta_i \neq 0$, dimana i=1,2,3,…,10

Taraf signifikan : $\alpha = 0,05$

Hasil pengujian parsial untuk status kepemilikan asuransi pemerintah disajikan pada Tabel 4.15.

Tabel 4.15 *Output* Pengujian Parsial Data Balanced SMOTE (Asuransi Pemerintah)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
Intercept	0,342	0,193	3,129	1	0,077
X_1	0,026	0,002	136,372	1	0,000
$[X_2 = 0]$	-1,095	0,096	129,138	1	0,000
$[X_3 = 0]$	-0,482	0,075	41,745	1	0,000
$[X_3 = 1]$	0,067	0,081	0,676	1	0,411
$[X_3 = 2]$	0,192	0,071	7,260	1	0,007
$[X_4 = 0]$	-0,321	0,155	4,273	1	0,039
$[X_4 = 1]$	-0,535	0,105	25,788	1	0,000
$[X_4 = 2]$	-0,819	0,150	30,002	1	0,000
$[X_4 = 3]$	-0,445	0,151	8,745	1	0,003
$[X_5 = 0]$	-0,495	0,060	68,129	1	0,000
$[X_5 = 1]$	0,561	0,182	9,509	1	0,002
$[X_5 = 2]$	1,064	0,250	18,067	1	0,000
$[X_6 = 0]$	0,354	0,149	5,616	1	0,018
$[X_6 = 1]$	-0,185	0,101	3,355	1	0,067
$[X_7 = 0]$	0,681	0,050	183,962	1	0,000
$[X_8 = 0]$	0,604	0,218	7,672	1	0,006
X_9	0,720	0,055	169,391	1	0,000
X_{10}	0,635	0,054	136,372	1	0,000

Hasil pengujian secara parsial terhadap status kepemilikan asuransi pada Tabel 4.15 menunjukkan bahwa semua variabel berpengaruh signifika. Hal ini dapat dilihat dari nilai *P-value* yang kurang dari taraf signifikan. Selanjutnya pengujian parsial terhadap status kepemilikan asuransi swasta dengan hipotesis sebagai berikut.

Hipotesis

H_0 : $\beta_i=0$

H_1 : $\beta_i \neq 0$, dimana i=1,2,3,…,10

Taraf signifikan : $\alpha = 0,05$

Output pengujian parsial untuk status kepemilikan asuransi swasta disajikan pada Tabel 4.16.

Tabel 4.16 *Output* Pengujian Parsial Data *Balanced* SMOTE (Asuransi Swasta)

Variabel	***Estimate Parameter***	***Std. Error***	***Wald***	**df**	***P-value***
Intercept	0,874	0,191	20,893	1	0,000
X_1	-0,001	0,002	0,337	1	0,561
$[X_2 = 0]$	-0,976	0,094	108,664	1	0,000
$[X_3 = 0]$	0,181	0,075	5,905	1	0,015
$[X_3 = 1]$	0,659	0,080	67,611	1	0,000
$[X_3 = 2]$	s0,385	0,073	27,878	1	0,000
$[X_4 = 0]$	-0,137	0,152	0,814	1	0,367
$[X_4 = 1]$	-0,253	0,105	5,777	1	0,016
$[X_4 = 2]$	-0,930	0,148	39,233	1	0,000
$[X_4 = 3]$	-0,253	0,154	2,706	1	0,100
$[X_5 = 0]$	-0,006	0,055	0,012	1	0,912
$[X_5 = 1]$	0,881	0,173	25,894	1	0,000
$[X_5 = 2]$	1,005	0,250	16,225	1	0,000
$[X_6 = 0]$	0,029	0,145	0,041	1	0,840
$[X_6 = 1]$	-0,256	0,100	6,622	1	0,010
$[X_7 = 0]$	0,480	0,048	98,831	1	0,000
$[X_8 = 0]$	0,343	0,224	2,348	1	0,125
X_9	0,702	0,054	166,237	1	0,000
X_{10}	0,516	0,054	91,154	1	0,000

Berdasarkan Tabel 4.16 terdapat satu variabel yang tidak berpengaruh signifikan terhadap status kepemilikan asuransi swasta yaitu variabel X_1, karena memiliki nilai *P-value* lebih dari taraf signifikan. Jadi dapat simpulkan terdapat 9 variabel yang signifikan terhadap status kepemilikan asuransi swasta yaitu jenis kelamin, pendidikan terakhir, pendapatan, pekerjaan, status, tempat tinggal, riwayat penyakit, serta frekuensi rawat inap dan rawat jalan.

D. Pemodelan Regresi Logistik Multinomial

Tahapan selanjutnya yaitu pemodelan regresi logistik multinomial. Hasil uji signifikansi parameter didapatkan variabel yang berpengaruh signifikan terhadap status kepemilikan asuransi pemerintah sebanyak 10 variabel, sedangkan untuk asuransi swasta sebanyak 9 variabel. Estimasi parameternya ditampilkan pada Tabel 4.17 sebagai berikut.

Tabel 4.17 Estimasi Parameter Data *Balanced* SMOTE

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Pemerintah	Intercept	0,342	
	X_1	0,026	1,026
	$[X_2 = 0]$	-1,095	0,335
	$[X_3 = 0]$	-0,482	0,618
	$[X_3 = 1]$	0,067	1,069
	$[X_3 = 2]$	0,192	1,211
	$[X_4 = 0]$	-0,321	0,725
	$[X_4 = 1]$	-0,535	0,585
	$[X_4 = 2]$	-0,819	0,441
	$[X_4 = 3]$	-0,445	0,641
	$[X_5 = 0]$	-0,495	0,610
	$[X_5 = 1]$	0,561	1,752
	$[X_5 = 2]$	1,064	2,897

Tabel 4.17 Estimasi Parameter Data *Balanced* SMOTE (Lanjutan)

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Pemerintah	$[X_6 = 0]$	0,354	1,425
	$[X_6 = 1]$	-0,185	0,831
	$[X_7 = 0]$	0,681	1,976
	$[X_8 = 0]$	0,604	1,829
	X_9	0,720	2,054
	X_{10}	0,635	1,887
As. Swasta	Intercept	0,874	
	$[X_2 = 0]$	-0,976	0,377
	$[X_3 = 0]$	0,181	1,199
	$[X_3 = 1]$	0,659	1,933
	$[X_3 = 2]$	0,385	1,469
	$[X_4 = 0]$	-0,137	0,872
	$[X_4 = 1]$	-0,253	0,776
	$[X_4 = 2]$	-0,930	0,395
	$[X_4 = 3]$	-0,253	0,777
	$[X_5 = 0]$	-0,006	0,994
	$[X_5 = 1]$	0,881	2,414
	$[X_5 = 2]$	1,005	2,732
	$[X_6 = 0]$	0,029	1,030
	$[X_6 = 1]$	-0,256	0,774
	$[X_7 = 0]$	0,480	1,616
	$[X_8 = 0]$	0,343	1,410
	X_9	0,702	2,018
	X_{10}	0,516	1,676

Dari Tabel 4.17 bisa didapatkan dua model logit yaitu model logit 1 untuk asuransi pemerintah dan model logit 2 untuk asuransi swasta sebagai berikut.

$$
\begin{aligned}
g_1(x) = \; & 0{,}342 + 0{,}026x_1 - 1{,}095x_2(0) - 0{,}482x_3(0) \\
& + 0{,}067x_3(1) + 0{,}192x_3(2) - 0{,}321x_4(0) \\
& - 0{,}535x_4(1) - 0{,}819x_4(2) - 0{,}445x_4(3) \\
& - 0{,}495x_5(0) + 0{,}561x_5(1) + 1{,}064x_5(2) \\
& + 0{,}354x_6(0) - 0{,}185x_6(1) + 0{,}681x_7(0) \\
& + 0{,}604x_8(0) + 0{,}720x_9 + 0{,}630x_{10}
\end{aligned} \tag{4.3}
$$

$$
\begin{aligned}
g_2(x) = \; & 0{,}874 - 0{,}976x_2(0) + 0{,}181x_3(0) + 0{,}659x_3(1) \\
& + 0{,}385x_3(2) - 0{,}137x_4(0) - 0{,}253x_4(1) \\
& - 0{,}93x_4(2) - 0{,}253x_4(3) - 0{,}006x_5(0) + 0{,}881x_5(1) \\
& + 1{,}005x_5(2) + 0{,}029x_6(0) - 0{,}256x_6(1) \\
& + 0{,}48x_7(0) + 0{,}342x_8(0) + 0{,}702x_9 + 0{,}516x_{10}
\end{aligned} \tag{4.4}
$$

Dimana

X_1 : Umur	X_6 : Status
X_2 : Jenis kelamin	X_7 : Tempat tinggal
X_3 : Pendidikan terakhir	X_8 : Riwayat pendidikan
X_4 : Pekerjaan	X_9 : frekuensi rawat inap
X_5 : Pendapatan	X_{10} : frekuensi rawat jalan

Pada Tabel 4.17 terdapat nilai Exp (B) yaitu nilai *odds ratio* masing-masing variabel. Berikut ini analisis *odds ratio* pada model logit pertama dan kedua.

a. Umur

Variabel X_1 atau umur pada model logit pertama menunjukkan nilai koefisien yang positif maka penduduk yang usianya semakin tua memiliki kecenderungan 1,026 kali lebih besar untuk menggunakan asuransi pemerintah daripada tidak memiliki asuransi.

b. Jenis kelamin

Variabel jenis kelamin pada model logit pertama menunjukkan nilai negatif yang berarti perbandingan antara penduduk yang berjenis kelamin laki-laki dan perempuan

berkecenderungan memilih menggunakan asuransi pemerintah sebesar 0,335 kali lebih kecil dibandingkan dengan tidak menggunakan asuransi. Untuk model logit 2 juga menunjukkan nilai koefisien positif yang berarti perbandingan antara penduduk yang berjenis kelamin laki-laki dan perempuan berkecenderungan memilih asuransi swasta sebesar 0,377 kali lebih kecils daripada tidak menggunakan asuransi.

c. Pendidikan terakhir

Pada model logit pertama nilai koefisien pada variabel pendidikan terakhir$_1$ dan pendidikan terakhir$_2$ adalah positif yang berarti perbandingan penduduk dengan pendidikan terakhir SMP dan SMA sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan untuk menggunakan asuransi pemerintah masing-masing sebesar 1,069 dan 1,211 kali lebih besar daripada tidak menggunakan asuransi. Sedangkan pada variabel pendidikan terakhir$_0$ memiliki koefisien negatif yang berarti perbandingan antara penduduk SD sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan menggunakan asuransi pemerintah 0,618 kali lebih kecil daripada tidak menggunakannya.

Nilai koefisien pada model logit kedua untuk pendidikan terakhir$_0$, pendidikan terakhir$_1$ dan pendidikan terakhir$_2$ positif yang berarti perbandingan penduduk SD, SMP dan SMA sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan untuk menggunakan asuransi swasta masing-masing sebesar 1,199, 19,33 dan 1,469 kali lebih besar daripada tidak menggunakan asuransi.

d. Pekerjaan

Pada model logit pertama dan kedua variabel pekerjaan$_0$, pekerjaan$_1$, pekerjaan$_2$ dan pekerjaan$_3$ menunjukkan nilai koefisien negatif. Hal ini menunjukkan bahwa perbandingan antara penduduk yang tergolong sekolah, bekerja, ibu rumah tangga dan

pensiun dengan yang tergolong menganggur memiliki kecenderungan untuk menggunakan asuransi pemerintah masing-masing sebesar 0,725, 0,585, 0,441 dan 0,641 lebih kecil daripada tidak memiliki asuransi. Sedangkan kecenderungan untuk asuransi swasta masing-masing sebesar 0,872, 0,776, 0,395 dan 0,777 lebih kecil daripada tidak memiliki asuransi.

e. Pendapatan

Pendapatan pada model logit 1 dan 2 untuk variabel pendapatan$_0$ menunjukkan nilai koefisien negatif yang berarti perbandingan antara penduduk yang pendapatannya kurang dari Rp 1.000.000,00 dengan yang pendapatan yang tergolong lebih dari Rp3.000.000,00 memiliki kecenderungan untuk menggunakan asuransi pemerintah dan swasta secara berturut-turut sebesar 0,61 dan 0,994 kali lebih kecil dibandingkan dengan tidak memiliki asuransi. Selanjutnya untuk pendapatan$_1$ dan pendapatan$_2$ menunjukkan nilai koefisien positif. Hal ini menunjukkan bahwa penduduk yang memiliki pendapatan sebesar Rp 1.000.000,00 - Rp 2.000.000,00 dan Rp 2.000.000,00 - Rp 3.000.000,00 dan yang tergolong pendapatan lebih dari Rp 3.000.000,00 memiliki kecenderungan menggunakan asuransi pemerintah masing-masing sebesar 1,752 dan 2,897 lebih besar dibandingkan tidak asuransi. Sedangkan kecenderungan untuk asuransi swasta masing-masing sebesar 2,414 dan 2,732 lebih besar dibandingkan tidak asuransi.

f. Status

Model logit untuk status$_0$ memiliki nilai positif yang berarti perbandingan antara penduduk yang belum kawin dengan yang tergolong cerai hidup/mati memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,425 kali, dan asuransi swasta sebesar 1,03 kali lebih besar daripada tidak menggunakannya. Sedangkan untuk penduduk yang sudah kawin dengan yang tergolong cerai hidup/mati memiliki kecenderungan

untuk menggunakan asuransi pemerintah dan swasta sebesar 0,831 kali dan 0,774 kali lebih kecil dibandingkan tidak asuransi.

g. Tempat tinggal

Variabel tempat tinggal memiliki koefisien positif yang menunjukkan bahwa perbandingan penduduk rural dan urban memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,976 kali lebih besar dibandingkan dengan tidak menggunakan asuransi, dan yang memilih asuransi swasta 1,616 kali lebih besar dibandingkan dengan tidak menggunakannya.

h. Riwayat penyakit

Model logit pertama dan kedua menunjukkan nilai koefisien positif yang berarti perbandingan antara penduduk yang memiliki riwayat penyakit dan tidak memiliki berkecenderungan memilih asuransi pemerintah sebesar 1,829 lebih besar dibandingkan dengan tidak menggunakannya, dan untuk asuransi swasta sebesar 1,41 kali.

i. Frekuensi rawat inap

Variabel frekuensi rawat inap menunjukkan nilai koefisien positif berarti semakin banyak frekuensi rawat inap maka kecenderungan penduduk untuk menggunakan asuransi pemerintah 2,054 kali lebih besar daripada tidak menggunakannya. Hal yang sama juga untuk asuransi swasta tetapi dengan perbandingan sebesar 2,018 kali.

j. Frekuensi rawat jalan

Semakin banyak frekuensi rawat jalan, kecenderungan memiliki asuransi pemerintah 1,887 kali lebih besar dibandingkan tidak asuransi. Sedangkan kecenderungan memiliki asuransi swasta 1,676 kali.

E. Uji Kesesuaian Model

Tahapan selanjutnya yaitu uji kesesuaian model dengan melihat nilai *goodness of fit*. Pengujian ini bertujuan untuk

mengecek model yang telah terbentuuk sudah baik atau belum. Hipotesis yang digunakan yaitu sebagai berikut.

Hipotesis

H_0 : Model fit (tidak ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

H_1 : Model belum fit (ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

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

Hasil pengujian *goodness of fit* ditampilkan pada Tabel 4.18.

Tabel 4.18 *Goodness of Fit* Model Data Balanced SMOTE

	Value	**Df**	***P-value***
Pearson Chi-Square	15490,159	7898	0,000

Nilai *goodness of fit test* dengan derajat bebas 7898 menghasilkan *P-value* sebesar 0,000. Jadi dapat disimpulkan tolak H_0 yang berarti terdapat perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi.

F. Ketepatan Klasifikasi

Selanjutnya ketepatan klasifikasi, tahapan ini bertujuan untuk mengetahui apakah data yang digunakan telah diklasifikasikan dengan benar. Berikut hasil *confusion matrix* dari data balanced SMOTE.

Tabel 4.19 *Confusion Matrix* pada *Balanced* SMOTE

Observed	***Predicted***		
	Tidak	**As.Pemerintah**	**As.Swasta**
Tidak	2323	712	1275
As.Pemerintah	1274	1612	1474
As. Swasta	1494	1170	2235

Tabel 4.19 menunjukkan bahwa klasifikasi untuk tiap kategori menyebar pada ketiga kelas. Kategori tidak asuransi, asuransi pemerintah dan swasta yang terklasifikasi dengan benar sebanyak 2323, 1612 dan 2235 data. Berdasarkan *confusion matrix* diatas, didapatkan nilai *sensitivity* untuk kelas tidak asuransi,

asuransi pemerintah dan asuransi swasta secara berturut-turut sebesar 53,9%, 36,97% dan 45,62%. Sedangkan untuk *specificity* sebesar 70,1%, 79,56% dan 68,29%%. Untuk nilai AUC yang didapatkan yaitu sebesar 63,05%.

4.2.3 Analisis Regresi Logistik Multinomial Status Kepemilikan Asuransi dengan Metode Balancing R*andom Undersampling* (RUS)

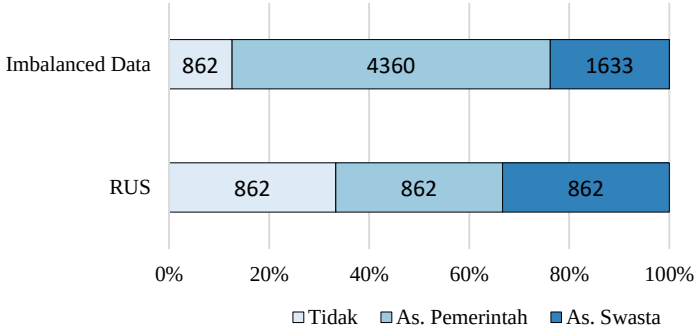
Analisis regresi logistik multinomial dilakukan melalui beberapa tahap meliputi pengujian serentak, pengujian parsial, *odd ratio*, uji kebaikan model dan ketepatan klasifikasi. Pada kasus status kepemilikan asuransi ini terjadi *imbalanced data* sehingga perlu dilakukan *preprocessing* data terlebih dahulu dengan metode *random undersampling*. Selanjutnya dilakukan pengujian asumsi serta analisis regresi. Hasil analisis akan disajikan sebagai berikut.

A. *Preprocessing* Data dengan Metode *Balancing Random Undersampling*

Kasus status kepemilikan asuransi pada *Indonesia Family Life Survey* (IFLS) merupakan salah satu contoh *imbalanced data* karena variabel respon pada penelitian ini memiliki komposisi jumlah yang berbeda. Data yang *imblanced* akan berdampak pada nilai AUC yang cenderung lebih buruk dibandingkan dengan data *balanced*. Oleh karena itu dilakukan *resampling* data dengan metode *Random Undersampling* atau RUS.

Variabel respon pada status kepemilikan asuransi terbagi dalam 3 kategori yaitu kategori tidak asuransi, memiliki asuransi pemerintah dan asuransi swasta. Jumlah data pada masing-masing kategori yaitu 862 data tidak asuransi, 4360 data asuransi pemerintah dan 1613 data asuransi swasta. Perbandingan jumlah data disetiap kategori cukup banyak, sehingga perlu dilakukan *resampling*. Metode *Random Undersampling* merupakan metode penghapusan secara *random* pada data mayoritas sehingga data menjadi *balanced* dan setiap kategori pada status kepemilikan

asuransi memiliki jumlah yang sama yaitu 862 data. Perbandingan komposisi data sebelum dan sesudah dilakukan *resampling* ditampilkan pada Gambar 4.9 sebagai berikut.



Gambar 4.7 Perbandingan Komposisi Data Sebelum dan Sesudah Dilakukan *Resampling* dengan *Random Undersampling*

B. Deteksi Multikolinieritas dan Independensi

Pengujian asumsi bertujuan untuk mengetahui ada tidaknya hubungan antar variabel. Pengujian ini dilakukan pada data *balanced* hasil *resampling* metode *random undersampling* sebanyak 2586 data. Terdapat dua pengujian asumsi yaitu uji multikolinieritas dan uji independensi.

Multikolinieritas dapat mendeteksi adanya hubungan antara variabel prediktor dengan melihat nilai *pearson correlation. Output* dari uji multikolinieritas pada data *balanced* metode *undersampling* disajikan pada Tabel 4.20 sebagai berikut.

Tabel 4.20 *Pearson Correlation* Data *Balanced* RUS

	X_1	X_9
X_9	-0,005	
X_{10}	-0,014	-0,023

Tabel 4.20 diatas menunjukkan nilai *pearson correlation* yang cukup kecil yaitu -0,005, -0,014 dan -0,023. Nilai korelasi yang kecil mengindikasikan tidak adanya hubungan antar variabel yang berarti tidak terdapat multikolinieritas antar variabel.

Selanjutnya uji independensi yang dapat dilakukan dengan *chi-square test.* Hipotesis yang digunakan yaitu sebagai berikut.

H_0 : Tidak ada hubungan antara variabel prediktor dengan variabel respon

H_1 : Ada hubungan antara variabel prediktor dengan variabel respon

Output uji independensi ditampilkan pada Tabel 4.21 sebagai berikut.

Tabel 4.21 Output Uji Independensi Data *Balanced* RUS

Variabel	*Chi-Square*	*P-value*
X_2	2,076	0,354
X_3	4,320	0,633
X_4	5,467	0,707
X_5	5,246	0,513
X_6	2,356	0,671
X_7	29,906	0,000
X_8	1,667	0,435

Hasil *P-value* uji independensi data *balanced* RUS pada Tabel 4.21 untuk variabel X_2 hingga X_8 secara berturut-turut yaitu 0,354, 0,633, 0,707, 0,513, 0,671, 0,000 dan 0,435. Dalam penelitian ini menggunakan taraf signifikansi atau tingkat kepercayaan sebesar 5%. Variabel yang memiliki nilai *P-value* kurang dari 5% yaitu variabel X_7. Sehingga dapat diambil kesimpulan bahwa terdapat hubungan yang signifikan antara status kepemilikan asuransi dengan variabel X_7 (tempat tinggal).

Pengujian asumsi multikolinieritas dan independensi menunjukkan bahwa terdapat empat variabel yang memenuhi asumsi yaitu variabel X_1, X_7, X_9 dan X_{10}. Jadi analisis regresi

logistik multinomial dilakukan dengan menggunakan empat variabel tersebut yaitu umur, tempat tinggal, frekuensi rawat inap dan rawat jalan.

C. Pengujian Signifikansi Parameter

Pengujian signifikansi parameter ini diawali dengan pengujian secara serentak yang bertujuan untuk melihat ada tidaknya pengaruh antara variabel prediktor dengan variabel respon. Hipotesis yang digunakan yaitu sebagai berikut.

Hipotesis

H_0 : $\beta_1 = \beta_7 = \beta_9 = \beta_{10} = 0$.

H_1 : Minimal terdapat satu $\beta_i \neq 0$, dimana i=1,7,9,10

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

Output pengujian serentak analisis regresi logistik multinomial pada data *balanced* dengan *preprocessing undersampling* disajikan pada Tabel 4.22.

Tabel 4.22 Output Uji Serentak pada Data *Balanced* RUS

	Value	**df**	***P-value***
Chi-Square	39,674	8	0,000

Hasil pengujian serentak didapatkan nilai value sebesar 39,674 dan *P-value* sebesar 0,000. Sehingga dapat disimpulkan tolak H_0 yang berarti minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap status kepemilikan asuransi. Sealnjutnya pengujian secara parsial terhadap status kepemilikan asuransi pemerintah dan swasta dengan hipotesis sebagai berikut.

Hipotesis

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$, dimana i=1,7,9,10

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

Output uji parsial untuk status kepemilikan asuransi pemerintah disajikan pada Tabel 4.20.

Tabel 4.23 Output Pengujian Parsial Data Balanced RUS (Asuransi Pemerintah)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
Intercept	-0,112	0,152	0,539	1	0,463
X_1	-0,002	0,003	0,369	1	0,544
$[X_7 = 0]$	0,477	0,103	21,631	1	0,000
X_9	0,000	0,086	0,000	1	0,998
X_{10}	0,160	0,080	3,963	1	0,047

Hasil pengujian parameter secara parsial didapatkan dua variabel memiliki nilai kurang dari taraf signifikansi 5% yaitu variabel X_7 dengan *P-value* sebesar 0,000 dan variabel X_{10} dengan *P-value* sebesar 0,047. Jadi dapat ditarik kesimpulan ansuransi pemerintah dengan *preprocessing undersampling* dipengaruhi oleh tempat tinggal dan frekuensi rawat jalan. Pengujian parsial selanjutnya yaitu status kepemilikan asuransi swasta dengan hipotesis sebagai berikut.

Hipotesis

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$, dimana i=1,7,9,10

Taraf signifikan : $\alpha = 0,05$

Hasil pengujian parsial untuk status kepemilikan asuransi swasta dengan *preprocessing undersampling* disajikan pada Tabel 4.24.

Tabel 4.24 Output Pengujian Parsial Data *Balanced* RUS (Asuransi Swasta)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
Intercept	-0,070	0,152	0,209	1	0,647
X_1	-0,004	0,003	1,461	1	0,227
$[X_7 = 0]$	0,515	0,103	25,184	1	0,000
X_9	0,110	0,080	1,855	1	0,173
X_{10}	0,155	0,081	3,694	1	0,055

Tabel 4.24 menunjukkan hasil uji signifikansi parameter secara parsial untuk variabel X_1, X_7, X_9 dan X_{10} secara berturut-turut sebesar 0,227, 0,000, 0,173 dan 0,055. Dengan taraf

signifikan 5% terdapat satu nilai *P-value* kurang dari 5% yaitu variabel X_7. Sehingga dapat ditarik kesimpulan bahwa asuransi swasta dengan *preprocessing undersampling* hanya dipengaruhi oleh tempat tinggal.

D. Pemodelan Regresi Logistik Multinomial

Pemodelan regresi logistik multinomial dilakukan terhadap variabel yang berpengaruh signifikan berdasarkan hasil uji signifikansi parameter, dimana status kepemilikan asuransi pemerintah dipengaruhi oleh tempat tinggal, sedangkan asuransi swasta hanya dipengaruhi oleh tempat tinggal. Estimasi parameter data *balanced undersampling* untuk asuransi pemerintah dan swasta ditampilkan pada Tabel 4.25 sebagai berikut.

Tabel 4.25 Estimasi Parameter Balanced Data RUS

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Pemerintah	Intercept	-0,112	
	[$X_7 = 0$]	0,477	1,612
	X_{10}	0,160	1,174
As. Swasta	Intercept	-0,070	
	[$X_7 = 0$]	0,515	1,673

Berdasarkan estimasi parameter Tabel 4.25 didapatkan dua model logit, model logit pertama untuk asuransi pemerintah dan model logit kedua untuk asuransi swasta. Model logit yang didapatkan yaitu sebagai berikut.

$$g_1(x) = -0{,}112 + 0{,}477x_7(0) + 0{,}16x_{10} \quad (4.5)$$

$$g_2(x) = -0{,}07 + 0{,}515x_7(0) \quad (4.6)$$

Dimana variabel X_7 merupakan tempat tinggal dan variabel X_{10} frekuensi rawat jalan. Nilai Exp (B) pada Tabel 4.25 merupakan nilai *odds ratio.* Pada data status kepemilikan asuransi pemerintah dengan *preprocessing undersampling* menunjukkan bahwa terdapat 2 variabel yang berpengaruh signifikan yaitu tempat tinggal dan frekuensi rawat jalan. Variabel tempat tinggal memiliki

nilai koefisien positif yang berarti penduduk rural dan urban yang memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,612 kali lebih besar dibandingkan dengan tidak menggunakan asuransi. Variabel tempat tinggal pada model logit pertama juga memiliki nilai koefisien positif yang berarti semakin banyak frekuensi rawat jalan, kecenderungan memiliki asuransi pemerintah 0,16 kali lebih besar dibandingkan tidak memilikinya.

Model logit 2 untuk variabel tempat tinggal memiliki nilai koefisien positif yang berarti penduduk rural dan urban yang memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,673 kali lebih besar dibandingkan dengan tidak menggunakan asuransi.

E. Uji Kesesuaian Model

Setelah didapatkan model logit, selanjutnya dilakukan pengujian apakah model yang dibentuk sesuai dengan nilai *goodness of fit*nya. Hipotesisnya adalah sebagai berikut.

Hipotesis

H_0 : Model fit (tidak ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

H_1 : Model belum fit (ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

Taraf signifikan : $\alpha = 0,05$

Output goodness of fit pada data *balanced undersampling* ditampilkan pada Tabel 4.26.

Tabel 4.26 *Goodness of Fit Model* Data *Balanced* RUS

	Value	**df**	***P-value***
Pearson Chi-Square	1079,289	1056	0,302

Tabel 4.26 menunjukkan hasil *goodness of fit test* dengan nilai *P-value* sebesar 0,302. Nilai *P-value* yang didapatkan lebih besar dari taraf signifikansi 5%. Jadi dapat ditarik kesimpulan bahwa model regresi logistik multinomial dengan *preprocessing*

undersampling tidak memiliki perbedaan yang nyata antara observasi dengan hasil prediksi.

F. Ketepatan Klasifikasi pada Data *Balanced RUS*

Ketepatan klasifikasi bertujuan untuk mengetahui apakah data yang digunakan telah terklasifikasikan dengan benar. Hasil *confusion matrix* dari data *balanced undersampling* ditampilkan pada Tabel 4.27 berikut.

Tabel 4.27 *Confusion Matrix* pada *Balanced* RUS

Observed	***Predicted***		
	Tidak	**As.Pemerintah**	**As.Swasta**
Tidak	571	143	148
As.Pemerintah	476	179	207
As. Swasta	467	162	233

Output confusion matrix pada Tabel 4.27 menunjukkan hasil klasifikasi yang menyebar pada ketiga kelas dengan hasil klasifikasi benar sebesar 571 data untuk tidak asuransi, 179 data untuk asuransi pemerintah, dan 233 data untuk asuransi swasta. Nilai *sensitivity* untuk setiap kelasnya berturut-turut sebesar 66,24%, 20,77% dan 27,03%. Sedangkan nilai AUC yang dihasilkan yaitu sebesar 54,78%.

4.2.4 Analisis Regresi Logistik Multinomial Status Kepemilikan Asuransi dengan Metode *Balancing Combine Sampling*

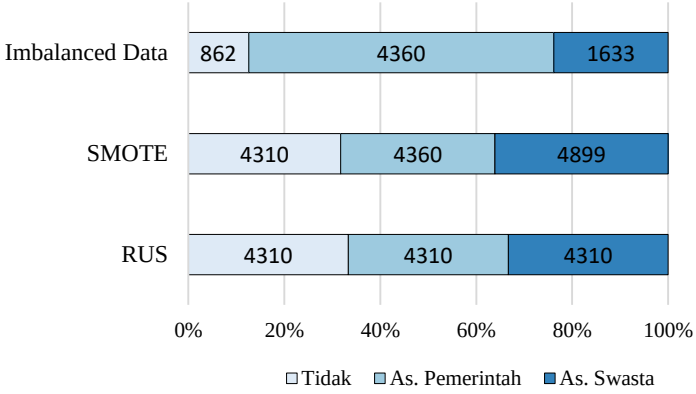
Analisis regresi logistik multinomial kali ini diawali dengan *preprocessing data* menggunakan *combine sampling* berupa SMOTE dan *undersampling*. Kemudian dilanjutkan dengan analisis regresi yang meliputi uji signifikansi parameter, pemodelan, uji kebaikan model serta ukuran ketepatan klasifikasi.

A. *Prepocessing* Data dengan Metode *Balancing Combine Sampling*

Status kepemilikan asuransi di Indonesia merupakan contoh salah satu data *imbalanced* dimana penggunaan asuransi

pemerintah mendominasi karena harganya yang terjangkau. Untuk mengatasi data *imbalanced* maka perlu dilakukan *resampling* dalam hal ini menggunakan metode *combine sampling.*

Metode *combine sampling* merupakan perpaduan metod *oversampling* dan *undersampling*, dimana metode *oversampling* yang digunakan yaitu *Synthetic Minority Oversampling Technique* (SMOTE) dan untuk metode *undersampling* yaitu *Random Undersampling* (RUS). Penggunaan kedua motode dilakukan secara berurutan. Langkah awal data *imbalanced* dilakukan resampling dengan menggunakan SMOTE sehingga data kelas minoritas akan seimbang dengan kelas mayoritas. Setelah itu, data hasil *resampling* dengan SMOTE dilanjutkan *resampling* kembali dengan menggunakan *Random Undersampling*. Proses *resampling* yang kedua dilakukan untuk mengapus secara random data pada kelas mayoritas sehingga data balanced dan status kepemilikan asuransi disetiap kategori memiliki jumlah yang sama yaitu 4310 data. Berikut ini perbandingan komposisi data sebelum dan sesudah dilakukan *resampling* dengan *combine sampling* yang ditampilkan pada Gambar 4.10.



Gambar 4.8 Perbandingan Komposisi Data Sebelum dan Sesudah dilakukan *Resampling* dengan *Combine Sampling*

B. Deteksi Mltikolinieritas dan Independensi

Selanjutnya dilakukan pengujian asumsi pada data hasil *resampling* dengan metode *combine sampling*. Uji asumsi ini bertujuan untuk mengetahui ada tidaknya hubungan antar variabel yang terdiri uji multikolinieritas dan uji independensi. Uji multikolinieritas merupakan uji asumsi untuk mengetahui adanya hubungan antara variabel prediktor dengan nilai *pearson correlation. Output pearson correlation* ditampilkan pada Tabel 4.28 sebagai berikut.

Tabel 4.28 *Pearson Correlation* Data *Balanced Combine Sampling*

	X_1	X_9
X_9	0,004	
X_{10}	-0,030	-0,021

Tabel 4.28 menunjukkan nilai *pearson correlation* pada data *balanced* metode *combine sampling.* Nilai *pearson correlation* sebesar 0,004, -0,030 dan -0,021. Nilai tersebut cukup kecil sehingga dapat ditarik kesimpulan bahwa tidak terdapat multikolinieritas antar variabel prediktor.

Setelah uji multikolinieritas, selanjutnya dilakukan uji independensi dengan melihat nilai *chi-square test.* Hipotesis yang digunakan yaitu sebagai berikut.

H_0 : Tidak ada hubungan antara variabel prediktor dengan variabel respon

H_1 : Ada hubungan antara variabel prediktor dengan variabel respon

Output uji independensi ditampilkan pada Tabel 4.29 sebagai berikut.

Tabel 4.29 *Output Uji Independensi* Data *Balanced Combine Sampling*

Variabel	*Chi-Square*	*P-value*
X_2	128,679	0,000
X_3	107,807	0,000

Tabel 4.29 *Output Uji Independensi Data* Balanced *Combine Sampling* (Lanjutan)

Variabel	*Chi-Square*	*P-value*
X_4	49,976	0,000
X_5	80,761	0,000
X_6	79,405	0,000
X_7	132,534	0,000
X_8	6,718	0,035

Nilai *P-value* yang dihasilkan pada uji independensi untuk variabel X_2 hingga X_8 kurang dari taraf signifikansi atau tingkat kepercayaan sebesar 5%. Sehingga dapat ditarik kesimpulan tolak H_0 yang berarti terdapat hubungan yang signifikan antar variabel.

Berdasarkan uji multikolinieritas dan uji independensi menunjukkan bahwa data *balanced combine sampling* memenuhi kedua asumsi. Sehingga dapat dilanjutkan analisis regresi logistik multinomial dengan seluruh variabel yaitu umur, jenis kelamin, pendidikan, pekerjaan, pendapatan, status, tempat tinggal, riwayat penyakit, serta frekuensi rawat inap dan rawat jalan.

C. Pengujian Signifikansi Parameter

Setelah dilakukan uji asumsi, selanjutnya dilakukan uji signifikansi parameter terhadap variabel yang memenuhi kedua asumsi tersebut. Pengujian signifikansi ini dilakukan secara serentak dan parsial. Pengujian secara serentak dilakukan terlebih dahulu untuk melihat ada tidaknya pengaruh variabel prediktor terhadap variabel respon. Hipotesisnya yaitu sebagai berikut.

Hipotesis

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_i = 0$.

H_1 : Minimal terdapat satu $\beta_i \neq 0$, dimana i=1,2,3,…,10

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

Hasil pengujian serentak ditampilkan pada Tabel 4.30 berikut.

Tabel 4.30 *Output* Uji Serentak pada Data *Balanced Combine Sampling*

	Value	**df**	***P-value***
Chi-Square	1308,883	36	0,000

Berdasarkan Tabel 4.30 nilai uji serentak menghasilkan nilai *P-value* sebesar 0,000. Jadi hasil uji serentak menunjukkan bahwa terdapat minimal satu variabel prediktor yang berpengaruh signifikan terhadap status kepemilikan asuransi. Selanjutnya dilakukan pengujian parsial terhadap status kepemilikan asuransi pemerintah dan swasta dengan hipotesis sebagai berikut.

Hipotesis

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$, dimana i=1,2,3,…,10

Taraf signifikan : $\alpha = 0,05$

Hasil pengujian parsial untuk status kepemilikan asuransi pemerintah disajikan pada Tabel 4.31.

Tabel 4.31 *Output* Pengujian Parsial Data *Balanced Combine Sampling* (Asuransi Pemerintah)

Variabel	***Estimate Parameter***	***Std. Error***	***Wald***	**df**	***P-value***
Intercept	0,303	0,194	2,428	1	0,119
X_1	0,026	0,002	154,087	1	0,000
$[X_2 = 0]$	-1,089	0,097	126,415	1	0,000
$[X_3 = 0]$	-0,478	0,075	40,699	1	0,000
$[X_3 = 1]$	0,068	0,082	0,690	1	0,406
$[X_3 = 2]$	0,198	0,072	7,658	1	0,006
$[X_4 = 0]$	-0,313	0,156	4,038	1	0,044
$[X_4 = 1]$	-0,533	0,106	25,411	1	0,000
$[X_4 = 2]$	-0,814	0,150	29,375	1	0,000
$[X_4 = 3]$	-0,459	0,151	9,233	1	0,002

Tabel 4.31 Output *Pengujian Parsial Data Balanced Combine Sampling (Asuransi Pemerintah)* (Lanjutan)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
$[X_5 = 0]$	-0,492	0,060	66,594	1	0,000
$[X_5 = 1]$	0,536	0,183	8,582	1	0,003
$[X_5 = 2]$	1,025	0,252	26,588	1	0,000
$[X_6 = 0]$	0,348	0,150	5,375	1	0,020
$[X_6 = 1]$	-0,178	0,102	3,084	1	0,079
$[X_7 = 0]$	0,678	0,050	180,440	1	0,000
$[X_8 = 0]$	0,613	0,218	7,884	1	0,005
X_9	0,718	0,056	167,243	1	0,000
X_{10}	0,640	0,055	137,100	1	0,000

Tabel 4.31 menunjukkan nilai uji signifikansi secara parsial terhadap status kepemilikan asuransi pemerintah dimana nilai *P-value* dihasilkan kurang dari taraf signifikansi 0,05. Jadi dapat disimpulkan seluruh variabel prediktor berpengaruh signifikan terhadap status kepemilikan asuransi pemerintah. Pengujian parsial selanjutnya dilakukan terhadap status kepemilikan asuransi swasta dengan hipotesis sebagai berikut.

Hipotesis

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$, dimana i=1,2,3,…,10

Taraf signifikan : $\alpha = 0,05$

Output pengujian parsial untuk status kepemilikan asuransi swasta disajikan pada Tabel 4.32.

Tabel 4.32 *Output* Pengujian Parsial Data *Balanced Combine Sampling* (Asuransi Swasta)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
Intercept	0,683	0,197	11,999	1	0,001
X_1	-0,001	0,002	0,114	1	0,736
$[X_2 = 0]$	-0,998	0,096	108,561	1	0,000
$[X_3 = 0]$	0,186	0,077	5,819	1	0,016

Tabel 32. *Output* Pengujian Parsial Data *Balaced Combine Sampling* (Asuransi Swasta) (Lanjutan)

Variabel	*Estimate Parameter*	*Std. Error*	*Wald*	df	*P-value*
[X_3 = 1]	0,672	0,083	66,104	1	0,000
[X_3 = 2]	0,379	0,076	25,130	1	0,000
[X_4 = 0]	-0,098	0,156	0,399	1	0,528
[X_4 = 1]	-0,210	0,109	3,723	1	0,054
[X_4 = 2]	-0,952	0,154	38,463	1	0,000
[X_4 = 3]	-0,219	0,158	1,920	1	0,166
[X_5 = 0]	0,018	0,057	0,104	1	0,747
[X_5 = 1]	0,846	0,178	22,667	1	0,000
[X_5 = 2]	0,998	0,254	15,454	1	0,000
[X_6 = 0]	0,060	0,149	0,164	1	0,685
[X_6 = 1]	-0,239	0,102	5,459	1	0,019
[X_7 = 0]	0,471	0,050	89,750	1	0,000
[X_8 = 0]	0,357	0,229	2,424	1	0,119
X_9	0,706	0,055	162,542	1	0,000
X_{10}	0,520	0,055	89,401	1	0,000

Tabel 4.32 menunjukkan nilai uji signifikansi parameter secara parsial terhadap asuransi swasta. Hasil pengujian menunjukkan bahwa terdapat satu variabel prediktor yang tidak berpengaruh yaitu variabel X_1 karena memiliki nilai *P-value* lebih dari taraf signifikan 5%. Jadi terdapat 9 variabel yang berpengaruh signifikan terhadap asuransi swasta yang meliputi jenis kelamin, pendidikan terakhir, pendapatan, pekerjaan, status, tempat tinggal, riwayat penyakit, serta frekuensi rawat inap dan rawat jalan.

D. Pemodelan Regresi Logistik Multinomial

Hasil uji signifikansi parameter terhadap status kepemilikan asuransi pemerintah yaitu terdapat 10 variabel yang berpengaruh signifikan, sedangkan untuk asuransi swasta terdapat 9 variabel. Estimasi parameter status kepemilikan asuransi ditampilkan pada Tabel 4.33.

Tabel 4.33 Estimasi Parameter Data *Balanced Combine Sampling*

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Pemerintah	Intercept	0,303	
	X_1	0,026	1,026
	$[X_2 = 0]$	-1,089	0,337
	$[X_3 = 0]$	-0,478	0,620
	$[X_3 = 1]$	0,068	1,070
	$[X_3 = 2]$	0,198	1,219
	$[X_4 = 0]$	-0,313	0,731
	$[X_4 = 1]$	-0,533	0,587
	$[X_4 = 2]$	-0,814	0,443
	$[X_4 = 3]$	-0,459	0,632
	$[X_5 = 0]$	-0,492	0,611
	$[X_5 = 1]$	0,536	1,710
	$[X_5 = 2]$	1,025	2,787
	$[X_6 = 0]$	0,348	1,416
	$[X_6 = 1]$	-0,178	0,837
	$[X_7 = 0]$	0,678	1,970
	$[X_8 = 0]$	0,613	1,845
	X_9	0,718	2,051
	X_{10}	0,640	1,897
As. Swasta	Intercept	0,683	
	$[X_2 = 0]$	-0,998	0,369
	$[X_3 = 0]$	0,186	1,205
	$[X_3 = 1]$	0,672	1,958
	$[X_3 = 2]$	0,379	1,460
	$[X_4 = 0]$	-0,098	0,906
	$[X_4 = 1]$	-0,210	0,811

Tabel 4.33 Estimasi Parameter Data *Balanced Combine Sampling* (Lanjutan)

Variabel Respon	Variabel Prediktor	B	Exp(B)
As. Swasta	$[X_4 = 2]$	-0,952	0,386
	$[X_4 = 3]$	-0,219	0,803
	$[X_5 = 0]$	0,018	1,019
	$[X_5 = 1]$	0,846	2,330
	$[X_5 = 2]$	0,998	2,713
	$[X_6 = 0]$	0,060	1,062
	$[X_6 = 1]$	-0,239	0,787
	$[X_7 = 0]$	0,471	1,602
	$[X_8 = 0]$	0,357	1,429
	X_9	0,706	2,025
	X_{10}	0,520	1,683

Berdasarkan Tabel 4.33 didapatkan dua model logit yaitu model logit pertama untuk asuransi pemerintah dan model logit kedua untuk asuransi swasta sebagai berikut.

$$\begin{aligned} g_1(x) = {} & 0{,}303 + 0{,}026x_1 - 1{,}089x_2(0) - 0{,}478x_3(0) \\ & + 0{,}068x_3(1) + 0{,}198x_3(2) - 0{,}313x_4(0) \\ & - 0{,}533x_4(1) - 0{,}814x_4(2) - 0{,}459x_4(3) \\ & - 0{,}492x_5(0) + 0{,}536x_5(1) + 1{,}025x_5(2) \\ & + 0{,}348x_6(0) - 0{,}178x_6(1) + 0{,}678x_7(0) \\ & + 0{,}613x_8(0) + 0{,}718x_9 + 0{,}64x_{10} \end{aligned} \tag{4.7}$$

$$\begin{aligned} g_2(x) = {} & 0{,}683 - 0{,}998x_2(0) + 0{,}186x_3(0) + 0{,}672x_3(1) \\ & + 0{,}379x_3(2) - 0{,}098x_4(0) - 0{,}21x_4(1) \\ & - 0{,}952x_4(2) - 0{,}219x_4(3) + 0{,}018x_5(0) \\ & + 0{,}846x_5(1) + 0{,}998x_5(2) + 0{,}06x_6(0) \\ & - 0{,}239x_6(1) + 0{,}471x_7(0) + 0{,}357x_8(0) \\ & + 0{,}706x_9 + 0{,}52x_{10} \end{aligned} \tag{4.8}$$

Dimana

X_1 : Umur	X_6 : Status
X_2 : Jenis kelamin	X_7 : Tempat tinggal
X_3 : Pendidikan terakhir	X_8 : Riwayat pendidikan
X_4 : Pekerjaan	X_9 : frekuensi rawat inap
X_5 : Pendapatan	X_{10} : frekuensi rawat jalan

Nilai *odds ratio* ditampilkan pada Tabel 4.33 yang berupa nilai Exp (B). Analisis nilai *odds ratio* pada model logit pertama dan kedua yaitu sebagai berikut.

a. Umur

Pada model logit pertama menunjukkan nilai koefisien yang positif yang berarti penduduk yang usianya semakin tua memiliki kecenderungan 1,026 kali lebih besar untuk menggunakan asuransi pemerintah daripada tidak menggunakannya.

b. Jenis kelamin

Variabel jenis kelamin pada model logit pertama menunjukkan nilai negatif yang berarti penduduk yang berjenis kelamin laki-laki dan perempuan berkecenderungan memilih menggunakan asuransi pemerintah sebesar 0,337 kali lebih kecil dibandingkan dengan tidak menggunakan asuransi. Pada model logit 2 untuk asuransi swasta juga menunjukkan nilai koefisien negatif yang berarti perbandingan antara penduduk yang berjenis kelamin laki-laki dan perempuan berkecenderungan memilih asuransi swasta sebesar 0,377 kali lebih kecil daripada tidak menggunakannya.

c. Pendidikan terakhir

Pada model logit pertama nilai koefisien pada variabel pendidikan terakhir$_1$ dan pendidikan terakhir$_2$ adalah positif yang berarti perbandingan penduduk dengan pendidikan terakhir SMP dan SMA sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan untuk menggunakan asuransi pemerintah

masing-masing sebesar 1,070 dan 1,219 kali lebih besar daripada tidak menggunakan asuransi. Sedangkan pada variabel pendidikan terakhir$_0$ memiliki koefisien negatif yang berarti perbandingan antara penduduk SD sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan menggunakan asuransi pemerintah 0,62 kali lebih kecil daripada tidak menggunakannya.

Nilai koefisien pada model logit kedua untuk pendidikan terakhir$_0$, pendidikan terakhir$_1$ dan pendidikan terakhir$_2$ positif yang berarti perbandingan penduduk SD, SMP dan SMA sederajat dengan yang tergolong perguruan tinggi memiliki kecenderungan untuk menggunakan asuransi swasta masing-masing sebesar 1.205, 1,958 dan 1,46 kali lebih besar daripada tidak menggunakan asuransi.

d. Pekerjaan

Pada model logit pertama dan kedua variabel pekerjaan$_0$, pekerjaan$_1$, pekerjaan$_2$ dan pekerjaan$_3$ menunjukkan nilai koefisien negatif. Hal ini menunjukkan bahwa perbandingan antara penduduk yang tergolong sekolah, bekerja, ibu rumah tangga dan pensiun dengan yang tergolong menganggur memiliki kecenderungan untuk menggunakan asuransi pemerintah masing-masing sebesar 0,731, 0,587, 0,443 dan 0,632 lebih kecil daripada tidak memiliki asuransi. Sedangkan kecenderungan untuk asuransi swasta masing-masing sebesar 0,906, 0,811, 0,386 dan 0,803 lebih kecil daripada tidak memiliki asuransi.

e. Pendapatan

Pada model logit pertama untuk variabel pendapatan$_0$ menunjukkan nilai koefisien negatif yang berarti perbandingan antara penduduk yang pendapatannya kurang dari Rp 1.000.000,00 dengan yang pendapatan yang tergolong lebih dari Rp 3.000.000,00 memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 0,611 kali lebih kecil dibandingkan dengan tidak memiliki asuransi. Sedangkan untuk model logit 2

memiliki nilai positif yang berarti kecenderungan menggunakan asuransi swasta sebesar 1,019 kali lebih besar daripada tidak asuransi.

Selanjutnya untuk variabel pendapatan$_1$ dan pendapatan$_s$ menunjukkan nilai koefisien positif. Hal ini menunjukkan bahwa penduduk yang pendapatanya Rp 1.000.000,00 - Rp 2.000.000,00 dan Rp 2.000.000,00 - Rp 3.000.000,00 dan yang tergolong pendapatan lebih dari Rp 3.000.000,00 memiliki kecenderungan menggunakan asuransi pemerintah masing-masing sebesar 1,71 dan 2,787 lebih besar dibandingkan tidak asuransi. Sedangkan kecenderungan untuk asuransi swasta masing-masing sebesar 1,019 dan 2,33 lebih besar dibandingkan tidak asuransi.

f. Status

Model logit untuk status$_0$ memiliki nilai positif yang berarti perbandingan antara penduduk yang belum kawin dengan yang tergolong cerai hidup/mati memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,416 kali, dan asuransi swasta sebesar 1,062 kali lebih besar daripada tidak menggunakannya. Sedangkan untuk penduduk yang sudah kawin dengan yang tergolong cerai hidup/mati memiliki kecenderungan untuk menggunakan asuransi pemerintah dan swasta sebesar 0,837 kali dan 0,787 kali lebih kecil dibandingkan tidak asuransi.

g. Tempat tinggal

Variabel tempat tinggal memiliki koefisien positif yang menunjukkan bahwa perbandingan penduduk rural dan urban memiliki kecenderungan untuk menggunakan asuransi pemerintah sebesar 1,97 kali lebih besar dibandingkan dengan tidak menggunakan asuransi, dan yang memilih asuransi swasta 1,602 kali lebih besar dibandingkan dengan tidak menggunakannya.

h. Riwayat penyakit

Model logit pertama dan kedua menunjukkan nilai koefisien positif yang berarti perbandingan antara penduduk yang memiliki riwayat penyakit dan uang tidak memiliki berkecenderungan memilih memnggunakan asuransi pemerintah sebesar 1,845 lebih besar dibandingkan dengan tidak menggunakannya, dan untuk asuransi swasta sebesar 1,429 kali.

i. Frekuensi rawat inap

Variabel frekuensi rawat inap menunjukkan nilai koefisien positif berarti semakin banyak frekuensi rawat inap maka kecenderungan penduduk untuk menggunakan asuransi pemerintah 2,051 kali lebih besar daripada tidak menggunakannya. Hal yang sama juga untuk asuransi swasta tetapi dengan perbandingan sebesar 2,025 kali.

j. Frekuensi rawat jalan

Semakin banyak frekuensi rawat jalan, kecenderungan memiliki asuransi pemerintah 1,897 kali lebih besar dibandingkan tidak asuransi. Sedangkan kecenderungan memiliki asuransi swasta 1,683 kali.

E. Uji Kesesuaian Model

Uji kesesuain model dilakukan untuk mengecek model yang telah terbentuuk sudah baik atau belum. Hipotesis yang digunakan yaitu sebagai berikut.

Hipotesis

H_0 : Model fit (tidak ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

H_1 : Model belum fit (ada perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi)

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

Hasil pengujian *goodness of fit* ditampilkan pada Tabel 4.34.

Tabel 4.34 *Goodness of Fit* Model Data *Balanced Combine Sampling*

	Value	**df**	***P-value***
Pearson Chi-Square	14792,289	7744	0,000

Berdasarkan Tabel 4.34 hasil dari *goodness of fit test* menunjukkan nilai P-value kurang dari taraf signifikansi 5%. Sehingga dapat disimpulkan terdapat perbedaan yang nyata antara hasil observasi dengan kemungkinan hasil prediksi.

F. Ketepatan Klasifikasi

Tahapan selanjutnya yaitu ketepatan klasifikasi yang bertujuan untuk mengetahui data yang digunakan telah diklasifikasikan dengan benar atau tidak. Hasil ketepatan klasifikasi berupa *confusion matrix* ditampilkan pada Tabel 4.35 sebagai berikut.

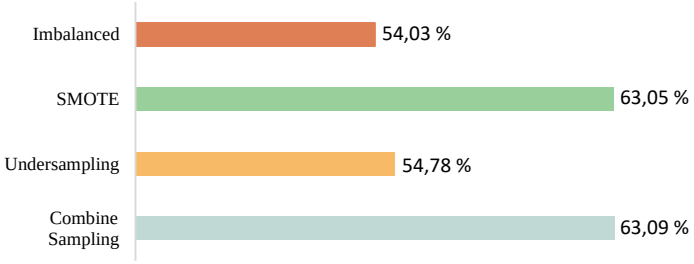
Tabel 4.35 *Confusion Matrix* pada *Balanced Combine Sampling*

Observed	***Predicted***		
	Tidak	**As.Pemerintah**	**As.Swasta**
Tidak	2790	873	647
As.Pemerintah	1550	1870	890
As. Swasta	1696	1253	1361

Tabel 4.35 menunjukkan bahwa hasil klasifikasi benar untuk kategori tidak asuransi, asuransi pemerintah dan swasta secara berturut-turut sebesar 2790, 1870 dan 1361 data. Dari *confusion matrix* diatas, didapatkan nilai *sensitivity* dan *specificity* sebesar 46,56% dan 73,28%. Sedangkan nilai AUC dari *preprocessing combine sampling* sebesar 63,09%.

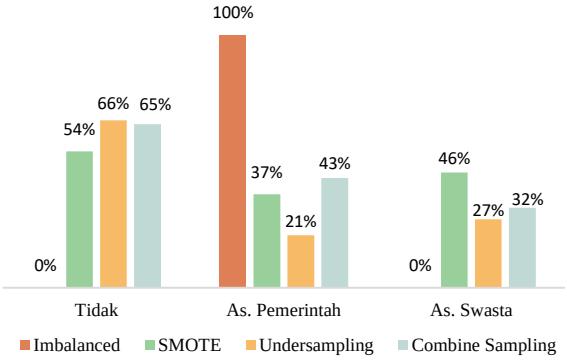
4.3 Perbandingan Performansi

Setelah dilakukan analisis status kepemilikan asuransi di Indonesia dengan metode *multinomial logistic regression*, selanjutnya akan dibandingkan ketepatan klasifikasi dari data *imbalanced* dan data *balanced* yang disajikan pada Gambar 4.9 berikut.



Gambar 4.9 Perbandingan AUC

Hasil perbandingan AUC pada Gambar 4.9 menunjukkan terjadi kenaikan yang signifikan dari data *imbalanced* ke data *balanced.* Sehingga metode *balancing* terbaik pada analisis status kepemilikan asuransi di Indonesia dilakukan dengan *random undersampling*, dimana model yang dihasilkan fit dengan AUC sebesar 54,78%. Meskipun tidak terjadi peningkatan nilai AUC yang cukup baik antara *imbalanced data* dan *balanced data* hasil *undersampling*, akan tetapi terjadi peningkatan nilai *sensitivity* yang signifikan disetiap kelasnya. Berikut ini visualisasi dari nilai *sensitivity* untuk setiap kelas.



Gambar 4.10 Perbandingan *Sensitivity* tiap Kelas

Berdasarkan Gambar 4.10 menunjukkan bahwa terdapat kenaikan nilai *sensitivity* antara *imbalanced data* dan *balanced data*

undersampling. Nilai *sensitivity* pada *imbalanced data* kelas minor yang cukup kecil yang berarti banyak terjadi kesalahan pada klasifikasi. Berdasarkan analisis status kepemilikan asuransi, status kepemilikan asuransi di Indonesia dipengaruhi oleh dua variabel. Untuk asuransi pemerintah dipengaruhi oleh tempat tinggal dengan nilai *oods ratio* sebesar 1,612 dan frekuensi rawat jalan dengan *odds ratio* sebesar 0,16. Sedangkan untuk asuransi swasta dipengaruhi oleh lokasi tempat tinggal dengan nilai *odds ratio* sebesar 1,673 yang berarti perbandingan penduduk rural dan urban memiliki kecenderungan untuk menggunakan asuransi swasta sebesar 1,673 kali lebih besar dibandingkan dengan tidak menggunakan asuransi.

(Halaman ini sengaja dikosongkan)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis dan pembahasan yang telah dilakukan pada bab 4, maka diperoleh kesimpulan sebagai berikut.

1. Status kepemilikan asuransi di Indonesia didominasi oleh asuransi pemerintah sebesar 64%. Sedangkan untuk asuransi swasta dan responden yang tidak memiliki asuransi secara berturut-turut sebesar 24% dan 12%. Berdasarkan faktor yang mempengaruhinya, Status kepemilikan asuransi di dominasi oleh penduduk yang berjenis kelamin laki-laki, tinggal di daerah urban, berpendidikan terakhir SD Sederajat dan bekerja. Penduduk yang
2. Analisis regresi logistik multinomial pada *imbalanced data* menghasilkan nilai AUC sebesar 54,03% dengan satu variabel yang berpengaruh signifikan terhadap status kepemilikan asuransi. Setelah dilakukan *balancing data* hasil *confusion matrix* tidak mengarah pada kelas mayor dan terjadi peningkatan nilai AUC. Pada metode *balancing random undersampling* didapatkan nilai AUC sebesar 54,78% dengan 2 variabel yang signifikan. Sedangkan untuk metode *balancing SMOTE* dan c*ombine sampling* semua variabel berpengaruh signifikan terhadap status kepemilikan asuransi dengan nilai AUC berturut-turut sebesar 63,05% dan 63,09%.
3. Hasil analisis status kepemilikan asuransi terbaik yaitu pada *balancing* dengan *random udersampling*, karena model yang dihasilkan fit dengan nilai AUC sebesar 54,78% dan *sensitivity* untuk kelas tidak asuransi, asuransi pemerintah dan asuransi swasta secara berturut-turut sebesar 53,9%, 36,97% dan 45,62%. Status kepemilikan asuransi dipengaruhi oleh dua variabel dimana untuk asuransi swasta dipengaruhi oleh lokasi tempat tinggal dengan nilai *odds*

ratio sebesar 1,673, dan untuk asuransi pemerintah dipengaruhi oleh lokasi tempat tinggal dan frekuensi rawat jalan dengan nilai *odds ratio* secara berturut-turut sebesar 1,612 dan 0,16.

5.2 Saran

Berdasarkan kesimpulan yang diperoleh, dapat dirumuskan saran sebagai pertimbangan penelitian selanjutnya yaitu menggunakan metode *balancing* data dan metode analisis yang lain sebagai pembanding untuk mendapatkan hasil yang lebih baik.

DAFTAR PUSTAKA

Aczel, A. D., dan Sounderpandian, J. (2008). *Complete Bussiness Statistics (7th ed.).* USA: The McGraw-Hill.

Agresti, A. (2013). *Categorical Data Analysis (3rd ed.).* New Jersey: John Wiley and Sons.

Asampana, G., Nantomah, K. K., dan Tungosiamu, E. A. (2017). Multinomial Logistic Regression Analysis of The Determinants of Students Academic Performance in Mathematics at Basic Education Certificate Examination. *Journal Science Publishing Group*, Vol. 2, 22-26.

Bernal, N., Carpio, M., dan Klein, T. (2017). The Effects of Access to Health Insurance : Evidence from a Regression Discontinuity Design in Peru. *Journal of Public Economics*, Vol. 154, 122-136.

Chawla, N., Bowyer, K., Hall, L., dan Kegelmeyer, W. (2002). SMOTE : Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 321-357.

Darmawi, H. (2004). *Manajemen Asuransi*. Jakarta: PT Bumi Aksara.

Erlangga, D., Ali, S., dan Bloor, K. (2019). The Impact of Public Health Insurance on Healthcare Utilisation in Indonesia : Evidence from Panel Data. *International Journal of Public Health*, Vol. 64, 603-613.

Feldstein, P. J. (1988). *Health Care Economics*. New York: John and Wiley Sons.

Gaudio, R., Batista, G., dan Branco, A. (2013). Coping with Highly Imbalanced Dataset : A Case Study with Definition Extraction in a Multilingual Setting. *Natural Language Engineering*, Vol. 19, 1-33.

Gujarati, D. N., dan Porter, D. C. (2009). *Basic Econometrics fifth edition.* New York: The McGraw-Hill.

Hand, D. J., dan Till, R. J. (2001). A Simple Generalisation of The Area Under The ROC Curve for Multiple Class Classification Problems. *Machine Learning*, Vol. 45, 171-186.

HIAA. (2000). *The Health Insurance Primer.* Amerika: HIAA.

Hidayat, B., Thabrany, H., Dong, H., dan Sauerborn, R. (2004). The Effects of Mandatory Health Insurance on Equity in Access to Outpatient Care in Indonesia. *Journal Health Policy and Planning*, Vol. 19, 322-335.

Hosmer, D.W., dan Lemeshow, J.S. (2000). *Applied Logistic Regression.* New York: John Wiley and Sons.

Japkowicz, N., dan Stephen, S. (2002). The Class Imbalanced Problem : A Systematic Study. *Intelligent Data Analysis*, Vol. 6, 429-449.

Johnson, R. A., dan Winchern, D. W. (1992). *Applied Multivariate Statistical Analysis.* United States: Prentice Hall.

Kumar, A.S., Kumar, S., Abraham, S., dan Rao, P. (2012). Leprosy Amng Tribal Population of Chhattisgarh State, India. *Indian Journal of Leprosy*, Vol. 84, 265-269.

Liu, K., Cook, B., dan Lu, C. (2019). Health Inequity and Community-based Health Insurance : a Case Study of Rural Rwanda with Repeated Cross-sectional Data. *International Journal of Public Health*, Vol. 64, 7-14.

Mudhu, Ashock, N., dan Balasubramanian, S. (2014). A Multinomial Logistic Regression Analysis to Study The Influence of Residence and Socio-economic Status on Breast Cancer

Incidances In Southern Karnataka. *International Journal of Mathematics and Statistics Invention*, Vol. 2, 01-08.

Roemer, M. I. (1981). More Data on Post-Surgical Deaths Related to The 1976 Los Angeles Doctor Slowdown. *Journal Social Science and Medicine*, Vol. 15, 161-163.

Salim, A. (2006). *Teori dan Paradigma Penelitian Sosial.* Yogyakarta: Tiara Wacana.

Solberg, A., dan Solberg, R. (1996). A Large Scale Evaluation of Features for Automatic Detection of Oil Spills in ERS SAR Images. *IEEE International Geoscience and Remote Sensing Symposium*, Vol. 3, 1484-1486.

Strauss, J., Witoelar, F., dan Sikoki, B. (2016). *The Fifth Wave of The Indonesia Family Life Survey : Overview and Field Report.* California: RAND Corporation.

Walpole, R. E. (2012). *Pengantar Metode Statistika Edisi ke-3. Diterjemahkan oleh : Bambang Sumantri.* Jakarta: Gramedia Pustaka Utama.

Yu, X., Zhou, M., Chen, X., Deng, L., dan Wang, L. (2013). Improving Protein-ATP Binding Residues Prediction by Boosting SVMs with Random Under-sampling. *Neurocomputing*, Vol. 104, 180-190.

Zerriaa, M., dan Noubbigh, H. (2016). Determinants of Life Insurance Demand in The Mena Region. *The Geneva Papers on Risk and Insurance*, Vol. 41, 491-511.

(Halaman ini sengaja dikosongkan)

LAMPIRAN

Lampiran 1. Data *Imbalanced* Status Kepemilikan Asuransi

Data	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0	59	0	0	0	0	1	0	0	0	0
2	0	29	1	0	2	0	1	0	1	1	0
3	0	59	0	0	1	3	2	0	1	0	0
4	0	36	0	0	1	3	1	0	1	0	0
5	0	55	0	0	0	3	1	0	1	0	1
6	0	34	0	3	1	3	1	0	1	1	0
7	0	24	0	3	0	2	0	1	1	0	0
8	0	23	0	3	0	0	0	1	1	0	0
9	1	50	0	0	1	3	1	0	1	0	0
10	1	30	0	1	1	3	1	0	1	0	0
11	1	58	0	0	2	0	1	0	1	1	0
12	1	27	0	0	1	3	1	0	1	0	0
13	2	29	0	0	1	3	1	0	1	0	2
14	2	25	0	0	4	3	1	0	1	0	0
15	1	29	0	1	1	3	1	1	0	0	0
16	1	19	1	1	1	3	0	1	1	1	0
17	1	42	0	0	1	3	1	0	1	0	2
18	1	46	0	0	1	0	1	0	1	0	0
19	1	67	1	0	3	0	2	0	1	0	0
20	1	46	0	2	1	0	1	1	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
6854	2	56	1	2	2	0	1	1	1	0	0
6855	1	64	0	0	3	0	1	1	1	0	0

Lampiran 2. Data *Balanced* SMOTE Status Kepemilikan Asuransi

Data	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0	59	0	0	0	0	1	0	0	0	0
2	0	29	1	0	2	0	1	0	1	1	0
3	0	59	0	0	1	3	2	0	1	0	0
4	0	36	0	0	1	3	1	0	1	0	0
5	0	55	0	0	0	3	1	0	1	0	1
6	0	34	0	3	1	3	1	0	1	1	0
7	0	24	0	3	0	2	0	1	1	0	0
8	0	23	0	3	0	0	0	1	1	0	0
9	1	50	0	0	1	3	1	0	1	0	0
10	1	30	0	1	1	3	1	0	1	0	0
11	1	58	0	0	2	0	1	0	1	1	0
12	1	27	0	0	1	3	1	0	1	0	0
13	2	29	0	0	1	3	1	0	1	0	2
14	2	25	0	0	4	3	1	0	1	0	0
15	1	29	0	1	1	3	1	1	0	0	0
16	1	19	1	1	1	3	0	1	1	1	0
17	1	42	0	0	1	3	1	0	1	0	2
18	1	46	0	0	1	0	1	0	1	0	0
19	1	67	1	0	3	0	2	0	1	0	0
20	1	46	0	2	1	0	1	1	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
13568	2	41	1	1	1	0	1	1	1	0	0
13569	2	41	1	1	1	0	1	1	1	0	0

Lampiran 3. Data *Balanced Undersampling* Status Kepemilikan Asuransi

Data	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0	64	0	1	3	0	1	1	1	0	0
2	0	32	0	1	1	2	1	0	1	0	0
3	0	36	0	0	1	3	1	0	1	0	0
4	0	37	0	1	1	3	1	1	1	0	0
5	0	39	0	3	1	3	1	1	1	0	0
6	0	35	0	3	1	3	1	1	1	1	0
7	0	24	1	3	4	3	0	1	1	1	0
8	0	43	0	2	1	3	1	1	1	0	0
9	0	25	1	2	1	3	1	1	1	1	0
10	0	26	0	0	1	3	1	0	1	1	0
11	0	58	0	1	1	3	1	1	1	0	1
12	0	42	0	2	1	3	1	1	1	1	1
13	0	58	0	0	1	3	1	1	1	0	0
14	0	56	0	3	1	3	1	1	1	0	0
15	0	57	0	1	0	0	1	1	1	0	0
16	0	43	0	0	1	3	1	0	1	0	0
17	0	49	0	2	1	0	1	1	1	0	0
18	0	61	0	2	4	0	1	1	1	0	0
19	0	43	0	0	1	3	1	0	1	0	0
20	0	24	0	3	0	2	0	1	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
2585	2	29	0	1	1	0	1	0	1	0	0
2586	2	56	0	0	1	3	1	1	1	1	1

Lampiran 4. Data *Balanced Combine Sampling* Status Kepemilikan Asuransi

Data	**Y**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**
1	0	52	0	0	1	3	1	1	1	0	0
2	0	37	0	2	1	3	1	1	1	1	0
3	0	35	0	2	2	0	1	0	1	0	0
4	0	29	1	0	2	0	1	0	1	0	0
5	0	36	0	1	1	3	1	1	1	1	0
6	0	46	0	0	1	3	1	0	1	0	0
7	0	56	0	0	1	1	2	1	1	0	0
8	0	42	0	2	1	3	1	1	1	0	0
9	0	16	0	2	1	3	0	1	1	0	0
10	0	34	0	2	1	3	1	1	1	0	0
11	0	35	0	3	1	3	1	1	1	0	0
12	0	54	0	0	1	0	1	0	1	0	0
13	0	56	0	0	1	0	2	1	1	0	0
14	0	39	0	2	1	3	1	1	1	0	0
15	0	36	0	1	1	3	1	1	1	1	0
16	0	30	0	0	1	3	1	1	1	0	0
17	0	51	0	0	1	0	1	0	1	0	0
18	0	35	0	0	1	3	1	1	1	0	0
19	0	54	0	0	1	0	1	0	1	0	0
20	0	36	0	1	1	3	1	0	1	0	0
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
12929	2	31	1	2	2	0	1	1	1	0	0
12930	2	30	0	2	1	3	1	1	1	0	0

Lampiran 5. Output SPSS Karakteristik Data

Pendidikan * Asuransi Crosstabulation

Count

		Asuransi			
		Tidak Asuransi	As. Pemerintah	As. Swasta	Total
Pendidikan	SD Sederajat	330	1629	585	2544
	SMP Sederajat	141	750	284	1175
	SMA Sederajat	242	1324	508	2074
	Perguruan Tinggi	149	657	256	1062
Total		862	4360	1633	6855

Pekerjaan * Asuransi Crosstabulation

Count

		Asuransi			
		Tidak Asuransi	As. Pemerintah	As. Swasta	Total
Pekerjaan	Sekolah	42	202	75	319
	Bekerja	640	3350	1244	5234
	Ibu Rumah Tangga	68	308	133	509
	Pensiun	51	229	74	354
	Menganggur	61	271	107	439
Total		862	4360	1633	6855

Lampiran 6. *Output* SPSS Data *Imbalanced*

MULTIKOLINIERITAS

Correlations

Control Variables			Umur	Rawat Inap	Rawat Jalan
Asuransi	Umur	Correlation	1,000	,002	-,028
		Significance (2-tailed)	.	,868	,021
		df	0	6852	6852
	Rawat Inap	Correlation	,002	1,000	-,020
		Significance (2-tailed)	,868	.	,091
		df	6852	0	6852
	Rawat Jalan	Correlation	-,028	-,020	1,000
		Significance (2-tailed)	,021	,091	.
		df	6852	6852	0

INDEPENDENSI

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	,100[a]	2	,951
Likelihood Ratio	,100	2	,951
Linear-by-Linear Association	,023	1	,879
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 131.15.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,596[a]	6	,470
Likelihood Ratio	5,575	6	,472
Linear-by-Linear Association	,510	1	,475
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 133.54.

Lampiran 6. Output *SPSS Data* Imbalanced (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,081[a]	8	,638
Likelihood Ratio	6,050	8	,642
Linear-by-Linear Association	,570	1	,450
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 40.11.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,011[a]	6	,675
Likelihood Ratio	4,023	6	,674
Linear-by-Linear Association	,632	1	,427
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.22.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,664[a]	4	,616
Likelihood Ratio	2,620	4	,623
Linear-by-Linear Association	2,073	1	,150
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 65.64.

Lampiran 6. Output *SPSS Data* Imbalanced (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	37,525[a]	2	,000
Likelihood Ratio	38,739	2	,000
Linear-by-Linear Association	13,889	1	,000
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 333.11.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,078[a]	2	,354
Likelihood Ratio	2,162	2	,339
Linear-by-Linear Association	,015	1	,902
N of Valid Cases	6855		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.81.

REGRESI LOGISTIK MULTINOMIAL

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	3209,246			
Final	3163,018	46,228	8	,000

Lampiran 6. Output *SPSS Data* Imbalanced (Lanjutan)

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1628,796	1594	,266
Deviance	1631,677	1594	,250

Pseudo R-Square

Cox and Snell	,007
Nagelkerke	,008
McFadden	,004

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	3163,018[a]	,000	0	.
X1	3166,738	3,720	2	,156
X9	3164,672	1,653	2	,437
X10	3165,089	2,071	2	,355
X7	3202,387	39,368	2	,000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Lampiran 6. Output *SPSS Data* Imbalanced (Lanjutan)

Parameter Estimates

Asuransi[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
As. Pemerintah	Intercept	1,511	,117	167,574	1	,000			
	X1	-,002	,002	,631	1	,427	,998	,993	1,003
	X9	,021	,064	,107	1	,744	1,021	,900	1,158
	X10	,092	,066	1,913	1	,167	1,096	,962	1,249
	[X7=0]	,497	,081	37,346	1	,000	1,643	1,401	1,927
	[X7=1]	0[b]	.	.	0	.	.	.	.
As. Swasta	Intercept	,666	,132	25,571	1	,000			
	X1	-,005	,003	3,142	1	,076	,995	,990	1,001
	X9	,073	,070	1,089	1	,297	1,076	,938	1,234
	X10	,072	,074	,965	1	,326	1,075	,931	1,242
	[X7=0]	,454	,091	25,065	1	,000	1,575	1,318	1,882
	[X7=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Tidak Asuransi.

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

Classification

Observed	Predicted			
	Tidak Asuransi	As. Pemerintah	As. Swasta	Percent Correct
Tidak Asuransi	0	862	0	0,0%
As. Pemerintah	0	4360	0	100,0%
As. Swasta	0	1633	0	0,0%
Overall Percentage	0,0%	100,0%	0,0%	63,6%

Output R

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2
         0    0    0    0
         1  862 4360 1633
         2    0    0    0
```

Lampiran 6. Output *SPSS Data* Imbalanced (Lanjutan)

```
Overall Statistics
               Accuracy : 0.636
                 95% CI : (0.6245, 0.6474)
    No Information Rate : 0.636
    P-Value [Acc > NIR] : 0.5055
                  Kappa : 0

 Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: 0 Class: 1 Class: 2
Sensitivity            0.0000    1.000   0.0000
Specificity            1.0000    0.000   1.0000
Pos Pred Value            NaN    0.636      NaN
Neg Pred Value         0.8743      NaN   0.7618
Prevalence             0.1257    0.636   0.2382
Detection Rate         0.0000    0.636   0.0000
Detection Prevalence   0.0000    1.000   0.0000
Balanced Accuracy      0.5000    0.500   0.5000

Call:
multiclass.roc.default(response = data$Y, predictor = predict
ed_scores)

Data: multivariate predictor predicted_scores with 3 levels o
f data$Y: 0, 1, 2.
Multi-class area under the curve: 0.5403
```

Lampiran 7. *Output* SPSS Data *Balanced* SMOTE

MULTIKOLINIERITAS

Correlations

Control Variables			Umur	Rawat Inap	Rawat Jalan
Asuransi	Umur	Correlation	1,000	,004	-,029
		Significance (2-tailed)	.	,640	,001
		df	0	13566	13566
	Rawat Inap	Correlation	,004	1,000	-,021
		Significance (2-tailed)	,640	.	,016
		df	13566	0	13566
	Rawat Jalan	Correlation	-,029	-,021	1,000
		Significance (2-tailed)	,001	,016	.
		df	13566	13566	0

INDEPENDENSI

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	130,445[a]	2	,000
Likelihood Ratio	140,359	2	,000
Linear-by-Linear Association	98,112	1	,000
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 555.54.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	111,495[a]	6	,000
Likelihood Ratio	111,944	6	,000
Linear-by-Linear Association	12,010	1	,001
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 599.70.

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	49,151[a]	8	,000
Likelihood Ratio	48,956	8	,000
Linear-by-Linear Association	,029	1	,864
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 186.77.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	84,311[a]	6	,000
Likelihood Ratio	92,608	6	,000
Linear-by-Linear Association	11,216	1	,001
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 54.32.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	76,527[a]	4	,000
Likelihood Ratio	78,027	4	,000
Linear-by-Linear Association	,681	1	,409
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 311.28.

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	138,127[a]	2	,000
Likelihood Ratio	140,651	2	,000
Linear-by-Linear Association	93,393	1	,000
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1573.89.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,562[a]	2	,038
Likelihood Ratio	6,478	2	,039
Linear-by-Linear Association	,742	1	,389
N of Valid Cases	13569		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 47.01.

REGRESI LOGISTIK MULTINOMIAL

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	19799,185			
Final	18464,799	1334,386	36	,000

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	15490,159	7898	,000
Deviance	16138,163	7898	,000

Pseudo R-Square

Cox and Snell	,094
Nagelkerke	,105
McFadden	,045

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	18464,799[a]	,000	0	.
X1	18699,216	234,416	2	,000
X9	18698,025	233,225	2	,000
X10	18639,112	174,313	2	,000
X2	18624,326	159,526	2	,000
X3	18691,126	226,326	6	,000
X4	18533,396	68,597	8	,000
X5	18619,460	154,661	6	,000
X6	18495,861	31,061	4	,000
X7	18662,029	197,230	2	,000
X8	18472,722	7,923	2	,019

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

Parameter Estimates

Asuransi[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Asuransi Pemerintah	Intercept	,342	,193	3,129	1	,077			
	X1	,026	,002	153,032	1	,000	1,026	1,022	1,030
	X9	,720	,055	169,391	1	,000	2,054	1,843	2,289
	X10	,635	,054	136,372	1	,000	1,887	1,696	2,099
	[X2=0]	-1,095	,096	129,138	1	,000	,335	,277	,404
	[X2=1]	0[b]	.	.	0	.	.	.	.
	[X3=0]	-,482	,075	41,745	1	,000	,618	,534	,715
	[X3=1]	,067	,081	,676	1	,411	1,069	,912	1,254
	[X3=2]	,192	,071	7,260	1	,007	1,211	1,054	1,393
	[X3=3]	0[b]	.	.	0	.	.	.	.
	[X4=0]	-,321	,155	4,273	1	,039	,725	,535	,983
	[X4=1]	-,535	,105	25,788	1	,000	,585	,476	,720
	[X4=2]	-,819	,150	30,002	1	,000	,441	,329	,591
	[X4=3]	-,445	,151	8,745	1	,003	,641	,477	,861
	[X4=4]	0[b]	.	.	0	.	.	.	.
	[X5=0]	-,495	,060	68,129	1	,000	,610	,542	,686
	[X5=1]	,561	,182	9,509	1	,002	1,752	1,227	2,501
	[X5=2]	1,064	,250	18,067	1	,000	2,897	1,774	4,730
	[X5=3]	0[b]	.	.	0	.	.	.	.
	[X6=0]	,354	,149	5,616	1	,018	1,425	1,063	1,909
	[X6=1]	-,185	,101	3,355	1	,067	,831	,681	1,013
	[X6=2]	0[b]	.	.	0	.	.	.	.
	[X7=0]	,681	,050	183,962	1	,000	1,976	1,791	2,180
	[X7=1]	0[b]	.	.	0	.	.	.	.
	[X8=0]	,604	,218	7,672	1	,006	1,829	1,193	2,805
	[X8=1]	0[b]	.	.	0	.	.	.	.
Asuransi Swasta	Intercept	,874	,191	20,893	1	,000			
	X1	-,001	,002	,337	1	,561	,999	,995	1,003
	X9	,702	,054	166,237	1	,000	2,018	1,814	2,245
	X10	,516	,054	91,154	1	,000	1,676	1,507	1,863
	[X2=0]	-,976	,094	108,644	1	,000	,377	,314	,453
	[X2=1]	0[b]	.	.	0	.	.	.	.
	[X3=0]	,181	,075	5,905	1	,015	1,199	1,036	1,387
	[X3=1]	,659	,080	67,611	1	,000	1,933	1,652	2,261
	[X3=2]	,385	,073	27,878	1	,000	1,469	1,274	1,695
	[X3=3]	0[b]	.	.	0	.	.	.	.
	[X4=0]	-,137	,152	,814	1	,367	,872	,648	1,174
	[X4=1]	-,253	,105	5,777	1	,016	,776	,632	,954
	[X4=2]	-,930	,148	39,233	1	,000	,395	,295	,528
	[X4=3]	-,253	,154	2,706	1	,100	,777	,575	1,050
	[X4=4]	0[b]	.	.	0	.	.	.	.
	[X5=0]	-,006	,055	,012	1	,912	,994	,892	1,108
	[X5=1]	,881	,173	25,894	1	,000	2,414	1,719	3,390
	[X5=2]	1,005	,250	16,225	1	,000	2,732	1,675	4,456
	[X5=3]	0[b]	.	.	0	.	.	.	.
	[X6=0]	,029	,145	,041	1	,840	1,030	,775	1,369
	[X6=1]	-,256	,100	6,622	1	,010	,774	,637	,941
	[X6=2]	0[b]	.	.	0	.	.	.	.
	[X7=0]	,480	,048	98,831	1	,000	1,616	1,470	1,776
	[X7=1]	0[b]	.	.	0	.	.	.	.
	[X8=0]	,343	,224	2,348	1	,125	1,410	,909	2,187
	[X8=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Tidak Asuransi.

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

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

Classification

Observed	Predicted			
	Tidak Asuransi	Asuransi Pemerintah	Asuransi Swasta	Percent Correct
Tidak Asuransi	2323	712	1275	53,9%
Asuransi Pemerintah	1274	1612	1474	37,0%
Asuransi Swasta	1494	1170	2235	45,6%
Overall Percentage	37,5%	25,7%	36,7%	45,5%

Output R

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2
         0 2323 1274 1494
         1  712 1612 1170
         2 1275 1474 2235

Overall Statistics
               Accuracy : 0.4547
                 95% CI : (0.4463, 0.4631)
    No Information Rate : 0.361
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.1806
 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:
                     Class: 0 Class: 1 Class: 2
Sensitivity            0.5390   0.3697   0.4562
Specificity            0.7010   0.7956   0.6829
Pos Pred Value         0.4563   0.4614   0.4484
Neg Pred Value         0.7656   0.7272   0.6897
Prevalence             0.3176   0.3213   0.3610
Detection Rate         0.1712   0.1188   0.1647
Detection Prevalence   0.3752   0.2575   0.3673
Balanced Accuracy      0.6200   0.5827   0.5696
```

Lampiran 7. Output *SPSS Data* Balanced *SMOTE* (Lanjutan)

```
Call:
multiclass.roc.default(response = data$Y, predictor = predict
ed_scores)
Data: multivariate predictor predicted_scores with 3 levels o
f data$Y: 0, 1, 2.

Multi-class area under the curve: 0.6305
```

Lampiran 8. *Output* SPSS Data *Balanced Random Undersampling*

MULTIKOLINIERITAS

Correlations

Control Variables			Umur	Rawat Inap	Rawat Jalan
Asuransi	Umur	Correlation	1,000	-,005	-,014
		Significance (2-tailed)	.	,815	,466
		df	0	2583	2583
	Rawat Inap	Correlation	-,005	1,000	-,023
		Significance (2-tailed)	,815	.	,246
		df	2583	0	2583
	Rawat Jalan	Correlation	-,014	-,023	1,000
		Significance (2-tailed)	,466	,246	.
		df	2583	2583	0

INDEPENDENSI

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,076[a]	2	,354
Likelihood Ratio	2,092	2	,351
Linear-by-Linear Association	,222	1	,637
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 129.67.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,320[a]	6	,633
Likelihood Ratio	4,330	6	,632
Linear-by-Linear Association	,314	1	,575
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 140.33.

Lampiran 8. Output *SPSS Data Balanced Random Undersampling* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,467[a]	8	,707
Likelihood Ratio	5,541	8	,698
Linear-by-Linear Association	1,247	1	,264
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 46.00.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,246[a]	6	,513
Likelihood Ratio	5,209	6	,517
Linear-by-Linear Association	1,425	1	,233
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.33.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,356[a]	4	,671
Likelihood Ratio	2,334	4	,675
Linear-by-Linear Association	,281	1	,596
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 65.33.

Lampiran 8. Output *SPSS Data Balanced Random Undersampling* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29,906[a]	2	,000
Likelihood Ratio	30,446	2	,000
Linear-by-Linear Association	24,014	1	,000
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 315.00.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,667[a]	2	,435
Likelihood Ratio	1,735	2	,420
Linear-by-Linear Association	,592	1	,442
N of Valid Cases	2586		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.67.

REGRESI LOGISTIK MULTINOMIAL

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2238,245			
Final	2198,571	39,674	8	,000

Lampiran 8. Output *SPSS Data Balanced Random Undersampling* (Lanjutan)

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1079,289	1056	,302
Deviance	1235,569	1056	,000

Pseudo R-Square

Cox and Snell	,015
Nagelkerke	,017
McFadden	,007

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	2198,571[a]	,000	0	.
X1	2200,033	1,463	2	,481
X9	2201,239	2,668	2	,263
X10	2203,667	5,097	2	,078
X7	2229,933	31,363	2	,000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Lampiran 8. Output *SPSS Data Balanced Random Undersampling* (Lanjutan)

Parameter Estimates

Asuransi[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
As. Pemerintah	Intercept	-,112	,152	,539	1	,463			
	X1	-,002	,003	,369	1	,544	,998	,992	1,004
	X9	,000	,086	,000	1	,998	1,000	,845	1,183
	X10	,160	,080	3,963	1	,047	1,174	1,002	1,374
	[X7=0]	,477	,103	21,631	1	,000	1,612	1,318	1,971
	[X7=1]	0[b]	.	.	0	.	.	.	.
As. Swasta	Intercept	-,070	,152	,209	1	,647			
	X1	-,004	,003	1,461	1	,227	,996	,990	1,002
	X9	,110	,080	1,855	1	,173	1,116	,953	1,306
	X10	,155	,081	3,694	1	,055	1,168	,997	1,368
	[X7=0]	,515	,103	25,184	1	,000	1,673	1,368	2,045
	[X7=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Tidak Asuransi.

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

Classification

Observed	Predicted			
	Tidak Asuransi	As. Pemerintah	As. Swasta	Percent Correct
Tidak Asuransi	571	143	148	66,2%
As. Pemerintah	476	179	207	20,8%
As. Swasta	467	162	233	27,0%
Overall Percentage	58,5%	18,7%	22,7%	38,0%

Output R

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0 571 476 467
         1 143 179 162
         2 148 207 233
```

Lampiran 8. Output *SPSS Data Balanced Random Undersampling* (Lanjutan)

```
Overall Statistics
               Accuracy : 0.3801
                 95% CI : (0.3614, 0.3992)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 3.251e-07
                  Kappa : 0.0702
 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:
                     Class: 0 Class: 1 Class: 2
Sensitivity            0.6624  0.20766   0.2703
Specificity            0.4530  0.82309   0.7941
Pos Pred Value         0.3771  0.36983   0.3963
Neg Pred Value         0.7285  0.67507   0.6852
Prevalence             0.3333  0.33333   0.3333
Detection Rate         0.2208  0.06922   0.0901
Detection Prevalence   0.5855  0.18716   0.2274
Balanced Accuracy      0.5577  0.51537   0.5322

Call:
multiclass.roc.default(response = data$Y, predictor = predict
ed_scores)

Data: multivariate predictor predicted_scores with 3 levels o
f data$Y: 0, 1, 2.
Multi-class area under the curve: 0.5478
```

Lampiran 9. *Output* SPSS Data *Balanced Combine Sampling*

MULTIKOLINIERITAS

Correlations

Control Variables			Umur	Rawat Inap	Rawat Jalan
Asuransi	Umur	Correlation	1,000	,004	-,030
		Significance (2-tailed)	.	,633	,001
		df	0	12927	12927
	Rawat Inap	Correlation	,004	1,000	-,021
		Significance (2-tailed)	,633	.	,016
		df	12927	0	12927
	Rawat Jalan	Correlation	-,030	-,021	1,000
		Significance (2-tailed)	,001	,016	.
		df	12927	12927	0

INDEPENDENSI

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	128,679[a]	2	,000
Likelihood Ratio	137,801	2	,000
Linear-by-Linear Association	98,694	1	,000
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 551,00.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	107,807[a]	6	,000
Likelihood Ratio	107,951	6	,000
Linear-by-Linear Association	10,872	1	,001
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 605,00.

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	49,976[a]	8	,000
Likelihood Ratio	49,581	8	,000
Linear-by-Linear Association	,014	1	,907
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 189,33.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	80,761[a]	6	,000
Likelihood Ratio	87,981	6	,000
Linear-by-Linear Association	11,646	1	,001
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 53,00.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79,405[a]	4	,000
Likelihood Ratio	80,690	4	,000
Linear-by-Linear Association	,663	1	,416
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 312,33.

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	132,534[a]	2	,000
Likelihood Ratio	134,770	2	,000
Linear-by-Linear Association	89,360	1	,000
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 1563,67.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,718[a]	2	,035
Likelihood Ratio	6,679	2	,035
Linear-by-Linear Association	1,061	1	,303
N of Valid Cases	12930		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 47,67.

REGRESI LOGISTIK MULTINOMIAL

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	19013,076			
Final	17704,192	1308,883	36	,000

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	14792,289	7744	,000
Deviance	15448,865	7744	,000

Pseudo R-Square

Cox and Snell	,096
Nagelkerke	,108
McFadden	,046

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	17704,192[a]	,000	0	.
X1	17929,623	225,431	2	,000
X9	17933,550	229,357	2	,000
X10	17879,727	175,534	2	,000
X2	17862,837	158,644	2	,000
X3	17927,077	222,885	6	,000
X4	17777,841	73,649	8	,000
X5	17853,646	149,454	6	,000
X6	17732,742	28,550	4	,000
X7	17896,259	192,066	2	,000
X8	17712,337	8,144	2	,017

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

Parameter Estimates

Asuransi[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Asuransi Pemerintah	Intercept	,303	,194	2,428	1	,119			
	X1	,026	,002	154,087	1	,000	1,026	1,022	1,031
	X9	,718	,056	167,243	1	,000	2,051	1,839	2,287
	X10	,640	,055	137,100	1	,000	1,897	1,704	2,111
	[X2=0]	-1,089	,097	126,415	1	,000	,337	,278	,407
	[X2=1]	0[b]	.	.	0	.	.	.	.
	[X3=0]	-,478	,075	40,699	1	,000	,620	,536	,718
	[X3=1]	,068	,082	,690	1	,406	1,070	,912	1,256
	[X3=2]	,198	,072	7,658	1	,006	1,219	1,059	1,402
	[X3=3]	0[b]	.	.	0	.	.	.	.
	[X4=0]	-,313	,156	4,038	1	,044	,731	,539	,992
	[X4=1]	-,533	,106	25,411	1	,000	,587	,477	,722
	[X4=2]	-,814	,150	29,375	1	,000	,443	,330	,595
	[X4=3]	-,459	,151	9,233	1	,002	,632	,470	,850
	[X4=4]	0[b]	.	.	0	.	.	.	.
	[X5=0]	-,492	,060	66,594	1	,000	,611	,543	,688
	[X5=1]	,536	,183	8,581	1	,003	1,710	1,194	2,448
	[X5=2]	1,025	,252	16,588	1	,000	2,787	1,702	4,564
	[X5=3]	0[b]	.	.	0	.	.	.	.
	[X6=0]	,348	,150	5,375	1	,020	1,416	1,055	1,900
	[X6=1]	-,178	,102	3,084	1	,079	,837	,686	1,021
	[X6=2]	0[b]	.	.	0	.	.	.	.
	[X7=0]	,678	,050	180,440	1	,000	1,970	1,784	2,174
	[X7=1]	0[b]	.	.	0	.	.	.	.
	[X8=0]	,613	,218	7,884	1	,005	1,845	1,203	2,830
	[X8=1]	0[b]	.	.	0	.	.	.	.
Asuransi Swasta	Intercept	,683	,197	11,999	1	,001			
	X1	-,001	,002	,114	1	,736	,999	,995	1,003
	X9	,706	,055	162,542	1	,000	2,025	1,817	2,257
	X10	,520	,055	89,401	1	,000	1,683	1,511	1,874
	[X2=0]	-,998	,096	108,561	1	,000	,369	,305	,445
	[X2=1]	0[b]	.	.	0	.	.	.	.
	[X3=0]	,186	,077	5,819	1	,016	1,205	1,036	1,402
	[X3=1]	,672	,083	66,104	1	,000	1,958	1,665	2,303
	[X3=2]	,379	,076	25,130	1	,000	1,460	1,259	1,693
	[X3=3]	0[b]	.	.	0	.	.	.	.
	[X4=0]	-,098	,156	,399	1	,528	,906	,668	1,230
	[X4=1]	-,210	,109	3,723	1	,054	,811	,655	1,003
	[X4=2]	-,952	,154	38,463	1	,000	,386	,286	,521
	[X4=3]	-,219	,158	1,920	1	,166	,803	,589	1,095
	[X4=4]	0[b]	.	.	0	.	.	.	.
	[X5=0]	,018	,057	,104	1	,747	1,019	,911	1,139
	[X5=1]	,846	,178	22,667	1	,000	2,330	1,645	3,300
	[X5=2]	,998	,254	15,454	1	,000	2,713	1,649	4,461
	[X5=3]	0[b]	.	.	0	.	.	.	.
	[X6=0]	,060	,149	,164	1	,685	1,062	,793	1,422
	[X6=1]	-,239	,102	5,459	1	,019	,787	,644	,962
	[X6=2]	0[b]	.	.	0	.	.	.	.
	[X7=0]	,471	,050	89,750	1	,000	1,602	1,453	1,766
	[X7=1]	0[b]	.	.	0	.	.	.	.
	[X8=0]	,357	,229	2,424	1	,119	1,429	,912	2,238
	[X8=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Tidak Asuransi.

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

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

Classification

Observed	Predicted			
	Tidak Asuransi	Asuransi Pemerintah	Asuransi Swasta	Percent Correct
Tidak Asuransi	2790	873	647	64,7%
Asuransi Pemerintah	1550	1870	890	43,4%
Asuransi Swasta	1696	1253	1361	31,6%
Overall Percentage	46,7%	30,9%	22,4%	46,6%

Output R

```
Confusion Matrix and Statistics
          Reference
Prediction    0    1    2
         0 2790 1550 1696
         1  873 1869 1253
         2  647  891 1361
Overall Statistics
               Accuracy : 0.4656
                 95% CI : (0.457, 0.4742)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.1984
 Mcnemar's Test P-Value : < 2.2e-16
Statistics by Class:
                     Class: 0 Class: 1 Class: 2
Sensitivity            0.6473   0.4336   0.3158
Specificity            0.6234   0.7534   0.8216
Pos Pred Value         0.4622   0.4678   0.4695
Neg Pred Value         0.7795   0.7268   0.7060
Prevalence             0.3333   0.3333   0.3333
Detection Rate         0.2158   0.1445   0.1053
Detection Prevalence   0.4668   0.3090   0.2242
Balanced Accuracy      0.6354   0.5935   0.5687
```

Lampiran 9. Output *SPSS Data Balanced Combine Sampling* (Lanjutan)

```
Call:
multiclass.roc.default(response = data$Y, predictor = predict
ed_scores)
Data: multivariate predictor predicted_scores with 3 levels o
f data$Y: 0, 1, 2.
Multi-class area under the curve: 0.6309
```

Lampiran 10. *Syntax Prepocessing* Data IFLS

```
#KEEP DATA#
keep hhid14 pidlink ar09 ar07 ar16 ar15c ar15b ar13
keep hhid14 sc05
keep hhid14 pidlink cd01
keep hhid14 pidlink rn02
keep hhid14 pidlink rj02

###MERGE DATA###


use "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC.dta", clear

merge 1:m hhid14 using "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\AR.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            89,382  (_merge==3)
    -----------------------------------------



. use "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR.dta", clear

. drop _merge

. merge 1:1 _n using "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\CD.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                       150,424
        from master                         0  (_merge==1)
        from using                    150,424  (_merge==2)

    matched                            89,382  (_merge==3)
    -----------------------------------------

. save "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD.dta", replace
file E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD.dta saved
```

Lampiran 10. Syntax Prepocessing *Data IFLS* (Lanjutan)

```
. use "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD.dta", clear

. drop _merge

. merge 1:1 _n using "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\RN.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                       231,446
        from master                   231,446  (_merge==1)
        from using                          0  (_merge==2)

    matched                             8,360  (_merge==3)
    -----------------------------------------

. save "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN.dta"
file E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN.dta saved



. use "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN.dta", clear

. drop _merge

. merge 1:1 _n using "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\RJ.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                       184,150
        from master                   184,150  (_merge==1)
        from using                          0  (_merge==2)

    matched                            55,656  (_merge==3)
    -----------------------------------------

. save "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN RJ.dta"
file E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN RJ.dta saved


. use "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN RJ.dta", clear

. drop _merge

. merge 1:1 _n using "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\AK.dta"

    Result                           # of obs.
    -----------------------------------------
    not matched                        38,415
        from master                    38,415  (_merge==1)
        from using                          0  (_merge==2)

    matched                           201,391  (_merge==3)
    -----------------------------------------

. save "E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN RJ AK.dta"
file E:\KULIAH\TUGAS AKHIR FIX\DATA\PREPROCESSING\1. PEMOTONGAN\SC AR CD RN RJ AK.dta saved
```

Lampiran 11. *Syntax* Eksplorasi Data

```
library(ggplot2)
data = DATA_TA
data$Y <- factor(data$Y)
data$X2 <- factor(data$X2)
data$X3 <- factor(data$X3)
data$X4 <- factor(data$X4)
data$X5 <- factor(data$X5)
data$X6 <- factor(data$X6)
data$X7 <- factor(data$X7)
data$X8 <- factor(data$X8)
data$X9 <- factor(data$X9)
#Boxplot X1#
boxplot(X1~Asuransi,data=dataTA_asli,
      xlab="Status Kepemilikan Asurnsi",
      ylab="Umur",
      col = c("Powder Blue","Royal blue","Lavender"))
#Boxplot X9#
boxplot(X9~Asuransi,data=dataTA_asli,
      xlab="Status Kepemilikan Asurnsi",
      ylab="Frekuensi Rawat Inap",
      col = c("Powder Blue","Royal blue","Lavender"))
#Boxplot X10#
boxplot(X10~Asuransi,data=dataTA_asli,
      xlab="Status Kepemilikan Asurnsi",
      ylab="Frekuensi Rawat Jalan",
      col = c("Powder Blue","Royal blue","Lavender"))
```

Lampiran 12. Syntax *Random Undersampling*

```
library(caret)
data <- DATA_TA
data$Y <- as.factor(data$Y)
Y <- data$Y
X <- data[, -1]
head(X)
table(Y)
down <- downSample(X,Y,FALSE)
write.csv(down, "E:/down.csv")
```

Lampiran 13. Syntax *Multinomial Logistic Regression*

```
library(foreign)
library(nnet)
library(caret)
library(e1071)
library(pROC)
data = read.csv("E:/dataTA.csv", sep = ";")
head(data)
data$Y <- factor(data$Y)
data$X2 <- factor(data$X2)
data$X3 <- factor(data$X3)
data$X4 <- factor(data$X4)
data$X5 <- factor(data$X5)
data$X6 <- factor(data$X6)
data$X7 <- factor(data$X7)
data$X8 <- factor(data$X8)
model <- multinom(Y~., data = data)
summary(model)
coef(model)
OR <- exp(coef(model))
z <- summary(model)$coefficients/summary(model)$standard.errors
p <- (1-pnorm(abs(z),0,1))*2
test <- data
predicted_scores <- predict (model, test, "probs")
predicted_class <- predict (model, test)
#confusion matrix#
confusionMatrix(predicted_class, data$Y)
#AUC#
multiclass.roc(data$Y, predicted_scores)
```

Lampiran 14. Surat Keterangan Pengambilan Data

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Departemen Statistika FMKSD ITS,

Nama : Rosikhu Ilmi Hidayati

NRP : 062115 4000 0077

menyatakan bahwa data yang digunakan dalam Tugas Akhir ini merupakan data sekunder yang diambil dari ~~penelitian / buku / Tugas Akhir / Thesis / Publikasi /~~ lainnya yaitu :

Sumber : Data dari website www.rand.org

Keterangan :

Data *Indonesia Family Life Survey (IFLS)* 5 periode september 2014 hingga maret 2015

1. Buku K (daftar anggota rumah tangga dan keterangan sampling)
2. Buku 3B (data asuransi kesehatan, kondisi kronis, rawat jalan dan rawat inap)

Surat pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Surabaya, Juli 2019

Mengetahui, Pembimbing Tugas Akhir	Mahasiswa
Dr.rer.pol. Heri Kuswanto, S.Si., M.Si	Rosikhu Ilmi Hidayati
NIP. 19820326 200312 1 004	NRP. 062115 4000 0077

(Halaman ini sengaja dikosongkan)

BIODATA PENULIS

Penulis bernama lengkap Rosikhu Ilmi Hidayati, biasa dipanggil Rosi lahir di Mojokerto pada tanggal 22 Juni 1996. Penulis merupakan anak pertama dari dua bersaudara dari seorang ayah bernama Mathudi & seorang ibu bernama Sumiati. Pendidikan formal penulis ditempuh di SDN Wringinrejo 1, SMPN 2 Kota Mojokerto, dan SMAN 1 Puri Mojokerto. Pada tahun 2015, penulis melanjutkan studi di Departemen Statistika ITS melalui jalur SBMPTN dengan beasiswa PPA selama 2 periode. Selama menempuh pendidikan formal, penulis mengikuti beberapa organisasi diantaranya adalah Staff Keilmiahan HIMASTA-ITS 2016/2017, Staff Humas Kopma Dr. Angka ITS 2016/2017, Staff Kawal PKM ITS 2017, Staff RISTEK BEM ITS 2017/2018, Asisten Dirjen RISIL RISTEK BEM ITS 2018/2019 serta Trainer Keilmiahan 2018/2019. Di bidang akademik, penulis diberi kesempatan untuk menjadi Semifinalis Elexcurtion ITB 2018, Finalis LKTIN ChEERS 2018 di Universitas Riau, Finalis LKTIN ErCOM NEON 2018 di Universitas Bengkulu, Finalis Galaksi 2018 di Universitas Negeri Surabaya, Semifinalis STC LOGIKA 2019 di Universitas Indonesia, dan Finalis *Regional Conference on Student Activism* (RECONSA) 2019 di Universiti Teknologi Petronas, Malaysia. Pencapaian penulis selama di perkuliahan yaitu Best Paper ISC ISCO 2018 di Semarang dan pendanaan penelitian dari Kemristekdikti berupa Program Kreativitas Mahasiswa bidang Sosial Humaniora (PKM PSH) 2018. Penulis juga pernah diberi kesempatan menjadi asisten dosen mata kuliah PKS serta menjadi surveyor Bappenas, serta SSH Surabaya dan Mojokerto. Bagi pembaca yang ingin berdiskusi, memberikan saran, dan kritik mengenai Tugas Akhir ini dapat disampaikan melalui email rosikhuilmi@gmail.com.