TUGAS AKHIR – SS141501

PENERAPAN *COMBINE UNDERSAMPLING* PADA KLASIFIKASI DATA *IMBALANCED* BINER (STUDI KASUS : DESA TERTINGGAL DI JAWA TIMUR TAHUN 2014)

RAHMA SHINTIA
NRP 062114 40000 032

Dosen Pembimbing
Dr. Santi Puteri Rahayu, M.Si.

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

TUGAS AKHIR – SS141501

PENERAPAN *COMBINE UNDERSAMPLING* PADA KLASIFIKASI DATA *IMBALANCED* BINER (STUDI KASUS : DESA TERTINGGAL DI JAWA TIMUR TAHUN 2014)

RAHMA SHINTIA
NRP 062114 4000 032

Dosen Pembimbing
Dr. Santi Puteri Rahayu, M.Si.

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

FINAL PROJECT – SS 141501

APPLICATION OF COMBINE UNDERSAMPLING IN IMBALANCED DATA BINARY CLASSIFICATION (CASE STUDY: UNDERDEVELOP VILLAGES IN EAST JAVA 2014)

RAHMA SHINTIA
SN 062114 4000 032

Supervisor
Dr. Santi Puteri Rahayu, M.Si.

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

LEMBAR PENGESAHAN

PENERAPAN *COMBINE UNDERSAMPLING* PADA KLASIFIKASI DATA *IMBALANCED* BINER (STUDI KASUS : DESA TERTINGGAL DI PROVINSI JAWA TIMUR TAHUN 2014)

TUGAS AKHIR

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

Oleh :
Rahma Shintia
NRP. 062114 4000 0032

Disetujui oleh Pembimbing :
Dr. Santi Puteri Rahayu, M.Si ( )
NIP. 19750115 199903 2 003

Mengetahui,
Kepala Departemen

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2018

(*Halaman ini sengaja dikosongkan*)

PENERAPAN *COMBINE UNDERSAMPLING*
PADA KLASIFIKASI DATA *IMBALANCED* BINER
(STUDI KASUS : DESA TERTINGGAL DI JAWA TIMUR TAHUN 2014)

Nama Mahasiswa : Rahma Shintia
NRP : 062114 40000 032
Departemen : Statistika
Dosen Pembimbing : Dr. Santi Puteri Rahayu

Abstrak

Regresi Logistik memiliki beberapa kelebihan dibandingkan metode klasifikasi lainnya yaitu sebagai classifier dengan akurasi yang cukup tinggi dan penggunaan algoritma yang tepat akan mampu menghasilkan waktu perhitungan yang lebih cepat khususnya pada data besar. Adanya permasalahan data yang tidak seimbang akan berpengaruh pada hasil ketepatan klasifikasi. Pada penelitian ini, metode resampling yang digunakan adalah Combine Undersampling dan metode classifier yang digunakan adalah Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel. Data yang diteliti adalah data status desa tertinggal di Jawa Timur tahun 2014 sebanyak 7.721 desa. Penerapan Combine Undersampling mampu meningkatkan ketepatan klasifikasi pada rata-rata sensitivititas secara siginifikan khususnya pada klasifikasi Regresi Logistik Ridge sebesar 42,4 kali dengan menggunakan semua variabel. Selain itu, hasil ketepatan klasifikasi terbaik menunjukkan nilai akurasi total dan AUC yang sama ketika menerapkan metode Combine Undersampling pada Klasifikasi Analisis Diskriminan Kernel yaitu 78.0 % sedangkan pada variabel signifikan metode Regresi Logistik Ridge menghasilkan ketepatan klasifikasi lebih baik dari metode lainnya yang memiliki nilai AUC sebesar 73,5% .

Kata Kunci : ***Combine Undersampling, Data Imbalanced, Desa Tertinggal***

(*Halaman ini sengaja dikosongkan*)

APPLICATION OF COMBINE UNDERSAMPLING IN IMBALANCED DATA BINARY CLASSIFICATION (CASE STUDY: UNDERDEVELOP VILLAGES IN EAST JAVA 2014)

Name : Rahma Shintia
Student Number : 062114 40000 032
Department : Statistics
Supervisor : Dr. Santi Puteri Rahayu

Abstract

Logistic regression has several advantages over other classification methods that is as a classifier with a fairly high accuracy and the use of appropriate algorithms will be able to produce faster calculation times, especially on large data. The existence of immbalanced data problems will affect the results of classification accuracy. In this research, the resampling method used is Combine Undersampling and the classifier method used are Logistic Regression, Ridge Logistic Regression, and Kernel Discriminant Analysis. The data studied is the status data of underdevelop villages in East Java in 2014 as many as 7,721 villages. The application of Combine Undersampling is able to increase the classification accuracy on the average sensitivity significantly in the Ridge Logistic Regression classification by 42.4 times using all the variables. In addition, the best classification accuracy results show the same total accuracy and AUC value when applying Combine Undersampling method in the Kernel Discriminant Classification Classification is 78.0% whereas in the significant variables the Ridge Logistic Regression method produces better classification accuracy than other methods which have AUC value of 73 , 5%.

Keywords: Combine - Undersampling, Imbalanced Data, Kernel Underdevelop Village

(*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 **"Penerapan *Combine Undersampling* pada Klasifikasi Data *Imbalanced* Biner (Studi Kasus : Desa Tertinggal di Jawa Timur Tahun 2014)"** 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. Bapak Dr. Suhartono selaku Kepala Departemen Statistika ITS dan Bapak Dr. Sutikno, M.Si selaku Ketua Program Studi Sarjana Departemen Statistika ITS yang telah menyediakan fasilitas guna kelancaran pengerjaan Tugas Akhir ini.
2. Santi Puteri Rahayu, Ph.D selaku dosen pembimbing Tugas Akhir yang telah meluangkan waktu dan dengan sangat sabar memberikan bimbingan, saran, dukungan serta motivasi selama penyusunan Tugas Akhir.
3. Dr. Ir. Setiawan, M.S. dan Dr. rer. pol. Dedy Dwi Prastyo, S.Si., M.Si. selaku dosen penguji yang telah banyak memberi masukan kepada penulis.
4. Dr. Muhammad Masuri, M.T selaku dosen wali yang telah banyak memberikan saran dan arahan dalam proses belajar di Departemen Statistika.
5. Keluarga penulis atas segala do'a, nasehat, kasih sayang, dan dukungan yang diberikan kepada penulis demi kesuksesan dan kebahagiaan penulis.
6. Teman-teman seperjuangan Tugas Akhir, khususnya Dewi Lutfia Pratiwi dan Canggih Shoffi Imanwardhani yang selama ini telah berjuang bersama dan saling memberikan semangat.
7. Teman-teman Statistika ITS angkatan 2014, Respect, yang selalu memberikan dukungan kepada penulis selama ini.
8. Semua pihak yang turut membantu dalam pelaksanaan Tugas Akhir yang tidak bisa penulis sebutkan satu persatu.

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

Surabaya, Juli 2018

Penulis

DAFTAR ISI

Halaman

(*Halaman ini sengaja dikosongkan*)

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

(*Halaman ini sengaja dikosongkan*)

BAB I

PENDAHULUAN

1.1 Latar Belakang

Data mining (DM) merupakan suatu pendekatan sistematis untuk menemukan sebuah pola dan tren dalam data. Metodologi yang digunakan terdiri dari visualisasi data, *machine learning*, dan teknik statistik (Curt, 1995). Aplikasi terkait yang menggunakan metodologi ini terdiri dari klasifikasi (*classification*), prediksi (*forecasting*), dan pengelompokkan (*clustering*) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Klasifikasi merupakan sebuah metode untuk menyusun data secara sistematis atau menurut beberapa aturan/kaidah yang telah ditetapkan. Secara harafiah bisa pula dikatakan bahwa klasifikasi adalah pembagian sesuatu menurut kelas-kelas (Azminuddin, Suhartono, & Himawan, 2017). Beberapa metode/algoritma yang umum dikenal dalam *classification* adalah *Linear Discrimination Analysis* (*LDA*), *Artificial Neural Netwowk* (*ANN*), *Fuzzy Clustering*, *Support Vector Machines* (*SVM*), Regresi Logistik, dan lain-lain. Regresi Logistik merupakan metode statistik yang bertujuan untuk mengetahui hubungan variabel respon (kategorik) dengan satu atau lebih variabel prediktor yang bersifat kategorik/kontinu (Agresti, 2002). Ada beberapa kelebihan klasifikasi Regresi Logistik dibandingkan metode klasifikasi lainnya yang telah dipelajari. Regresi Logistik memberikan probabilitas dan mencakup masalah klasifikasi multi kelas sehingga disebut sebagai *classifier* yang tangguh (Hastie, Tibshirani, & Friedman, 2001). Regresi Logistik hanya memerlukan pemecahan masalah *unconstraid optimization* sehingga penggunaan algoritma yang tepat akan mampu menghasilkan waktu perhitungan yang lebih cepat dibandingkan metode lain, seperti *Support Vector Machine* (Maalouf & Siddiqi, 2014). Kelebihan menggunakan Regresi Logistik adalah metode ini telah dipelajari secara ektensif dan telah dikembangkan menggunakan metode *Truncated Newton* (Maalouf & Trafalis, 2010). Pengembangan metode tersebut, menunjukkan bahwa

Regresi Logistik lebih baik dari *Support Vector Machine* karena sangat efektif sebagai *classifier* untuk mengoptimasi data besar dan ketepatan klasifikasi dari kedua metode ini *comparable* atau tidak jauh berbeda (Maalouf, 2011).

Adanya pengembangan dari metode Regresi Logistik mampu menangani permasalahan yang mengurangi efektifitas sebagai *classifier*. Salah satu pengembangan dari Regresi Logistik adalah adalah Regresi Logistik Ridge. Regresi Logistik Ridge yaitu suatu metode Regresi Logistik yang dikembangkan untuk mengakomodasi permasalahan korelasi tinggi antara beberapa variabel bebas atau multikolinearitas (Draper & Smith, Applied Regression Analysis, 1998). Namun, metode ini kurang tepat digunakan pada data *imbalanced* dan data *rare event* karena akan menghasilkan estimasi parameter yang bias dan standar *error* yang buruk (Agresti, 2007). Penelitian sebelumnya dilakukan oleh Sunyoto, Setiawan & Zain (2009) yaitu Regresi Logistik Ridge pada keberhasilan murid SMA Negeri 1 Kediri yang diterima di Perguruan Tinggi. Selain itu, Putra & Ratnasari (2015) melakukan pemodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur dengan menggunakan metode Regresi Logistik Ridge, dimana ditemukan kasus multikolinearitas pada faktor-faktor yang mempengaruhi IPM di Jawa Timur dan diatasi dengan menggunakan metode *backward elimination*, didapatkan ketepatan klasifikasi sebesar 97.73% dan terdiri dari 6 kabupaten/kota tergolong IPM menengah bawah dan 33 lainnya tergolong IPM menengah atas. Liu, Kawamoto, Morita, Yoshinari, & Honda (2017) melakukan penelitian menggunakan model Regresi Logistik Ridge untuk mengidentifikasi gen-gen yang bertanggung jawab terhadap hipertrofi hati dan hepatokarsinogenesis hipertrofik. Hasil penelitian menunjukkan akurasi yang didapatkan dari model masing-masing gen sebesar 94.8% dan 82.7%.

Salah satu permasalahan klasifikasi data adalah komposisi data yang tidak seimbang. Dengan perbandingan data yang tidak seimbang maka akan mempengaruhi metode klasifikasi dari data mining untuk memprediksi kelas data minoritas, akurasi tinggi

sering dicapai dalam klasifikasi tanpa penanganan data *imbalanced* karena hanya berfokus pada data mayoritas dan untuk data minoritas dianggap sebagai data langka atau data tidak sengaja (Haibo, Member, & Edwardo, 2009). Dengan adanya penanganan menggunakan sampling terhadap data *imbalanced*, tingkat data yang tidak seimbang akan semakin kecil dan klasifikasi dapat dilakukan dengan tepat. *Sampling-based approaches* merupakan pendekatan sampling dengan memodifikasi distribusi dari data *training* sehingga kedua kelas data (negatif maupun positif) dapat dipresentasikan dengan baik dalam data *training*. Sampling terdiri dari dua jenis yaitu *undersampling* dan *oversampling*. Metode *undersampling* dilakukan dengan cara mengurangi jumlah data kelas mayor sedangkan *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan jumlah data kelas minor. Namun, masalah yang muncul dari metode *oversampling* adalah masalah *over-fitting* secara umum akan terjadi, yang menyebabkan aturan klasifikasi menjadi semakin spesifik meskipun akurasi untuk data *training* semakin membaik, sedangkan kelebihan *undersampling* adalah penghapusan beberapa data dapat secara signifikan mengurangi ukuran data sehingga dapat menurunkan biaya *run-time* terutama dalam kasus data yang besar (Elhassan T. , Aljurf, Al-Mohanna, & Shoukri, 2016).

Salah satu metode penghapusan data *noise* untuk menangani data *imbalanced* adalah *Tomek Links*, metode ini diperkenalkan oleh *Tomek* (1997). Cara kerja *Tomek Links* adalah menghapus data kelas negatif ataupun positif yang memiliki kesamaan karakteristik. Park & Sklansky (1990) melakukan penelitian untuk merancang linear *tree classifier* dengan menggunakan *Tomek Links* untuk mengatasi data *imbalanced* dan menghasilkan nilai akurasi yang sama dengan klasifikasi k-NN. Penelitian lainnya adalah pembersihan data dari data *noise* atau tidak konsisten dan menunjukkan bahwa *Tomek Links* merupakan pola yang baik untuk pembersihan data pada kasus tersebut (Jeatrakul, Wong, & Fung, 2010). *Tomek Links* mampu meningkatkan ketelitian pada

klasifikasi *Decision Tree* dan kinerja pada klasifikasi *Support Vector Machine* pada kasus pengenalan simpangan sambatan dalam barisan DNA (Lorena, Batista, Carvalho, & Monard, 2002). Selain itu, Sain dan Purnami (2015) menyatakan pada penelitiannya bahwa pada beberapa kasus data *imbalanced* yang sangat ekstrim yaitu prosentase kelas minoritas kurang dari 10%, metode *Tomek Links* lebih baik dari *Combine Sampling* atau pun SMOTE. Devi, Biswas, & Purkayastha (2017) menunjukkan bahwa transfusi darah memiliki akurasi rata-rata yang bagus menggunakan metode *Tomek Links Naïve Bayes* untuk mengatasi data *imbalanced*.

Setelah melakukan penghapusan data dengan metode *Tomek Linkss*, penelitian ini menggunakan metode *Random Undersampling* untuk menyeimbangkan data. *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 (Yu, Zhou, Chen, Deng, & Wang, 2017). RUS memiliki waktu prediksi yang lebih cepat dibandingkan ROS dan SMOTE (Park, Oh, & Pedrycz, 2013). Pengaplikasian RUS pada algoritma prediksi cacat *software* bisa mengurangi pengeluaran biaya (Arar & Ayan, 2015). Metode RUS lebih mudah dan cepat diaplikasikan dibandingkan metode yang lainnya (Yu, et al., 2013). Penelitian yang dilakukan oleh Yu, Tang, Shen, Yang, dan Yang, J, melakukan penelitian untuk meningkatkan akurasi prediksi cacat *software* mengkombinasikan SVM dengan cara modifikasi *Adaboost* dengan *random undersampling* (RUS). SVM memiliki kelemahan didalam mengukur kelas minoritas dan penentuan keputusan final dataset *training*, RUS memiliki efektifitas yang tinggi untuk mengatasi masalah pada kelas mayoritas dan minoritas pada SVM. Hasilnya menunjukan AUC terendah 0.75 dan terbesarnya 0.86. Penelitian lainnya dilakukan oleh Elhassan, Aljurf, Al-Mohanna, & Shoukri (2016) dengan mengkombinasikan metode *Tomek Links* dan *Random Undersampling*, hasilnya

menunjukkan bahwa adanya peningkatan sensitivitas (akurasi pada kelas minoritas), *G-mean* dan F-statistik.

Kasus desa tertinggal di Provinsi Jawa Timur merupakan salah satu kasus *real* dari data *imbalanced* yang menarik untuk dijadikan topik penelitian. Berdasarkan data Sensus Penduduk BPS (2010), Provinsi Jawa Timur ada pada urutan kedua dengan jumlah penduduk terbanyak, yaitu 37.476.757 jiwa dan terdiri dari 5.674 desa. Selain itu, provinsi Jawa Timur merupakan satu-satunya provinsi yang masih memiliki beberapa wilayah tertinggal di Pulau Jawa. Data Indeks Pembangunan Desa Tahun 2014 Pulau Jawa-Bali menunjukkan bahwa persentase desa tertinggal di Jawa Timur sebesar 2,69% dan desa tidak tertinggal (berkembang & mandiri) sebesar 97.31 %. Hal ini mengindikasikan bahwa data desa tertinggal merupakan data *imbalanced* dengan rasio yang cukup tinggi yaitu 1 : 36,17 dengan data minoritas adalah desa tertinggal dan mayoritas desa tidak tertinggal.

Desa tertinggal adalah adalah desa yang belum memenuhi Standar Pelayanan Minimal Desa (SPM Desa) pada aspek kebutuhan sosial dasar, infrastruktur dasar, sarana dasar, pelayanan umum, dan penyelenggaraan pemerintahan (Bappenas, 2015). Pemerataan pembangunan dibidang infrastruktur maupun suprastruktur tidak merata disetiap desa yang tertuang dalam Indeks Desa Membangun (IDM). IDM dikembangkan sebagai tolok ukur untuk memperkuat upaya pencapaian sasaran pembangunan desa dan kawasan perdesaan sebagaimana yang tertuang dalam Buku Rencana Pembangunan Jangka Menengah Nasional 2015-2019, yakni mengurangi jumlah desa tertinggal sampai dengan 500 desa dan meningkatkan jumlah desa mandiri sedikitnya 2.000 desa pada tahun 2019 (Sukrajap & Harahap, 2017). Adapun kriteria dan indikator penetapan daerah tertinggal berdasarkan Pemendesa PDTT No.3 Tahun 2016 tentang Petunjuk Teknis Penentuan Indikator Dalam Penerapan Daerah Tertinggal Secara Nasional, yaitu perekonomian masyarakat, sumber daya manusia, sarana dan prasarana, kemampuan keuangan daerah, aksebilitas, dan karakteristik wilayah. Beberapa penelitian yang

telah dilakukan mengenai identifikasi desa tertinggal adalah penggunaan *Geographically Weighted Regression-Kriging* untuk klasifikasi desa tertinggal (Dimulyo, 2009). Sambodo, Purnami, dan Rahayu (2014) menganalisis ketepatan klasifikasi status ketertinggalan desa dengan pendekatan *Reduce Support Vector Machine* (*RSVM*) di Provinsi Jawa Timur sebanyak 8.502 desa dengan rasio data, yaitu 1 : 1 (desa tertinggal: desa tidak tertinggal) dan menghasilkan akurasi 71,65%. Penelitian lainnya adalah penggunaan *Geographically Weighted Regression* (*GWR*) dengan pembobot *Gauss kernel* untuk klasifikasi desa miskin (Rahmawati, Djuraidah, & Aidi, 2010), sedangkan Sulasih, Purnami & Rahayu (2016) melakukan penelitian tentang *Rare Event Weighted Logistic Regression* untuk klasifikasi *imbalanced data* (Studi Kasus : klasifikasi desa tertinggal di provinsi Jawa Timur) yang terdiri dari 7.721 desa dan rasio data sebesar 1:36 (desa tertinggal : desa tidak tertinggal), serta menghasilkan rata-rata akurasi 98,06% (λ=2).

Pada penelitian ini, efektifitas dari metode klasifikasi Regresi Logistik Ridge akan dibandingkan dengan dengan metode lainnya, yaitu klasifikasi Analisis Diskriminan Kernel. Analisis Diskriminan Kernel merupakan suatu metode yang fleksibel karena tidak harus memenuhi asumsi tertentu pada analisis diskriminan parametrik, yaitu asumsi normal multivariat dan matriks ragam peragam yang homogen (Hardle, 1990). Metode kernel yang dikemukakan oleh Rosenblatt (1956) digunakan untuk menduga fungsi kepadatan peluang secara nonparametrik. Kelebihan penduga kernel adalah bentuknya lebih fleksibel dan bentuk matematis dari penduganya lebih mudah disesuaikan. Akan tetapi, kesulitan metode ini terutama untuk data peubah ganda adalah menentukan penduga parameter penghalus. Penelitian sebelumnya dilakukan oleh Aitchison dan Aiken (1976) menggunakan metode kernel pada Analisis Diskriminan Kernel data biner. Analisis Diskriminan Kernel memberikan hasil yang bagus untuk pengenalan pola, seperti Li, Gong, & Liddell (2003) melakukan identifikasi wajah dengan Analisis Diskriminan Kernel. Djuraidah dan Aunuddin (2004) meneliti pengelompokkan warna

menggunakan Analisis Diskriminan Kernel dan memberikan kesalahan klasifikasi 0%. Penelitiannya lainnya dilakukan Meilianawati, Sumarminingsih, & Wardhani (2013) menyatakan bahwa pendekatan Analisis Diskriminan Kernel mengklasifikasikan data dengan respon ordinal lebih tepat daripada pendekatan model *proportional odd*. Fungsi kernel yang digunakan pada penelitian ini adalah fungsi kernel *Gaussian* dikarenakan lebih halus dibandingkan fungsi Kernel lain (Seber, 1984). Menurut Hsu, Chang, & Lin (2004), fungsi kernel yang direkomendasikan untuk diuji pertama kali adalah fungsi kernel RBF karena dapat memetakan hubungan tidak linier RBF lebih robust terhadap outlier. Hal ini dikarenakan fungsi kernel RBF berada antara selang $-\infty, \infty$ sedangkan fungsi kerenl yang lain memiliki rentang antara (-1 sampai dengan 1). Penelitian yang dilakukan oleh Yang, Jin, Yang, Zhang, & Frangi (2004) menggunakan dua kernel untuk membandingkan *recognition rate*, dimana kernel *Gaussian RBF* memiliki *rate* yang lebih tinggi dari hasil kernel *Polynomial*.

Pada penelitian ini, pengklasifikasian status data desa tertinggal menggunakan indikator-indikator Indeks Pembangunan Desa menurut data Potensi Desa Badan Pusat Statistik Tahun 2014 yang terdiri dari 8 variabel berskala numerik dengan menggunakan *Combine Undersampling* pada Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel untuk menangani data *imbalanced*. Tujuan dari penelitian ini adalah membandingkan efektifitas dari kedua metode *classifier* yang menghasilkan ketepatan klasifikasi terbaik untuk data desa tertinggal di Jawa Timur tahun 2014. Hasil dari dari penelitian, diharapkan menghasilkan akurasi data yang tinggi, sehingga dapat memberikan informasi agar kebijakan atau keputusan yang ditetapkan efektif, efisien, dan tepat sasaran.

1.2 Rumusan Masalah

Kondisi data *imbalanced* merupakan salah satu masalah dalam klasifikasi karena *classifier* akan memprediksi ke kelas data *training* yang banyak (mayoritas) dibandingkan ke kelas minoritas.

Hal itu berdampak pada hasil akurasi prediksi yang baik terhadap kelas data data *training* mayoritas sedangkan untuk kelas data *training* minoritas akan menghasilkan akurasi prediksi yang buruk. Selain itu, variabel-variabel data desa tertinggal di Jawa Timur darah diduga memiliki hubungan yang kuat satu sama lainnya sehingga mengakibatkan adanya kasus multikolinearitas. Apabila kasus multikolinearitas tidak diatasi, maka variansi dari hasil estimasi akan berdampak pada besarnya nilai *error* dan interval kepercayaan menjadi lebar. Efektifitas terbaik dari *classifier* dapat diketahui dengan membandingkan performansi dari ketiga metode klasifikasi yang digunakan. Berdasarkan uraian diatas, permasalahan dalam penelitian ini adalah bagaimana hasil ketepatan klasifikasi terbaik dalam menangani data *imbalanced* menggunakan Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel.

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut :

1. Mendeskripsikan karakteristik dari data desa tertinggal di Jawa Timur tahun 2014.
2. Mengetahui ketepatan klasifikasi Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel sebagai teknik pemodelan berbasis *machine learning* pada data *imbalanced*.
3. Membandingkan tingkat ketepatan klasifikasi untuk status ketertinggalan desa di Provinsi Jawa Timur tahun 2014 melalui pendekatan Combine Undersampling pada klasifikasi Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel.
4. Mengetahui efektivitas hasil penanganan data *imbalanced* menggunakan pendekatan metode *Combine Undersampling (Tomek Links + Random Undersampling)* pada klasifikasi Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel.

1.4 Manfaat

Hasil penelitian ini diharapkan dapat memberikan tambahan informasi dan pengetahuan dalam penanganan data *imbalanced* menggunakan *Combine Undersampling* dengan metode klasifikasi Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel. Selain itu, hasil penelitian dari klasifikasi status desa tertinggal di Jawa Timur dapat dijadikan sebagai tambahan informasi bagi pemerintah dalam menetapkan kebijakan pembangunan desa terutama untuk memprediksi desa-desa di Jawa Timur yang mengalami pemekaran sehingga kebijakan yang ditetapkan efektif dan tepat sasaran.

1.5 Batasan Masalah

Data yang diolah pada penelitian ini menggunakan 8 variabel berskala numerik dan terdeteksi adanya kasus multikolinearitas pada data *balanced*. Metode *classifier* yang digunakan terdiri dari Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel. Fungsi kernel yang digunakan pada Analisis Diskriminan Kernel adalah fungsi kernel *Gaussian RBF* sedangkan penentuan Ridge parameter pada klasifikasi Regresi Logistik Rigde dipilih secara *default* menggunakan *software R*.

(Halaman ini sengaja dikosongkan)

BAB II

TINJAUAN PUSTAKA

Dalam bab dua ini akan dijelaskan mengenai landasan teori yang digunakan pada penelitian, terdiri dari metode *Tomek Links*, *Random Undersampling*, Regresi Logistik Ridge, Analisis Diskriminan Kernel, Ketepatan Klasifikasi, dan Desa Tertinggal di Jawa Timur tahun 2014.

2.1 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 dan Solberg, 1996). *Sampling-based approaches* yaitu memodifikasi distribusi dari data *training* sehingga kedua kelas data (negatif maupun positif) dipresentasikan dengan baik di dalam data *training*. *Sampling* sendiri dibedakan menjadi 2 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 *Tomek Links* dan *Random Undersampling*, sebagai berikut :

2.1.1 *Tomek Links*

Dengan adanya penerapan *sampling* pada data yang *imbalanced*, tingkat data *imbalanced* semakin kecil dan klasifikasi dapat dilakukan dengan tepat (Solberg & Solberg , 1996). *Sampling* dibedakan menjadi 2 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.

Tomek Links dapat didefinisikan sebagai berikut, diberikan dua sampel x dan y milik kelas yang berbeda, dan $d(x,y)$ adalah jarak x dan y. Sepasang (x,y) disebut *Tomek Links* jika tidak ada sampel z, sehingga $d(x,z) < d(x,y)$ atau $d(y,z) < d(y,x)$ (Batista, Bazzan , & Monard, 2003). Jika dua sampel membentuk *Tomek Linkss*, maka salah satu dari kedua sampel adalah data *noise* atau kedua contoh adalah *borderline*. *Tomek Links* dapat digunakan sebagai metode *undersampling*, dimana hanya sampel dari kelas negatif yang akan deliminasi atau sebagai metode pembersihan yaitu kedua sampel dari kedua kelas yang berbeda akan dihapus. Sistem kerja *Tomek Links* akan berhenti ketika tidak ada lagi data *noise* atau data *borderline* pada data kelas mayoritas. Contoh penggunaan metode *Tomek Links* pada subbab selanjutnya (Sain, 2013) .

Tabel 2.1 Data Ilustrasi untuk Metode Tomek Links

No.	X_1	X_2	Y	No.	X_1	X_2	Y
1	2	2	0	10	4	4	1
2	3	6	0	11	5	1	1
3	4	2	0	12	5	3	1
4	6	5	0	13	5	6	1
5	1	2	1	14	6	2	1
6	1	4	1	15	6	4	1
7	3	1	1	16	2	3	0
8	3	3	1	17	2,5	2,5	0
9	3	4	1	18	2	1,5	0

Nilai $Y = 0$ merupakan sampel dari kelas negatif dan $Y = 1$ merupakan sampel dari kelas positif sehingga contoh hasil penerapannya sebagai berikut :

Rumus yang digunakan adalah jarak Euclidian pada persamaan (2.1).

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (2.1)$$

1. Misalkan $y_5 = (1,2)$ dan $y_1 = (2,2)$ dideteksi bukan merupakan kasus *Tomek Links*.
   $d(y_5, y_1) = \sqrt{(1-2)^2 + (2-2)^2 = 1}$
   Titik lain yang terdeteksi berada dekat dengan antara kedua titik y_5 dan y_1 adalah $y_{18} = \big(2, (1,5)\big)$ sehingga
   $d(y_5, y_{18}) = \sqrt{(1-2)^2 + (2-1,5)^2 = 1,12}$
   $d(y_1, y_{18}) = \sqrt{(2-2)^2 + (2-1,5)^2 = 0,5}$
   Kesimpulannya $d(y_5, y_{18}) = 1,12 > d(y_5, y_1) = 1$ atau $d(y_1, y_{18}) = 0,5 > d(y_5, y_1) = 1$, dengan demikian kedua titik y_5 dan y_1 bukan merupakan kasus *Tomek Links* karena memenuhi syarat dari definisi kasus *Tomek Links*.
2. Misalkan $y_8 = (3,3)$ dan $y_{17} = \big((2,5), (2,5)\big)$ dideteksi merupakan kasus *Tomek Links*.
   $d(y_8, y_{17}) = \sqrt{(3-2,5)^2 + (3-2,5)^2 = 0,707}$
   Titik lain yang terdeteksi berada dekat dengan antara kedua titik y_8 dan y_{17} adalah $y_9 = (3,4)$ sehingga
   $d(y_8, y_9) = \sqrt{(3-3)^2 + (3-4)^2 = 1}$
   $d(y_{17}, y_9) = \sqrt{(2,5-3)^2 + (2,5-4)^2 = 1,58}$
   Kesimpulannya $d(y_8, y_9) = 1 > d(y_8, y_{17}) = 0,707$ atau $d(y_{17}, y_9) = 1,58 > d(y_8, y_{17}) = 0,707$, dengan demikian kedua titik y_8 dan y_{17} merupakan kasus *Tomek Links* karena tidak memenuhi syarat dari definisi kasus *Tomek Links* sehingga titik $y_8 = (3,3)$ akan dihapus.

2.1.2 *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 (Yu, Zhou, Chen, Deng, & Wang, 2017).

2.2 Multikolinearitas

Deteksi multikolinieritas digunakan untuk menunjukkan adanya hubungan linier diantara variabel independen. Multikolinieritas disini dimaksudkan untuk menunjukkan derajat kolinieritas yang tinggi diantara variabel independen. Model yang baik seharusnya tidak terdapat korelasi yang tinggi diantara variabel independen. Deteksi multikolinieritas dilakukan menggunakan kriteria nilai *VIF* (*Variance Inflation Factor*). Apabila nilai *VIF* lebih dari 5, maka dapat diindikasikan terdapat kasus multikolinieritas. Selain itu, jika nilai *VIF* lebih besar dari 10, maka hal tersebut mengindikasikan bahwa variabel prediktor memiliki kasus multikolinieritas yang serius dan harus diatasi. Nilai *VIF* dinyatakan sebagai berikut (Hocking, 2003) :

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1,2, \dots, p \tag{2.2}$$

dengan R_j^2 adalah koefisien determinasi antara satu variabel independen X_j dengan variabel independen lainnya. R_j^2 dapat dinyatakan sebagai berikut :

$$R_j^2 = 1 - \frac{SSE}{SST} \tag{2.3}$$

dengan $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ dan $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Hal – hal yang akan terjadi apabila kasus multikolinieritas tidak diatasi adalah sebagai berikut :

1. Variansi estimasi menjadi besar.
2. Interval kepercayaan menjadi lebar, dikarenakan variansi dan standar *error* besar.
3. Pengujian signifikansi secara parsial menjadi tidak signifikan.
4. Koefisien determinasi R^2 tinggi, tetapi tidak banyak variabel prediktor yang signifikan.

2.3 Regresi Logistik

Regresi Logistik merupakan salah satu metode *dichotomous* (berskala nominal atau ordinal dengan dua kategori) atau *polychotomous* (mempunyai skala nominal atau ordinal dengan lebih dari dua kategori) dengan satu atau lebih variabel prediktor. Variabel respon pada Regresi Logistik bersifat kontinyu atau kategorik (Agresti, 1990).

Regresi Logistik biner merupakan salah satu metode regresi yang menggambarkan hubungan antar variabel respon (*outcome* atau *dependent*) dengan satu atau lebih variabel prediktor (*explanatory* atau *independent*). Regresi Logistik biner hanya digunakan untuk kasus khusus yaitu apabila variabel respon (Y) merupakan variabel kualitatif yang bersifat biner atau dikotumus. Variabel dikotumus adalah variabel yang hanya mempunyai dua kemungkinan nilai, misalnya sukses dan gagal.

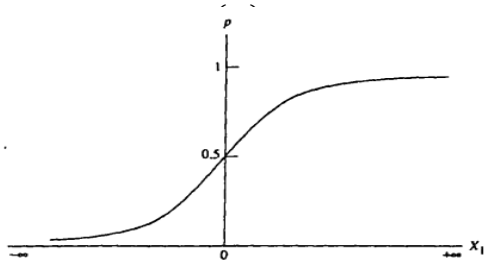
Menurut Agresti (2007) untuk memudahkan, maka variabel respon/dependen diberi notasi *Y* dan varibel prediktor/independen dinotasikan *X*. Apabila *Y* menghasilkan dua kategori, misalnya "1" jika sukses dan "0" jika gagal. Variabel respon *Y* berdistribusi Binomial dengan parameter π_i, dimana untuk setiap pengamatan ke-*i* ditulis pada persamaan (2.4).

$$y_i \sim \text{Binomial}\,(1, \pi_i)\,, \tag{2.4}$$

dengan fungsi probabilitas seperti dibawah ini :

$$f(y_i) = (\pi(\mathbf{x}_i))^{y_i}\,(1-\pi(\mathbf{x}_i))^{1-y_i}\,,\ y_i = 0,1 \tag{2.5}$$

dimana $\pi_i = \pi(\mathbf{x}_i)$ adalah probabilitas dari kejadian ke-i. Apabila $y_i = 1$, maka $f(y_i) = \pi(\mathbf{x}_i)$ dan apabila $y_i = 0$, maka $f(y_i) = 1 - \pi(\mathbf{x}_i)$. Dalam regresi logistik, hubungan antara variabel prediktor dan variabel respon bukanlah suatu fungsi linier. Gambar 2.1 merupakan visualisasi untuk *logistic curve* yang menjadi dasar Regresi Logistik.



Gambar 2.1 *Logistic Curve* (Sharma, 1996)

Bentuk persamaan Regresi Logistik ditunjukkan pada persamaan (2.5)

$$\pi(\mathbf{x}_i) = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)}$$
$$\pi(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i{}^T \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i{}^T \boldsymbol{\beta}\right)} \quad (2.6)$$

dengan β_0 = konstanta, β_j = koefisien regresi dan j = banyaknya variabel prediktor. Probabilitas ini dapat digunakan untuk mengklasifikasikan pengamatan ke dalam dua kelompok. Klasifikasi pengamatan ke dalam kelompok didasarkan pada nilai *cutoff* yang biasanya diasumsikan 0,5. Semua pengamatan yang lebih besar dari atau sama dengan 0,5 diklasifikasikan sebagai kelompok sukses dan yang nilainya kurang dari 0,5 diklasifikasikan sebagai yang gagal (Sharma, 1996). Menurut Yan & Su (2009), terdapat suatu bentuk alternatif dari persamaan Regresi Logistik seperti persamaan (2.7) yang merupakan tranformasi logit dari $\pi(\mathbf{x}_i)$.

$$\text{Logit}\left[\pi(\mathbf{x}_i)\right] = \ln\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right] \quad (2.7)$$

$$\text{Logit}\left[\pi\left(\mathbf{x}_i\right)\right]=\ln\left[\frac{\dfrac{\exp\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}\right)}{1+\exp\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}\right)}}{1-\dfrac{\exp\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}\right)}{1+\exp\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}\right)}}\right]$$

$$\text{Logit}\left[\pi\left(\mathbf{x}_i\right)\right]=\ln\left[\frac{\dfrac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}{1-\dfrac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}\right]=\ln\left[\frac{\dfrac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}{\dfrac{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)-\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}\right]$$

$$\text{Logit}\left[\pi\left(\mathbf{x}_i\right)\right]=\ln\left[\frac{\dfrac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}{\dfrac{1}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}}\right]=\ln\left[\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\right]$$

$$\text{Logit}\left[\pi\left(\mathbf{x}_i\right)\right]=\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}. \qquad (2.8)$$

Menurut Agresti (2007) dalam menaksir parameter dalam model Regresi Logistik digunakan metode *Maximum Likelihood Estimator* (*MLE*). Metode *MLE* digunakan karena distribusi dari variabel respon telah diketahui. *MLE* didapatkan dengan cara memaksimumkan logaritma fungsi *likelihood*. Dari persamaan (2.8) didapatkan fungsi *likelihood* pada persamaan (2.9).

$$L\left(\mathbf{X},\boldsymbol{\beta}\right)=\prod_{i=1}^{n}\left[\pi\left(\mathbf{x}_i\right)\right]^{y_i}\left[1-\pi\left(\mathbf{x}_i\right)\right]^{1-y_i}$$

$$\ln L\left(\mathbf{X},\boldsymbol{\beta}\right)=l\left(\mathbf{X},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}y_i\ln\left[\pi\left(\mathbf{x}_i\right)\right]+\sum_{i=1}^{n}\left(1-y_i\right)\ln\left[1-\pi\left(\mathbf{x}_i\right)\right]$$

$$\ln L\left(\mathbf{X},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}y_i\ln\left[\pi\left(\mathbf{x}_i\right)\right]+\sum_{i=1}^{n}\left(\ln\left[1-\pi\left(\mathbf{x}_i\right)\right]-y_i\ln\left[1-\pi\left(\mathbf{x}_i\right)\right]\right)$$

$$\ln L\left(\mathbf{X},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}y_i\ln\left[\pi\left(\mathbf{x}_i\right)\right]-\sum_{i=1}^{n}y_i\ln\left[1-\pi\left(\mathbf{x}_i\right)\right]+\sum_{i=1}^{n}\ln\left[1-\pi\left(\mathbf{x}_i\right)\right]$$

$$\ln L(\mathbf{X},\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \left(\ln\left[\pi(\mathbf{x}_i)\right] - \ln\left[1-\pi(\mathbf{x}_i)\right] \right) + \sum_{i=1}^{n} \ln\left[1 + \frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right]$$

$$\ln L(\mathbf{X},\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} \right] + \sum_{i=1}^{n} \ln\left[\frac{1}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right]$$

$$\ln L(\mathbf{X},\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right) + \ln\left(1+\exp \mathbf{x}_i{}^T\boldsymbol{\beta}\right)^{-1} \right]$$

$$\ln L(\mathbf{X},\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right) - \ln\left(1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\right) \right]. \qquad (2.9)$$

Menurut Ryan (1997), Berdasarkan persamaan (2.9) dilakukan penurunan terhadap β menjadi persamaan (2.10).

$$\frac{\partial \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i - \frac{\mathbf{x}_i \exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right] = \sum_{i=1}^{n} \mathbf{x}_i \left[y_i - \frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right]$$

$$\frac{\partial \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \left[y_i - \pi(\mathbf{x}_i) \right] \qquad (2.10)$$

$$\frac{\partial \ln(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{y}\text{-}\hat{\boldsymbol{\pi}})$$

dimana $\boldsymbol{y} =$ vektor pengamatan pada variabel respon yang berukuran $n \times 1$, sedangan $\mathbf{X} =$ matriks variabel prediktor yang berukuran $n \times (p+1)$.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \qquad (2.11)$$

Tahapan selanjutnya adalah mendapatkan nilai $\boldsymbol{\beta}$ yang menyebabkan nilai dari fungsi *likelihood* bernilai ekstrim, dengan cara mencari turununan pertama $L(\mathbf{X},\boldsymbol{\beta})$ terhadap $\boldsymbol{\beta}$. Menurut Ryan (1997), hasil penurunan kedua ditunjukkan pada persamaan (2.12).

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \frac{\partial}{\partial \boldsymbol{\beta}^T} \left(\frac{\partial \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) = \frac{\partial}{\partial \boldsymbol{\beta}^T} \left(\sum_{i=1}^{n} \left[y_i \mathbf{x}_i - \frac{\mathbf{x}_i \exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right] \right)$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = 0 - \sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}_i{}^T \exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\left[1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\right] - \mathbf{x}_i\mathbf{x}_i{}^T\left[\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\right]^2}{\left[1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)\right]^2}$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}^T \left(\frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} - \left[\frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}\right]^2 \right)$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}^T \left(\frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right)\left(1 - \frac{\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i{}^T\boldsymbol{\beta}\right)} \right)$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i{}^T \pi(\mathbf{x}_i)\left(1-\pi(\mathbf{x}_i)\right)$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \mathbf{X}^T\mathbf{W}\mathbf{X} \quad . \tag{2.12}$$

Apabila dilakukan ekspansi berdasarkan Deret Taylor disekitar nilai β, maka didapatkan persamaan (2.13).

$$\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}} = \left.\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}}\right|_{\hat{\beta}=\hat{\beta}_0} + \left.\frac{\partial^2 L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}^T}\right|_{\hat{\beta}=\hat{\beta}_0} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right)$$

Jika $\dfrac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}} = 0$, maka :

$$\left.\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}}\right|_{\hat{\beta}=\hat{\beta}_0} + \left.\frac{\partial^2 L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}^T}\right|_{\hat{\beta}=\hat{\beta}_0} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right) = 0 \tag{2.13}$$

$$\left.\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}}\right|_{\hat{\beta}=\hat{\beta}_0} = \left.\frac{\partial^2 L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}^T}\right|_{\hat{\beta}=\hat{\beta}_0} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right)$$

Menurut Hosmer & Lemeshow (2000), Hasil substitusi persamaan (2.11) dan (2.12) ke dalam persamaan (2.13) menghasilkan estimasi parameter $\widehat{\boldsymbol{\beta}}$ ditunjukkan pada persamaan (2.14).

$$\mathbf{X}^T\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right) = -\left(-\mathbf{X}^T\mathbf{W}\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right)$$

$$\mathbf{X}^T\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right) = \mathbf{X}^T\mathbf{W}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right)$$

$$\mathbf{X}^T\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right) = \mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_0$$

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{X}^T\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)$$

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{X}^T\mathbf{W}\mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)$$

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\left[\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)\right]$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left[\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)\right]$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \qquad (2.14)$$

dengan ***z*** merupakan vektor *n* x 1 dan **W** merupakan pembobot dengan fungsi seperti dibawah ini :

$$\mathbf{W} = \begin{bmatrix} \hat{p}_1\left(1-\hat{p}_1\right) & 0 & \cdots & 0 \\ 0 & \hat{p}_2\left(1-\hat{p}_2\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{p}_n\left(1-\hat{p}_n\right) \end{bmatrix} \qquad (2.15)$$

$$z_i = \text{Logit}\left[\hat{\pi}\left(\mathbf{x}_i\right)\right] + \frac{y_i - \hat{\pi}\left(\mathbf{x}_i\right)}{\hat{\pi}\left(\mathbf{x}_i\right)\left[1-\hat{\pi}\left(\mathbf{x}_i\right)\right]}. \qquad (2.16)$$

Matriks kovarian untuk $\hat{\boldsymbol{\beta}}$ ditampilkan pada persamaan (2.17).

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^T diag\left[\hat{\pi}\left(1-\hat{\pi}\right)\right]\mathbf{X}\right)^{-1}. \qquad (2.17)$$

sedangkan fungsi dari standar *error* adalah $SE\left(\hat{\beta}\right) = \sqrt{Var\left(\hat{\beta}\right)}$.

2.4 Regresi Ridge

Regresi Ridge merupakan pengembangan metode kuadrat terkecil (*least square*) yang dapat digunakan untuk mengatasi masalah multikolinieritas yang disebabkan adanya korelasi yang tinggi antara beberapa variabel prediktor dalam model regresi, yang dapat menghasilkan hasil estimasi dari parameter menjadi tidak stabil (Draper & Smith, Applied Regression Analysis, 1998). Model regresi linier dinyatakan dengan persamaan :

$$\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.18)$$

didapatkan error, $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}^T\boldsymbol{\beta}$. Melalui metode *least square* dengan meminimalkan jumlah kuadrat *error* , dimana fungsi jumlah kuadrat *error* ada pada persamaan (2.19).

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \left(\mathbf{Y} - \mathbf{X}^T\boldsymbol{\beta}\right)^T\left(\mathbf{Y} - \mathbf{X}^T\boldsymbol{\beta}\right). \tag{2.19}$$

Metode *least square* bertujuan untuk mengusahakan turunan pertama terhadap vektor $\boldsymbol{\beta}$ sama dengan nol (Draper & Smith, Applied Regression Analysis, 1998).

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.20}$$

Namun, persamaan (2.20) tidak mampu mengakomodasi adanya kasus multikolinearitas sehingga diperlukan pengembangan dari metode ini.

Menurut Drapper & Smith (1998), dalam mengestimasi parameter regresi ridge digunakan metode *Least Square* (*LS*) dengan cara menambahkan bilangan positif kecil θ atau ridge parameter pada diagonal matriks $\mathbf{X}^T\mathbf{X}$, sehingga bias yang terjadi dapat dikendalikan atau kasus multikolinearitas dapat teratasi. Nilai estimasi untuk parameter Regresi Ridge dalam bentuk matriks dituliskan pada persamaan (2.21).

$$\hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^T\mathbf{X} + \theta\boldsymbol{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.21}$$

yang mana didapatkan dengan meminimalkan fungsi obyektif pada persamaan (2.22).

$$\phi\left(\hat{\boldsymbol{\beta}}^*\right) = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*\right) + \theta\hat{\boldsymbol{\beta}}^{*T}\hat{\boldsymbol{\beta}}^*. \tag{2.22}$$

Besarnya bilangan positif kecil θ bernilai antara 0 dan 1 yang mencerminkan besarnya bias pada estimasi regresi ridge. Apabila nilai θ adalah 0, maka estimasi regresi logistik akan sama dengan estimasi LS pada Regresi Linier. Jika nilai θ lebih dari 0, maka estimasi ridge akan bias terhadap parameter β, tetapi cenderung lebih stabil.

2.5 Regresi Logistik Ridge

Fungsi obyektif untuk Regresi Ridge yang didapatkan dari model linier $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ pada persamaan (2.22) subbab sebelumnya.

$$\phi\left(\hat{\boldsymbol{\beta}}^{*}\right) = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{*}\right)^{T}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{*}\right) + \theta\hat{\boldsymbol{\beta}}^{*T}\hat{\boldsymbol{\beta}}^{*}$$

sedangkan Regresi Logistik didapatkan bentuk fungsi obyektif dari persamaan (2.9).

$$\phi\left(\hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} y_i \ln\left[\pi\left(\mathbf{x}_i\right)\right] + \sum_{i=1}^{n}\left(1 - y_i\right)\ln\left[1 - \pi\left(\mathbf{x}_i\right)\right].$$

Kemudian, Vago & Kemeny (2006) menerapkan teknik Regresi Ridge pada Regresi Logistik, didapatkan fungsi obyektif untuk Regresi Logistik Ridge pada persamaan (2.23).

$$\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right) = \sum_{i=1}^{n} y_i \ln\left[\pi\left(\mathbf{x}_i\right)\right] + \sum_{i=1}^{n}\left(1 - y_i\right)\ln\left[1 - \pi\left(\mathbf{x}_i\right)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus}. \qquad (2.23)$$

Dari fungsi obyektif pada persamaan (2.23), didapatkan persamaan (2.24).

$$\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right) = \sum_{i=1}^{n} y_i \ln\left[\pi\left(\mathbf{x}_i\right)\right] + \sum_{i=1}^{n}\left(1 - y_i\right)\ln\left[1 - \pi\left(\mathbf{x}_i\right)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus}$$

$$\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right) = \sum_{i=1}^{n} y_i \ln\left[\pi\left(\mathbf{x}_i\right)\right] + \sum_{i=1}^{n}\ln\left[1 - \pi\left(\mathbf{x}_i\right)\right] - \sum_{i=1}^{n} y_i \ln\left[1 - \pi\left(\mathbf{x}_i\right)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus}$$

$$\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right) = \sum_{i=1}^{n} y_i \ln\left[\frac{\pi\left(\mathbf{x}_i\right)}{1 - \pi\left(\mathbf{x}_i\right)}\right] + \sum_{i=1}^{n}\ln\left[1 - \pi\left(\mathbf{x}_i\right)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus}$$

$$\phi(\hat{\boldsymbol{\beta}}^{\oplus}) = \sum_{i=1}^{n} y_i \ln\left[\pi(\mathbf{x}_i)\right] + \sum_{i=1}^{n}(1 - y_i)\ln\left[1 - \pi(\mathbf{x}_i)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus} \qquad (2.24)$$

dengan $\beta^{\oplus}$ merupakan koefisien parameter untuk Regresi Logistik Ridge. Sedangkan y_i merupakan respon berupa kategorik yang mengikuti distribusi Binomial $(1, \pi_i)$ dan $\mathbf{x}_i$ merupakan vektor untuk setiap observasi yang diambil dari matriks variabel prediktor berukuran $n \times (p + 1)$.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Selanjutnya diturunkan secara parsial terhadap $\hat{\boldsymbol{\beta}}^{\oplus}$ (Vago & Kemeny, 2006). :

$$\frac{\partial \phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right)}{\partial \hat{\boldsymbol{\beta}}^{\oplus}} = \frac{\partial}{\partial \hat{\boldsymbol{\beta}}^{\oplus}}\left[\sum_{i=1}^{n}\left[y_i\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right) - \ln\left(1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)\right)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus}\right]$$

$$= \sum_{i=1}^{n}\left[y_i\mathbf{x}_i - \frac{\mathbf{x}_i\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}\right] - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}$$

$$= \sum_{i=1}^{n}\mathbf{x}_i\left[y_i - \frac{\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}\right] - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}$$

$$= \sum_{i=1}^{n}\mathbf{x}_i\left[y_i - \pi_i(\mathbf{x}_i)\right] - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}$$

$$= \mathbf{X}^T(\mathbf{y}-\boldsymbol{\pi}) - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}.$$

Kemudian dilanjutkan pada turunan kedua (Vago & Kemeny, 2006).

$$\frac{\partial^2\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right)}{\partial\hat{\boldsymbol{\beta}}^{\oplus}\partial\hat{\boldsymbol{\beta}}^{\oplus T}} = \frac{\partial}{\partial\hat{\boldsymbol{\beta}}^{\oplus T}}\left[\frac{\partial\phi\left(\hat{\boldsymbol{\beta}}^{\oplus}\right)}{\partial\hat{\boldsymbol{\beta}}^{\oplus}}\right] = \frac{\partial}{\partial\hat{\boldsymbol{\beta}}^{\oplus T}}\left[\sum_{i=1}^{n}\left[\mathbf{x}_i^T y_i - \frac{\mathbf{x}_i^T\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}\right] - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}\right]$$

$$= \frac{\partial}{\partial\hat{\boldsymbol{\beta}}^{\oplus T}}\left[\sum_{i=1}^{n}\left[\mathbf{x}_i^T y_i - \frac{\mathbf{x}_i^T\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)}\right] - 2\theta\hat{\boldsymbol{\beta}}^{\oplus}\right]$$

$$= -\sum_{i=1}^{n}\frac{\mathbf{x}_i^T\mathbf{x}_i\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)\left[1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right) - \mathbf{x}_i^T\mathbf{x}_i\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)\right]^2}{\left[1+\exp\left(\mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}^{\oplus}\right)\right]^2} - 2\theta$$

$$= -\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \left[\frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)} - \left[\frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)} \right]^2 \right] - 2\theta$$

$$= -\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \left[\frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)} \right] \left[1 - \frac{\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)}{1+\exp\left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{\oplus}\right)} \right] - 2\theta$$

$$= -\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \pi_i(\mathbf{x}_i)\left[1-\pi_i(\mathbf{x}_i)\right] - 2\theta$$

$$= -\mathbf{X}^T \mathbf{W} \mathbf{X} - 2\theta \mathbf{I}$$

dengan $\mathbf{W} = \text{diag}\left[\hat{\pi}\left(x_i\right)\left[1-\hat{\pi}\left(x_i\right)\right]\right]$.

Estimasi parameter Regresi Logistik Ridge menggunakan metode MLE dengan iterasi *Newton-Raphson* yang akan digunakan untuk memaksimumkan fungsi obyektif pada persamaan (2.24). Kemudian diekspansikan di sekitar $\beta^{\oplus}$menurut Deret *Taylor* dan didapatkan persamaan (2.25).

$$\left.\frac{\partial \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0{}^{\oplus}} = -\left.\frac{\partial^2 \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus} \partial \hat{\boldsymbol{\beta}}^{\oplus T}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0{}^{\oplus}} (\hat{\boldsymbol{\beta}}^{\oplus} - \hat{\boldsymbol{\beta}}_0{}^{\oplus}) \cdot \qquad (2.25)$$

Menurut Vago & Kemeny (2006), hasil penurunan di substitusikan ke dalam persamaan (2.25) menghasilkan estimasi parameter Regresi Logistik Ridge pada persamaan (2.26).

$$\hat{\boldsymbol{\beta}}^{\oplus} = (\mathbf{X}^T \mathbf{W} \mathbf{X} + 2\theta \mathbf{I})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z} . \qquad (2.26)$$

dengan $\beta^{\oplus}$ adalah parameter ridge untuk Regresi Logistik Ridge yang merupakan bilangan positif kecil. *z* merupakan vector berukuran *n* × 1, dengan $z_i = \text{Logit}[\hat{\pi}(\mathrm{x}_i)] + \frac{y_i - \hat{\pi}(\mathrm{x}_i)}{\hat{\pi}(\mathrm{x}_i)[1-\hat{\pi}(\mathrm{x}_i)]}$ dan $\mathbf{W} = \text{diag}[\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))]$.

dengan menambahkan *ridge parameter* untuk Regresi Logistik Ridge pada elemen diagonal matriks kovarian dari Regresi Logistik, maka variansi Regresi Logistik Ridge dapat dihiung dengan formula pada persamaan (2.27).

$$Var(\hat{\boldsymbol{\beta}}^{\oplus}) = \left(\mathbf{X}^T \text{ diag}[\hat{\pi}_i(1-\hat{\pi}_i)]\mathbf{X} + \theta^{\oplus}\mathbf{I}\right)^{-1}. \quad (2.27)$$

2.6 Pengujian Parameter

Pengujian parameter digunakan untuk mengetahui variabel prediktor mana saja yang berpengaruh terhadap variabel respon. Pengujian ini dilakukan dua kali secara berurutan, yaitu uji serentak (bersama-sama) dan uji parsial (sendiri-sendiri). Pengujian signifikansi parameter secara serentak dilakukan dengan menggunakan *Likelihood Ratio Test* dengan hipotesis sebagai berikut.

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

H_1 : minimal ada satu $\beta_j \neq 0,\ j = 1, 2, ..., p$

Statistik uji yang digunakan adalah :

$$G = -2\log\left(\frac{L_0}{L_1}\right) = -2[\ln(L_0) - \ln(L_1)]. \quad (2.28)$$

Nilai $-2[\ln(L_0) - \ln(L_1)]$ mengikuti distribusi *chi-square* dengan derajat bebas (*df*) = *p*. Pengambilan keputusan akan ditolak apabila $-2[\ln(L_0) - \ln(L_1)] \geq \chi^2_{(p,\alpha)}$, yang artinya model berpengaruh signifikan. Sehingga dapat dilanjutkan pengujian signifikansi parameter secara parsial.

Pengujian signifikansi secara parsial dilakukan dengan metode *Wald Test* untuk mengetahui variabel-variabel prediktor yang signifikan terhadap peluang sukses. Hipotesis yang digunakan untuk uji ini adalah :

$H_0 : \beta_j = 0$

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

Statistik uji :

$$Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (2.29)$$

dengan $SE(\hat{\beta}_j) = \sqrt{\text{var}(\hat{\beta}_j)}$.

Daerah kritis : H_0 ditolak, jika nilai $|Z| > Z_{\alpha/2}$ atau nilai *p-value* < α. Artinya, variabel ke-*j* berpengaruh signifikan terhadap pembentukan model.

2.7 Analisis Diskriminan

Menurut Hair, Black, Babin, & Anderson (2006), analisis diskriminan merupakan salah satu metode dalam analisis multivariat dengan metode depedensi (dimana hubungan antar variabel sudah bisa dibedakan mana variabel terikat dan mana variabel bebas). Analisis diskriminan digunakan pada kasus dimana variabel bebas berupa metrik (interval atau rasio) dan variabel terikat berupa data nonmetrik (nominal atau ordinal).

Analisis diskriminan dapat digunakan menentukan fungsi yang membedakan antar kelompok dan mengkelaskan obyek baru ke dalam kelompoknya (2007). Ketika variabel dependen terdiri dari kedua kelompok, teknik analisis diskriminan ini disebut analisis diskriminan dua kelompok. Ketika variabel dependen terdiri dari tiga atau lebih kelompok, teknik analisis diskriminan ini disebut analisis diskriminan berganda (*multiple discriminant analysis*). Pengelompokkan pada analisis diskriminan bersifat *mutually axclusive*, yaitu jika suatu pengamatan telah masuk pada salah satu kelompok maka tidak dapat menjadi anggota dari kelompok yang lain (Hair, Black, Babin, & Anderson, 2006). Terdapat beberapa asumsi yang harus dipenuhi dalam analisis diskriminan yaitu asumsi homogenitas, asumsi distribusi normal multivariat dengan pendekatan *Expected Cost of Misclassification* (*ECM*) (Johnson & Winchern, 2007).

2.7.1 Uji Normal Multivariat

Pengujian distribusi Normal Multivariat dilakukan dengan menggunakan metode *mardia's test on multinormality*. Uji dengan metode *mardia's test* menggunakan nilai *skewness* dan nilai *kutosis* untuk menguji apakah suatu data berdistribusi normal multivariat. Nilai dari *skewness* dan *kurtosis* data multivariat dapat dihitung dengan persamaan sebagai berikut :

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\left(\mathbf{x}_i - \bar{\mathbf{x}} \right)^{\mathrm{T}} S^{-1} \left(\mathbf{x}_j - \bar{\mathbf{x}} \right) \right]^3 \qquad (2.30)$$

dan

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\left(\mathbf{x}_i - \bar{\mathbf{x}} \right)^{\mathrm{T}} S^{-1} \left(\mathbf{x}_j - \bar{\mathbf{x}} \right) \right]^2 \qquad (2.31)$$

dengan $\mathbf{S} = \dfrac{\sum_{j=1}^{n} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}}{n}$

Central moment orde ketiga untuk normal multivariat adalah nol, sehingga $b_{1,\mathrm{p}}$ akan bernilai nol ketika $\boldsymbol{x}$ berdistribusi normal dengan parameter μ dan σ^2. Jika $\boldsymbol{x}$ berdistribusi normal maka $b_{2,\mathrm{p}}$ akan menjadi $p(p+2)$. Hipotesis yang digunakan dalam pengujian ini adalah sebagai berikut (Rancher, 2002):

H_0 : Data mengikuti distribusi normal multivariat

H_1 : Data tidak berdistribusi normal multivariat

dengan statistik uji yang digunakan adalah :

$$z_1 = \frac{(p+1)(n+1)(n+3)}{6\left[(n+1)(p+1)-6\right]} b_{1,p} \,. \qquad (2.32)$$

Hipotesis awal akan ditolak jika nilai $z_1 \geq \chi^2_{0,05,\frac{1}{6}p(p+1)(p+2)}$ dan statistik uji untuk z_2 adalah sebagai berikut :

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \,. \qquad (2.33)$$

Nilai z_2 diharapkan tidak terlalu kecil dan tidak terlalu besar. Nilai z_2 menggambarkan bentuk puncak distribusi. Jika nilainya terlalu besar atau terlalu kecil akan menunjukkan puncak distribusi yang terlalu lancip atau terlalu landai.

2.7.2 Uji Homogenitas

Asumsi lain yang harus terpenuhi adalah matriks varians kovarians antar kelompok yang homogen. Statistik uji yang digunakan adalah Box's M. apabila terdapat dua kelompok, maka

hipotesis yang digunakan adalah sebagai berikut (Johnson & Winchern, 2007) :

$H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ (matriks varians kovarians bersifat homogen)

$H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ (matriks varians kovarians tidak homogen).

Statistik Uji *Box's M* dihitung dari persamaan (2.34) :

$$\chi^2 = -2(1-c_i)\left[\frac{1}{2}\sum_{i=1}^{2} v_i ln\left|\boldsymbol{S}_i\right| - \frac{1}{2} ln\left|\boldsymbol{S}_{pool}\right|\sum_{i=1}^{2} v_i\right] \qquad (2.34)$$

dengan

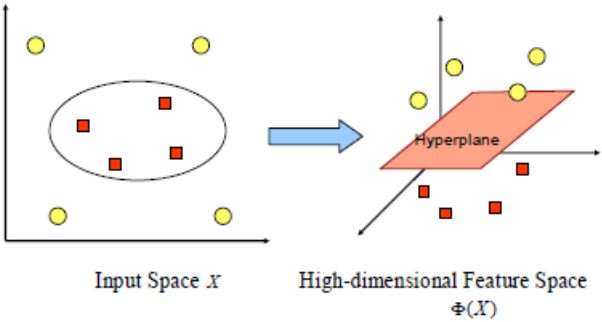
$$S_{pool} = \frac{\sum_{i=1}^{2} v_i S_i}{\sum_{i=1}^{2} v_i}, v_i = n_i \text{ -1, dan } S_i = \frac{\sum_{j=1}^{n} (\bar{\mathbf{x}}_{1j} - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_{2j} - \bar{\mathbf{x}}_2)}{n-1}$$

$$c_1 = \left[\sum_{i=1}^{2} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{2} v_i}\right]\left[\frac{2p^2 + 3p - 1}{6(\text{p+1})}\right].$$

Tolak H_0 jika $\chi^2 \geq \chi^2_{\frac{1}{2}p(\text{p+1})}$ maka dapat dikatakan matriks kovarian telah tidak homogen.

2.8 Analisis Diskriminan Kernel

Salah satu metode yang dapat digunakan untuk klasifikasi pada kasus *non-linear* adalah Analisis Diskriminan Kernel. Sama halnya dengan analisis diskriminan linier, pada Analisis Diskriminan Kernel digunakan pendekatan *Fisher*. Data yang diasumsikan mengikuti suatu distribusi yang bukan merupakan distribusinya akan menghasilkan klasifikasi yang kurang baik. Menurut Mika, Ratsch, Jason, Scholkopf, & Muller (1999), langkah pertama dari Analisis Diskriminan Kernel adalah memetakan data *non-linear* kedalam *feature space* $\mathcal{F}$. Pemetaan ini dapat dilihat pada Gambar 2.2



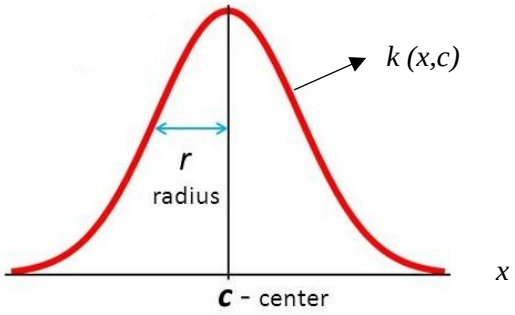
Gambar 2.2 Pemetaan data non-linear kedalam feature space F

Pada ruang vector yang baru ini, hyperplane yang memisahkan kedua kelas tersebut dapat dikontruksikan (Nugroho, Witarto, & Handoko, 2003). Penggunaan fungsi kernel memungkinkan analisis diskriminan linier bekerja secara efisien dalam suatu kernel *space* berdimensi tinggi yang linier.

Penelitian ini menggunakan pendekatan kernel *Gaussian RBF* pada Analisis Diskriminan Kernel, fungsi kernel *Gaussian RBF* dapat dilihat pada persamaan (2.35).

$$k(\mathbf{x},\mathrm{c}) = \exp\left(-(x-c)^2 / r^2\right) \cdot \qquad (2.35)$$

(Yang, Jin, Yang, Zhang, & Frangi, 2004).
Kernel *Gaussian RBF* dapat divisualisasikan pada Gambar 2.3.



Gambar 2.3 Gaussian RBF Kernel

Langkah pertama dari analisis diskriminan kernel adalah memetakan data *non-linear* kedalam *feature space F*. Misal Φ adalah pemetaan non-linier dari *feature space F*, diskriminan linier F akan didapatkan dengan memaksimumkan persamaan (2.36) :

$$J(\boldsymbol{\omega}) = \frac{\omega^T S_B^{\Phi} \omega}{\omega^T S_W^{\Phi} \omega} \tag{2.36}$$

dengan $\boldsymbol{\omega} \in \mathcal{F}$ dan $S_B^{\Phi} = (m_1^{\Phi} - m_2^{\Phi})(m_1^{\Phi} - m_2^{\Phi})^T$ serta $S_W^{\Phi} = \sum_{i=1,2} \sum_{x \in \chi_i} (\Phi(x) - m_i^{\Phi})(\Phi(x) - m_i^{\Phi})^T$ dengan persamaan $m_i^{\Phi} = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j^i)$.

Diskriminan kernel dengan pendekatan *Fisher* dihitung dengan memasukkan fungsi kernel kedalam persamaan (2.36) dan fungsi perluasan dari $\boldsymbol{\omega}$ pada persamaan (2.37).

$$\boldsymbol{\omega} = \sum_{i=1}^{l} \alpha_i \Phi(x_i) . \tag{2.37}$$

Persamaan (2.37) dan persamaan m_i^{Φ} menghasilkan persamaan (2.38)

$$\boldsymbol{\omega}^T m_i^{\Phi} = \frac{1}{l_i} \sum_{j=1}^{l} \sum_{k=1}^{l_i} \alpha_j \, \mathrm{k}(x_j, x_k^i) = \alpha^{\mathrm{T}} M_i , \tag{2.38}$$

dengan $(M_i)_j = \frac{1}{l_i} \sum_{k=1}^{l_i} \mathrm{k}(x_j, x_k^i)$. Melalui persamaan (2.38) diperoleh persamaan baru dari $\boldsymbol{\omega}^T S_B^{\Phi} \boldsymbol{\omega}$.

$$\boldsymbol{\omega}^T S_B^{\Phi} \boldsymbol{\omega} = \alpha^T M \alpha , \tag{2.39}$$

dengan $M = (M_1 - M_2)(M_1 - M_2)^T$. Persamaan $\boldsymbol{\omega}^T S_W^{\Phi} \boldsymbol{\omega}$ juga berubah menjadi persamaan (2.40)

$$\boldsymbol{\omega}^T S_W^{\Phi} \boldsymbol{\omega} = \alpha^T N \alpha , \tag{2.40}$$

dengan $N = \sum_{j=1,2} K_j (\boldsymbol{I} - 1_{lj}) K_j^{\,T}$. Diketahui K_j. adalah matriks $l \times l_j$ dengan $(K_j)_{nm} = k(x_n, x_m^j)$, *I* adalah matriks identitas, dan 1_{lj} adalah

semua entri dari $1/l_j$. Persamaan Analisis Diskriminan Kernel dengan pendekatan Fisher didapatkan dengan memaksimumkan persamaan (2.41).

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}. \qquad (2.41)$$

Pola baru dari *x* akan diproyeksikan kedalam dengan fungsi yang dituliskan pada persamaan (2.42).

$$(\boldsymbol{\omega}, \Phi(x)) = \sum_{i=1}^{l} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}). \qquad (2.42)$$

Aturan klasifikasi pada Analisis Diskriminan Kernel menggunakan aturan Bayes berdasarkan peluang posterior terbesar. Berdasarkan fungsi kepadatan peluang, maka peluang posterior dari kelompok **x** dapat dihitung. Menurut Khattree (2000), misalkan $\mathbf{x_1}, ..., \mathbf{x}_{n_t}$ adalah sampel acak dari populasi Π_t dan **x** adalah sebuah amatan tambahan dari populasi Π_t yang mana tidak diketahui fungsi kepadatan peluang $f_t(\mathbf{x})$. Fungsi kepadatan peluang $f_t(\mathbf{x})$ dapat diestimasi dengan :

$$\hat{f}_t(\mathbf{x}) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_t(\mathbf{x} - \mathbf{x}_i)$$

Dimana kuantitas $K_t(\mathbf{x})$ disebut fungsi kernel kelompok ke-*t*.

Misalkan pada data dikotomus, dimana $\hat{f}_1(\mathbf{x})$ adalah penduga fungsi kernel dari kelompok Π_1, dan P_1 adalah peluang awal dari kelompok Π_1. Peluang posterior suatu **x** berasal dari kelompok Π_1, adalah

$$P(\Pi_1 \mid \mathbf{x}) = \frac{P_1 \hat{f}_1(\mathbf{x})}{P_1 \hat{f}_1(\mathbf{x}) + P_2 \hat{f}_2(\mathbf{x})}, \text{ dimana } P_1 = \frac{n_1}{n_1 + n_2}$$

Sedangkan, peluang posterior suatu **x** berasal dari kelompok Π_2 adalah

$$P(\Pi_2 \mid \mathbf{x}) = 1 - P(\Pi_1 \mid \mathbf{x}) = \frac{P_2 \hat{f}_2(\mathbf{x})}{P_1 \hat{f}_1(\mathbf{x}) + P_2 \hat{f}_2(\mathbf{x})} \text{ dimana } P_2 = \frac{n_2}{n_1 + n_2}$$

Jika $P(\Pi_1 \mid \mathbf{x}) > P(\Pi_2 \mid \mathbf{x})$ maka pengamatan **x** diklasifikasikan ke Π_1, demikian sebaliknya (Johnson & Wichern, 2007).

2.9 Evaluasi Performansi Ketepatan Klasifikasi

Data aktual dan data hasil prediksi dari model klasifikasi disajikan dengan menggunakan tabilasi silang (*confusion matrix*), yang mengandung informasi tentang kelas data yang aktual dipresentasikan pada baris matriks dan kelas data hasil prediksi suatu algoritma dipresentasikan pada kolom matriks klasifikasi (Kohavi dan Provost, 1998 diacu dalam Sain, 2013). Tabel klasifikasi dapat dilihat pada Tabel 2.2 berikut :

Tabel 2.2 Tabel Klasifikasi

		Nilai Prediksi	
	Kelas	Positif	Negatif
Nilai Aktual	Positif	TP	FN
	Negatif	FP	TN

keterangan :

TP : *True Positive*, data aktual positif dan diklasifikasikan positif

FP : *False Positive*, data aktual negatif dan diklasifikasikan positif

FN : *False Negative*, data aktual positif, namun diklasifikasikan negatif

TN : *True Negative*, data aktual negatif dan diklasifikasikan negatif

Berdasarkan Tabel 2.2, dapat dilakukan perhitungan akurasi klasifikasi, sensitivitas, dan spesifisitas. Sensitivitas adalah tingkat positif benar atau akurasi kelas yang positif, sedangkan spesifisitas adalah tingkat negatif benar atau akurasi kelas negatif. Berikut ini rumus perhitungan akurasi klasifikasi, sensitivitas, dan spesifisitas (Morton, Hebel, & McCarter, 2008) :

$$\text{Rata-rata } akurasi = \frac{\sum_{k=1}^{K} \frac{TP_k + TN_k}{TP_k + TN_k + FN_k + FP_k}}{k}; k = 1, 2, \ldots, K \qquad (2.43)$$

$$\text{Rata-rata } sensitivitas = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{k}; k = 1, 2, \ldots, K \tag{2.44}$$

$$\text{Rata-rata } spesifisitas = \frac{\sum_{k=1}^{K} \frac{TN_k}{TN_k + FP_k}}{k}; k = 1, 2, \ldots, K . \tag{2.45}$$

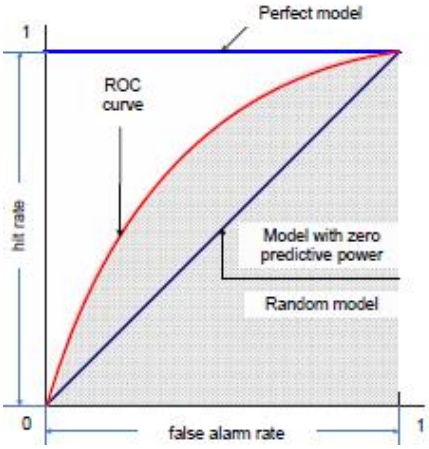
Selanjutnya evaluasi performansi model klasifikasi menggunakan $G - Mean$. $G - Mean$ merupakan rata-rata geometrik sensitivitas dan spesifisitas. Apabila semua kelas data positif tidak dapat diprediksi maka $G - Mean$ akan bernilai nol., sehingga diharapkan suatu algoritma klasifikasi mencapai nilai $G - Mean$ yang tinggi (Kubat & Matwin, 1997). Dengan rumus sebagai berikut :

$$\text{Rata-rata } G - Mean = \sqrt{\frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{k} \times \frac{\sum_{k=1}^{K} \frac{TN_k}{TN_k + FP_k}}{k}}; k = 1, 2, \ldots, K \tag{2.46}$$

(Yuchun, Ya-Qing, Chawla, & Sven, 2002).

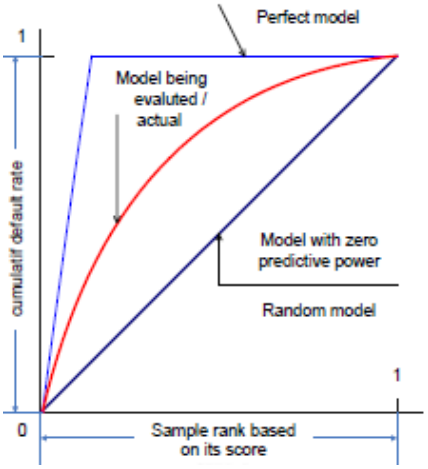
Metode lain dalam mengukur performa klasifikasi adalah menggunakan kurva *ROC* (*Receiving Operating Characteristic*). Area dibawah kurva *ROC* biasa disebut *Area Under The ROC Curve* (*AUC*). Umumnya, *AUC* digunakan untuk mengukur klasifikasi apabila data *imbalanced*. Hal ini dikarenakan *AUC* menggunakan sensitivitas atau spesitifitas sebagai dasar pengukuran. Nilai *AUC* berada diantara 0 dan 1. Apabila nilai *AUC* semakin mendekati 1, maka model klasifikasi yang terbentuk semakin akurat. Kurva *ROC* yang baik berada disebelah atas dari garis diagonal (0,0) dan (1,1), sehingga tidak ada nilai *AUC* yang lebih kecil dari 0,5.

Perhitungan *AUC* dilakukan melalui beberapa pendekatan, yang paling banyak digunakan adalah *trapeziodal method*. Pendekatan tersebut berbasis metode geometris yang bedasarkan interpolasi linear antara masing – masing titik pada kurva *ROC*. Kurva *ROC* dapat divisualisasikan pada Gambar 2.4.



Gambar 2.4 ROC Curve (Haerdle, et.al., 2014)

Selain itu, *ROC* memiliki konsep yang mirip dengan *Cumulative Accuracy Prole* (*CAP*) sedangkan wilayah dibawah kurva pada *CAP* disebut *Accuracy Ratio* (*AR*). Metode klasifikasi dapat dikatakan baik, jika *AR* bernilai tinggi atau mendekati 1. Kurva *CAP* dapat dilihat pada Gambar 2.5.



Gambar 2.5 CAP Curve (Haerdle, et.al., 2014)

Apabila $Y = (0,1)$, maka *Accuracy Ratio* didapatkan dari persamaan (2.47).

$$AR = \frac{\int_0^1 Y_{actual}\, F\, dF - \frac{1}{2}}{\int_0^1 Y_{perfect}\, F\, dF - \frac{1}{2}}. \tag{2.47}$$

Selanjutnya, dari hubungan antara *AR* dan *AUC* didapatkan persamaan (2.48)

$$AR = 2AUC - 1 \tag{2.48}$$

sehingga didapatkan rumus rata-rata *AUC* pada persamaan (2.49).

$$\text{Rata-rata } AUC = \frac{\sum_{k=1}^{K} \frac{1}{2}\left(AR_k + 1\right)}{k}; k = 1,2,\ldots,K \tag{2.49}$$

(Haerdle, *et.al.*, 2014).

Khusus untuk kasus biner, nilai *AUC* dapat didekati dengan nilai *Balanced Accuracy* (Bekkar, Djemaa, & Alitouche, 2013).

$$AUC = \frac{1}{2}(sensitivity + specitifity). \tag{2.50}$$

Kategori berdasarkan nilai *AUC* dapat disajikan ke dalam tabel berikut :

Tabel 2.3 Kategori Pengklasifikasian Model Berdasarkan Nilai AUC

Nilai *AUC*	Model Diklasifikasikan Sebagai
0,90-1,00	*Excellent*
0,80-0,90	*Very Good*
0,70-0,80	*Good*
0,60-0,70	*Fair*
0,50-0,60	*Poor*

Sumber : Bekkar, Djemaa, & Alitouche (2013)

2.10 *Stratified k-fold Cross Validation*

Pada *k-fold cross-validation* data akan dipartisi secara acak menjadi *k* bagian atau *folds* yaitu D_1, D_2, ..., D_k dengan masing-masing ukuran yang hampir sama. Validasi menggunakan *training* dan *testing* dilakukan sebanyak *k* kali. Pada iterasi ke-*i* , partisi D_i akan diatur sebagai data *testing* dan partisi yang tersisa lainnya

akan digunakan sebagai data *training* untuk memperoleh model. Artinya, pada iterasi yang pertama, partisi D_2, D_3, ..., D_k akan menjadi data *training* untuk medapatkan model yang pertama yang akan diuji dengan data pada partisi D_1. Pada iterasi kedua partisi $D_1, D_3, ..., D_k$ akan menjadi data *training* kemudian D_2 akan menjadi data *testing*, begitu seterusnya (Han, Kamber, & Pei, 2012). Hal ini dilakukan agar semua observasi dari setiap data yang digunakan dalam penelitian, bisa dijadikan data *training* dan data *testing*.

Suatu cara pengecekan sederhana yang dapat dilakukan agar data *training* dan data *testing* representatif yaitu dengan cara memastikan bahwa setiap kelas dalam dataset penuh harus terwakili dalam proporsi yang tepat untuk data *training* dan data *testing*. Jika semua sampel dengan kelas tertentu dihilangkan dari *training set*, *classifier* tidak dapat diharapkan belajar dengan baik dari data yang tersedia dalam melakukan klasifikasi pada *testing set*. Maka harus dipastikan bahwa pengambilan sampel dilakukan dengan cara *random* yang menjamin bahwa setiap kelas terwakili baik pada *training* dan *testing*. Prosedur ini dinamakan stratifikasi, namun stratifikasi hanya menyediakan perlindungan yang lemah terhadap kelas yang tidak representatif dalam *training* dan *testing* set. Adapun kelebihan dari *stratified k-fold cross validation* adalah menghindari adanya *overfitting* pada data *training* (Zhang, Wu, & Wang, 2011).

2.11 Desa Tertinggal

Desa tertinggal merupakan desa yang ketersediaan sarana dan prasarana dasar wilayahnya kurang atau tidak tersedia sehingga menghambat pertumbuhan dan perkembangan kehidupan masyarakatnya di bidang ekonomi dan sosial. Jika kondisi desa tertinggal dibiarkan, maka desa tersebut mengalami ketertinggalan daripada desa lainnya dan menghambat pertumbuhan ekonomi di suatu wilayah (Adisasmita, 2005).

Desa tertinggal adalah desa-desa yang kondisinya secara ekonomi relatif tertinggal dibandingkan desa-desa lainnya. Kemajuan atau ketertinggalan suatu desa dicerminkan oleh indikator utama, yaitu tinggi rendahnya rata-rata pengeluaran per

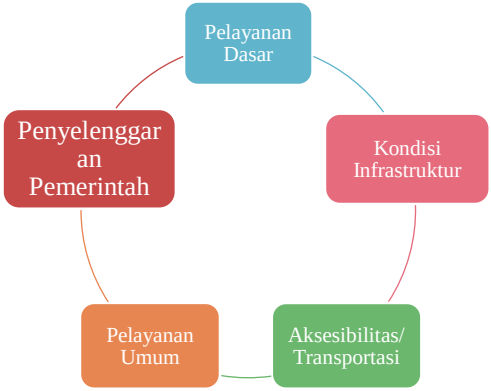
kapita penduduk desa. Beberapa faktor diduga menjadi penyebab kemajuan atau ketertinggalan suatu desa, yaitu faktor alam/lingkungan, faktor kelembagaan, faktor sarana/prasarana dan akses, serta faktor sosial ekonomi penduduk (BPS, 2015).

Pembangunan desa merupakan konsep multidimensional yang kompleks. Pengukuran tingkat kemajuan pembangunan desa diharapkan tetap mengacu pada kompleksitas konsep tersebut meskipun perlu diupayakan adanya penyederhanaan dalam hal instrumen dan teknis pengukurannya. Beberapa dimensi disusun untuk mencakup sekaligus beberapa variabel dan indikator. Antar dimensi diharapkan bersifat saling melengkapi untuk menggambarkan tingkat kemajuan pembangunan di setiap desa. Hasil sintesis oleh BPS membagi dimensi IPD menjadi 5 dimensi dengan disesuaikan dengan ketersediaan data/variabel dalam data Potensi Desa 2014 sebagai berikut :

1. **Pelayanan Dasar :** Aspek pelayanan dasar untuk mewujudkan bagian dari kebutuhan dasar, khusus untuk pendidikan dan kesehatan. Variabel yang termasuk sebagai komponen penyusunnya meliputi ketersediaan dan akses terhadap fasilitas pendidikan seperti TK, SD, SMP, dan SMA; serta ketersediaan dan akses terhadap fasilitas kesehatan seperti rumah sakit, rumah sakit bersalin, puskesmas/pustu, tempat praktik dokter, poliklinik/balai pengobatan, tempat praktik bidan, poskesdes, polindes, dan apotik.
2. **Kondisi Infrastruktur :** Mewakili Kebutuhan Dasar; Sarana; Prasarana; Pengembangan Eknomi Lokal; dan Pemanfaatan Sumberdaya Alam secara Berkelanjutan dengan memisahkan aspek aksesibilitas/transportasi. Variabel-variabel penyusunnya mencakup ketersediaan infrastruktur ekonomi seperti: kelompok pertokoan, minimarket, maupun toko kelontong, pasar, restoran, rumah makan, maupun warung/ kedai makanan, akomodasi hotel atau penginapan, serta bank; ketersediaan infrastruktur energi; ketersediaan infrastruktur air bersih dan sanitasi; serta ketersediaan dan kualitas infrastruktur komunikasi dan informasi.

3. **Aksesibilitas/Transportasi :** Sarana dan prasarana transportasi memiliki kekhususan dan prioritas pembangunan desa sebagai penghubung kegiatan sosial ekonomi dalam desa. Variabel-variabel penyusunnya meliputi ketersediaan dan akses terhadap sarana transportasi; dan aksesibilitas transportasi.
4. **Pelayanan Umum :** Upaya pemenuhan kebutuhan pelayanan atas barang, jasa, dan/atau pelayanan administratif dengan tujuan memperkuat demokrasi, kohesi sosial, perlindungan lingkungan, dan sebagainya. Pelayanan dalam dimensi ini mewakili aspek lingkungan dan aspek pemberdayaan masyarakat serta mengacu pada ketersediaan data Potensi Desa 2014. Aspek lingkungan dalam hal ini terkait dengan kesehatan lingkungan masyarakat, sedangkan aspek pemberdayaan masyarakat diwakili dengan keberadaan kelompok masyarakat.
5. **Penyelenggaran Pemerintahan :** Mewakili indikasi kinerja pemerintahan desa merupakan bentuk pelayanan administratif yang diselenggarakan penyelenggara pelayanan bagi warga yang dalam hal ini adalah pemerintah.

   Berikut Gambar dari 5 dimensi dengan disesuaikan dengan ketersediaan data/variabel dalam data Potensi Desa 2014 :



Gambar 2.6 Lima Dimensi Indeks Pembangunan Desa

BAB III

METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari data Potensi Desa tahun 2014 yaitu data yang berkaitan dengan indikator-indikator desa tertinggal di Jawa Timur. Jumlah Variabel penelitian yang digunakan sebanyak 8 variabel berskala numerik dan diklasifikasikan menjadi dua kelas yaitu desa tertinggal atau desa tidak tertinggal (desa berkembang dan desa mandiri). Jumlah data desa Jawa Timur sebanyak 7.721 desa yang dikelompokkan menjadi 207 desa tertinggal (kelas minoritas) dan 7514 desa tidak tertinggal (kelas mayoritas) sehingga rasio data sebesar 1 : 32,3.

Dalam kurun waktu 10 tahun atau 2 sampai 3 tahun sebelum pelaksanaan Sensus, Potensi Desa (Podes) telah mekakukan pendataan penduduk sebanyak tiga kali. Potensi Desa tahun 2014 dilaksanakan selama bulan April, mencakup seluruh wilayah administrasi pemerintahan antara lain desa/kelurahan, kecamatan, dan kabupaten/kota. Pendataan berikutnya akan dilakukan pada tahun 2018.

3.2 Variabel Penelitian

Variabel respon (Y) merupakan variabel yang berisi kelas yang terdiri atas dua kategori yaitu {0} untuk desa tidak tertinggal (desa berkembang dan desa mandiri) dan {1} untuk desa tertinggal. Variabel Prediktor (X) merupakan variabel yang dapat mempengaruhi variabel respon. Berikut adalah variabel respon dan variabel prediktor yang digunakan dalam penelitian ini :

a. Variabel respon :

   Y : Status ketertinggalan desa yang terdiri dari desa tertinggal {0} dan desa tidak tertinggal {1}.

b. Variabel prediktor :

1. Pelayanan Dasar

   X_1 : Rasio banyaknya SD/MI tehadap total murid SD/MI.

X_2 : Rasio banyaknya praktik bidan terhadap total penduduk.
X_3 : Rasio banyaknya poskesdes terhadap total penduduk.

2. Faktor Infrastruktur
   X_4 : Rasio banyaknya toko/warung kelontong terhadap total penduduk.
   X_5 : Rasio banyaknya keluarga pengguna listrik terhadap total rumah tangga.
3. Aksesibilitas/Transportasi
   X_6 : Jarak tempuh per kilometer ke kantor camat.
4. Pelayanan Umum
   X_7 : Rasio banyaknya penderita gizi buruk terhadap total penduduk.
5. Penyelenggaran Pemerintah
   X_8 : Rasio Pendapatan Asli Desa (PAD) terhadap total penduduk.

Adapun penjelasan dari variabel-variabel yang digunakan dalam penelitian ini adalah sebagai berikut :

Konsep dan definisi yang digunakan mengacu pada BPS yaitu :

a. Status ketertinggalan desa. Desa tertinggal adalah desa-desa yang kondisinya relatif tertinggal dibandingkan desa-desa lainnya. Beberapa faktor diduga menjadi penyebab kemajuan atau ketertinggalan suatu desa, yaitu faktor alam/lingkungan, faktor kelembagaan, faktor sarana/prasarana dan akses, serta faktor sosial penduduk.
b. Rasio banyaknya SD/MI, yaitu jumlah sekolah SD dibagi total jumlah murid yang sekolah dikali 100.
c. Rasio banyaknya praktik bidan adalah jumlah tepat praktik bidan dibagi total penduduk dikali 100. Bidan adalah petugas paramedik yang berdomisli atau tinggal di desa atau kelurahan atau yang bertugas sebagai bidan di desa dengan SK.
d. Rasio banyaknya poskesdes adalah jumlah banyaknya poskesdes dibagi total penduduk dikali 100. Poskedes adalah upaya kesehatan bersumber daya masyarakat (UKBM) yang dibentuk di desa dalam rangka menyediakan pelayanan kesehatan dasar masyarakat desa.

e. Rasio banyaknya toko/warung kelontong yaitu jumlah toko/warung kelontong dibagi total penduduk dikali 100. Toko/warung kelontong adalah bangunan (kedai) yang menjual beraneka barang secara eceran.
f. Rasio banyaknya keluarga pengguna listrik adalah jumlah keluarga pengguna listrik PLN dan Non PLN dibagi total penduduk dikali 100.
g. Jarak tempuh ke kantor camat adalah jarak yang harus ditempuh oleh penduduk dari kantor kepala desa/lurah ke kantor camat dalam kilometer.
h. Rasio banyaknya penderita gizi buruk adalah jumlah penderita gizi buruk selama 3 bulan tahun terakhir dibagi total penduduk
i. Rasio Pendapatan Asli Desa (PAD) jumlah pendapatan asli desa berupa hasil usaha, hasil aset, swadaya, partisipasi, gotong royong, bagian dari hasil pajak daerah. Dana hibah dari pihak ketiga maupun pemerintah dan lain-lain dibagi total penduduk.

Struktur data yang digunakan pada penelitian ini adalah sebagai berikut :

Tabel 3.1 Struktur Data Penelitian

Desa	Respon	Prediktor			
	Y	X_1	X_2	…	X_p
1	0	1,261	0,03	…	0,473
2	0	2,632	0	…	0,583
3	0	1,863	0,029	…	0,345
⋮	⋮	⋮	⋮	⋮	⋮
7721	0	0,23	0,022	…	4,946

3.3 Langkah Analisis

Langkah-langkah analisis yang dilakukan pada penelitian ini sebagai berikut :

1. Mendeskripsikan karakteristik desa berdasarkan variabel yang diduga mempengaruhi status ketertinggalan desa.
   a. Melakukan *preprocessing data* dengan pemilihan data variabel, *filtering* data sesuai jenis data menggunakan *software R.*

b. Melakukan analisis statistika deskriptif pada variabel-variabel yang diduga berpengaruh pada data desa tertinggal di Jawa Timur tahun 2014.

2. Melakukan klasifikasi data *imbalanced* dengan beberapa metode, yaitu :
   a. Metode Klasifikasi Regresi Logistik
      i. Mendeteksi kasus multikolinearitas dengan menggunakan *VIF*/korelasi untuk setiap variabel prediktor.
      ii. Membagi data menjadi data *training* dan *testing* dengan menggunakan stratifikasi 10-*fold cross validation*.
      iii. Melakukan klasifikasi data *training* menggunakan Regresi Logistik.
      iv. Menghitung ketepatan klasifikasi pada data *testing* dan data *training* dengan menggunakan persamaan (2.43), (2.44), (2.45), (2.46), dan (2.49).
      v. Melakukan pengujian parameter yang signifikan.
      vi. Melakukan pemilihan variabel signifikan berdasarkan metode *backward elimination*.
      vii. Mengulangi langkah i, ii, iii, dan iv menggunakan variabel signifikan yang didapat pada langkah ke vi.
   b. Metode Klasifikasi Regresi Logistik Ridge
      i. Membagi data menjadi data *training* dan *testing* dengan menggunakan stratifikasi 10-*fold cross validation*.
      ii. Melakukan klasifikasi data *training* menggunakan Regresi Logistik Ridge.
      iii. Menghitung ketepatan klasifikasi pada data *testing* dan data *training* dengan menggunakan persamaan (2.43), (2.44), (2.45), (2.46), dan (2.49).
      iv. Melakukan analisis seperti langkah 2.a (vi).
   c. Metode Klasifikasi Analisis Diskriminan Kernel
      i. Membagi data menjadi data *training* dan *testing* dengan menggunakan stratifikasi 10-*fold cross validation*.

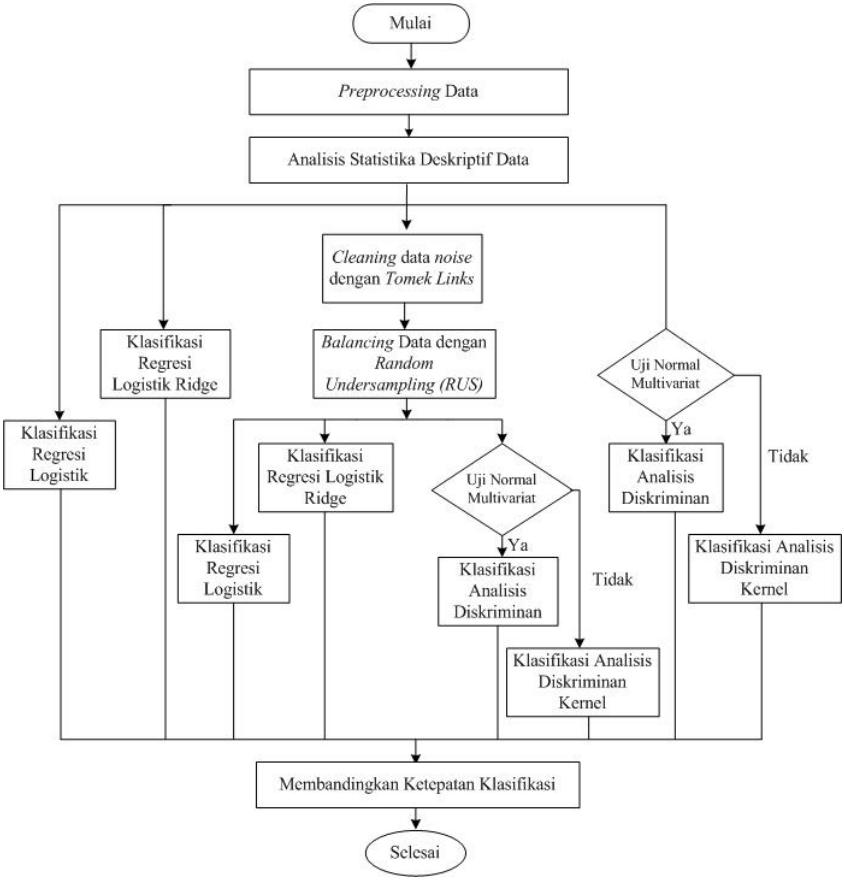
ii. Melakukan uji asumsi yang meliputi uji asumsi normal multivariat (Bab 2.7.1), uji asumsi homogenitas (Bab 2.7.2).
iii. Jika seluruh asumsi dipenuhi, maka status desa tertinggal diklasifikasikan menggunakan Analisis Diskriminan Linier. Jika seluruh asumsi tidak dipenuhi, maka dilakukan klasifikasi data *training* menggunakan Analisis Diskriminan Kernel dengan pendekatan *Fisher*.
iv. Menghitung ketepatan klasifikasi pada data *testing* dan data *training* dengan menggunakan persamaan (2.43), (2.44), (2.45), (2.46), dan (2.49).
v. Melakukan analisis seperti langkah 2.a (vi).

3. Melakukan penanganan kondisi data *imbalanced* dengan metode *Combine Undersampling* dan menentukan metode *classifier* terbaik untuk memprediksi status desa tertinggal di Jawa Timur tahun 2014 menggunakan *software R*, sebagai berikut :
   a. Prosedur dari metode *Tomek Links* adalah bekerja dengan pengecekan setiap data dari kelas yang berbeda. Apabila ditemukan sepasang data yang memiliki kelas label berbeda dan merupakan kasus *Tomek Links*, maka data dari kelas mayoritas akan dihapus dari data *training* sampai menghasilkan data *training* yang bersih dari *noise* dan *borderline*.
   b. Prosedur dari *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.
   c. Melakukan klasifikasi menggunakan data *balanced* dengan metode klasifikasi Regresi Logistik seperti langkah 2.a. dan
   d. Melakukan klasifikasi menggunakan data *balanced* dengan metode klasifikasi Regresi Logistik Ridge seperti langkah 2.b.

e. Melakukan klasifikasi menggunakan data *balanced* dengan metode klasifikasi Analisis Diskriminan Kernel seperti langkah 2.c.

4. Membandingkan efektivitas metode Combine Undersapling pada data menggunakan semua variabel dan variabel signifikan.

Langkah-langkah analisis tersebut dapat digambarkan dalam diagram alir sebagaimana yang ditampilkan pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

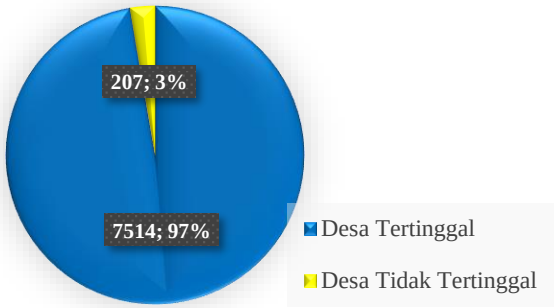
BAB IV

ANALISIS DAN PEMBAHASAN

Analisis dan pembahasan yang akan diuraikan pada bab ini mencakup hasil klasifikasi status desa tertinggal dengan menggunakan *Combine Undersampling (Tomek Links + Random Undersampling)* sebagai metode untuk menyeimbangkan data dan dua metode klasifikasi, yaitu Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel. Klasifikasi yang dihasilkan dari kedua metode ini akan dibandingkan untuk memilih metode yang dianggap terbaik untuk mengklasifikasikan status desa tertinggal dan tidak tertinggal.

4.1 Karakteristik Data Potensi Desa di Provinsi Jawa Timur Tahun 2014

Dalam menentukan klasifikasi status ketertinggalan desa, variabel penelitian yang dipilih mencakup lima dimensi. Berdasarkan Indeks Pembangunan Desa oleh BPS, lima dimensi tersebut terdiri dari pelayanan dasar, kondisi infrastruktur, aksesibilitas/transportasi, pelayanan umum, dan penyelenggaraan pemerintah. Analisis statistika deskriptif ini menggunakan data variabel yang belum dijadikan data rasio. Berikut hasil analisis statistika deskriptif data Potensi Desa Jawa Timur Tahun 2014 :



Gambar 4.1 Diagram lingkaran status desa tertinggal

Gambar 4.1 menunjukkan adanya ketimpangan antara jumlah desa tertinggal terhadap jumlah desa tidak tertinggal. Jumlah desa tertinggal sebanyak 207 desa atau sebesar 3% dari total desa di Jawa Timur sedangkan desa dengan status tidak tertinggal sebanyak 7514 desa atau sebesar 97%. Hal ini berarti rasio antara status desa tertinggal dengan desa tertinggal dapat dikatakan cukup tinggi, yaitu 1 : 32,3. Selain itu, 8 variabel skala numerik dijadikan sebagai indikator dalam mengklasifikasikan status desa tertinggal. Berikut tabel analisis statistika deskriptif dari data jumlah SD/MI, tempat praktik bidan, poskesdes, toko kelontong, rumah tangga pengguna listrik, penderita gizi buruk, pendapatan asli desa, dan jarak ke kantor camat tersebut :

Analisis statistika deskriptif dari masing-masing status ketertinggalan desa terdiri dari desa tertinggal dengan kode {1} dan desa tidak tertinggal dengan kode {0}.

Tabel 4.1 Statistika Deskriptif Masing-Masing Status

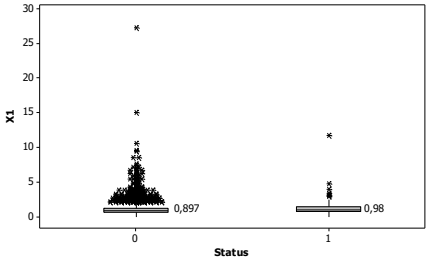
Variabel	Status	Mean	Min	Max
Jumlah SD/MI	0	3,1	0	27
	1	2,96	0	9
Jumlah Tempat Praktik Bidan	0	1,23	0	16
	1	0,74	0	4
Jumlah Poskesdes	0	0,58	0	11
	1	0,49	0	2
Jumlah Toko Kelontong	0	40,45	0	806
	1	15,51	0	61
Jumlah Keluarga Pengguna Listrik	0	1287,7	10	9505
	1	913,9	4	3046
Jarak ke Kantor Camat	0	4,854	1	197
	1	11,048	1	164
Jumlah Penderita Gizi Buruk	0	0,78	0	98
	1	0,65	0	16

Tabel 4.1 Statistika Deskriptif Masing-Masing Status (lanjutan)

Variabel	Status	Mean	Min	Max
Jumlah	0	129,57	0	7106
Pendapatan Asli Desa (jutaan rupiah)	1	31,28	0	1800

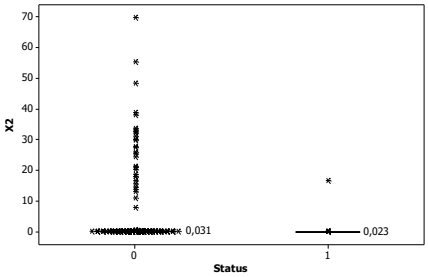
Dari Tabel 4.1 diketahui bahwa rata-rata jumlah SD/MI pada setiap desa tidak tertinggal adalah 3 sekolah dengan jumlah SD/MI terbanyak ada di desa Tlambah, Kabupaten Sampang sebanyak 27 sekolah dan masih ada 39 desa yang belum memiliki SD/MI. Dilihat dari variabel jumlah SD/MI perbedaan dari karakteristik data antara desa tertinggal dengan desa tidak tertinggal adalah dari jumlah SD/MI terbanyak, yaitu ada di desa Pajeruan dan Palenggiyan, Kabupaten Sampang sebanyak 9 sekolah serta rata-rata jumlah SD/MI pada desa tidak tertinggal ada 4 sekolah per desa. Selanjutnya, setiap desa tidak tertinggal rata-rata memiliki 2 tempat praktik desa dan jumlah tempat praktik terbanyak ada di desa Kolor, Kabupaten Sumenep sebanyak 16 tempat praktik bidan. Namun, jika dilihat dari nilai minimum masih ada 432 desa yang belum memiliki tempat praktik bidan. Pada desa tertinggal jumlah tempat praktik bidan terbanyak ada pada desa Larangan Timur, Kabupaten Bangkalan Rata-rata jumlah tempat praktik pada desa tertinggal adalah 1 unit. Rata-rata jumlah poskesdes disetiap desa tidak tertinggal ada 1 unit dengan jumlah terbanyak ada di desa Payudan Dundang, Kabupaten Sumenep sebanyak 11 unit sedangkan nilai minimum menunjukkan 0 atau masih ada 712 desa belum memiliki poskesdes. Selain itu, disetiap desa tertinggal rata-rata tidak memiliki poskesdes dan jumlah poskesdes terbanyak ada di desa Montorna dan Sawah Sumur, Kabupaten Sumenep sebanyak 2 desa. Dilihat dari variabel jumlah toko kelontong pada desa tidak tertinggal, desa Waru Barat di Kabupaten Pamekasan memiliki jumlah toko kelontong terbanyak yaitu 806 toko. Rata-rata jumlah toko kelontong ada 41 toko. Pada desa tertinggal, rata-rata jumlah toko kelontong ada 16 toko dan terbanyak sebanyak 61 toko di desa Kalisari, Kabupaten Situbondo. Selain itu, pada desa

tertinggal masih ada sebanyak 8 desa yang tidak memiliki toko kelontong. Jumlah keluarga pengguna listrik pada desa tidak tertinggal terbanyak sebesar 9505 keluarga di desa Suci, Kabupaten Gresik dan terendah ada di desa Besuki, Kabupaten Sidoarjo sebanyak 10 keluarga. Rata-rata jumlah keluarga pengguna listrik disetiap desa sebanyak 1288 keluarga. Kemudian, jumlah keluarga pengguna listrik pada desa tertinggal terbanyak sebesar 3046 keluarga di desa Tanjung, Kabupaten Pamekasan dan terendah ada di desa Kedung Bendo, Kabupaten Sidoarjo sebanyak 4 keluarga. Rata-rata jumlah keluarga pengguna listrik disetiap desa sebanyak 914 keluarga. Rata-rata jarak dari setiap desa pada desa tidak tertinggal ke kantor camat sejauh 4,85 km dengan jarak terjauh ditempuh desa Karamian, Kabupaten Sumenep yaitu 164 km dan jarak terdekat sejauh 1 km. Pada desa tertinggal jarak terjauh harus ditempuh sejauh 164 km dari desa Masakambing, Kabupaten Sumenep dan jarak terdekat minimal harus menempuh 1 km sedangkan rata-rata jarak desa ke kantor camat yaitu 11,05 km. Selanjutnya, jumlah penderita gizi buruk terbanyak pada desa tidak tertinggal ada 98 orang di desa Sapeken, Kabupaten Sumenep. Rata-rata penderita gizi buruk pada desa tidak tertinggal adalah 1 orang. Sama halnya dengan desa tertinggal, beberapa desa sudah tidak memiliki penderita gizi buruk. Jumlah penderita gizi buruk terbanyak ada di desa Kembangsari, Kabupaten Situbondo sebanyak 16 orang. Jumlah PAD terbesar pada desa tidak tertinggal ada di desa Winong, Kabupaten Tulungagung sebesar Rp7.160.000.000,- sedangkan beberapa desa ada yang tidak memiliki PAD dan rata-rata pendapatan setiap desa sebesar Rp129.570.000,-. Pada desa tertinggal, jumlah PAD terbesar ada di desa Ngadas, Kabupaten Malang sebesar Rp1.800.000.000,- dan beberapa desa ada yang tidak memiliki PAD sedangkan rata-rata jumlah PAD disetiap desa adalah Rp31.280.000,-. Selain itu, karakteristik status desa tertinggal dapat dilihat dari nilai rasio masing-masing variabel menggunakan *boxplot*. Berikut adalah *boxplot* dari rasio banyaknya jumlah SD/MI terhadap total murid SD/MI :



Gambar 4.2 Boxplot Rasio Banyaknya SD/MI

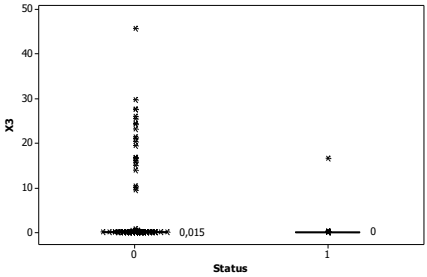
Nilai median pada variabel rasio banyaknya SD/MI terhadap total murid/SD antara desa berstatus tertinggal (1) sebesar 0,98 dan tidak tertinggal (0) sebesar 0,897 menunjukkan bahwa selisih dari kedua nilai median yang kecil. Selain itu, nilai rata-rata pada masing-masing status desa tersebut berkisar antara nilai 1 sampai 2 serta nilai *variance* pada desa tertinggal lebih besar dari desa tidak tertinggal. Selanjutnya, *boxplot* pada rasio banyaknya tempat praktik bidan terhadap total penduduk.



Gambar 4.3 Boxplot Rasio Banyaknya Tempat Praktik Bidan

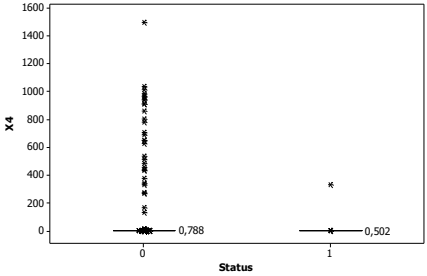
Gambar 4.3 menunjukkan bahwa nilai rata-rata pada desa berstatus desa tidak tertinggal dan desa desa tertinggal sangat kecil yaitu berkisar antara nilai 0-1 sedangkan nilai median dan standar deviasi menunjukkan bahwa keragaman dari desa tidak tertinggal lebih besar dibandingkan desa tertinggal. Kemudian dapat dilihat

boxplot dari rasio banyaknya jumlah poskesdes terhadap penduduk pada Gambar 4.4.



Gambar 4.4 Boxplot Rasio Banyaknya Poskesdes

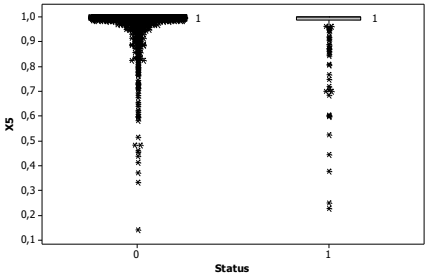
Pada rasio jumlah banyaknya poskesdes terhadap jumlah penduduk menunjukan nilai median dari desa berstatus tidak tertinggal lebih besar yaitu 0,015 sedangkan median dari desa berstatus tertinggal adalah 0. Selain itu, nilai standar deviasi dari desa tidak tertinggi lebih tinggi sehingga mengindikasikan keragaman data tersebut besar dibandingkan desa tertinggal. Selanjutnya *boxplot* dari variabel X_4 pada gambar 4.5.



Gambar 4.5 Boxplot Rasio Banyaknya Toko Kelontong

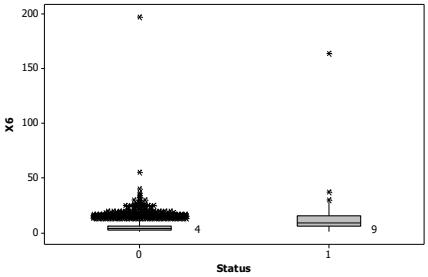
Berdasarkan Gambar 4.5 dapat dilihat bahwa nilai median pada desa tidak tertinggal lebih tinggi dibandingkan desa tertinggal dengan nilai median berurutan adalah 0,788 dan 0,502 serta data pada desa tidak tertinggal lebih beragam karena memiliki nilai standar deviasi lebih tinggi. Selain itu, nilai rata-rata dari variabel rasio jumlah banyaknya toko terhadap total penduduk sangat kecil.

Selain itu, rasio jumlah banyaknya rumah tangga pengguna listrik terhadap total rumah tangga dapat dilihat dari *boxplot* pada Gambar 4.6.



Gambar 4.6 Boxplot Rasio Banyaknya Keluarga Pengguna Listrik

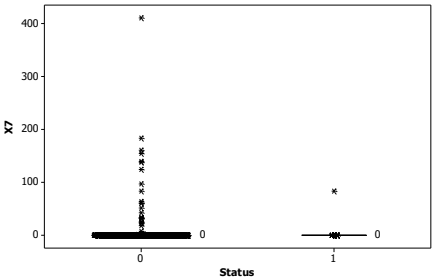
Gambar 4.6 menunjukkan bahwa nilai median dari desa berstatus desa tertinggal dan desa tidak tertinggal sama, yaitu 1 sedangkan nilai rata-rata pada setiap status mendekati nilai 1. Namun, data status desa tertinggal pada variabel X_6 memiliki variabilitas yang tinggi dibandingkan data desa tidak tertinggal. Kemudian dapat dilihat *boxplot* dari variabel jarak desa ke kantor camat per km pada Gambar 4.7.



Gambar 4.7 Boxplot Jarak Tempuh Ke Kantor Camat

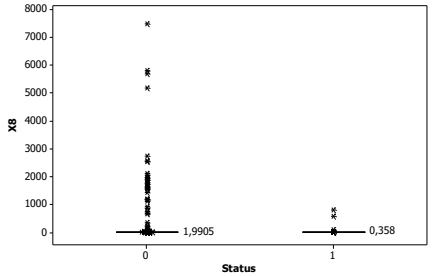
Nilai *variance* menunjukkan keragaman data dari desa berstatus tertinggal lebih besar dari desa tidak tertinggal. Tidak berbeda dengan nilai rata-rata, nilai tengah atau median pada desa tertinggal juga lebih tinggi, yaitu 9 km dan nilai median pada desa tidak tertinggal adalah 4 km. Hal ini membuktikan bahwa semakin

jauh jarak yang ditempuh dari desa ke kantor camat (km), maka desa tersebut berpeluang masuk ke dalam kelompok desa tertinggal Selanjutnya, *boxplot* variabel dari rasio jumlah penderita gizi buruk terhadap total penduduk pada Gambar 4.8.



Gambar 4.8 Boxplot Rasio Banyaknya Penderita Gizi Buruk

Nilai median dari desa tidak tertinggal dan desa tertinggal adalah 0. Selain tu, nilai rata-rata dari setiap status desa mendekati nol, maka dapat dikatakan rata-rata penderita gizi buruk cenderung sedikit untuk disetiap desa di Jawa Timur. Kemudian, dapat dilihat pula boxplot dari variabel rasio jumlah banyaknya pendapatan asli desa terhadap jumlah penduduk pada Gambar 4.9.



Gambar 4.9 Boxplot Rasio Pendapatan Asli Desa

Pada Gambar 4.9 diketahui bahwa nilai median dan desa tidak tertinggal lebih besar, yaitu 1,99 dan median desa tertinggal adalah 0,358. Nilai rata-rata dari kedua status hampir bernilai sama sedangkan keragaman data dari desa tidak tertinggal lebih besar.

4.2 Analisis Klasifikasi Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel pada Data *Imbalanced*

Data *imbalanced* status desa tertinggal di Jawa Timur tahun 2014 akan dianalisis menggunakan 3 metode *classifier*, yaitu : Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel pada subbab selanjutnya.

4.2.1 Regresi Logistik pada Data *Imbalanced*

Regresi Logistik merupakan salah satu metode yang digunakan untuk pengklasifikasian status desa tertinggal dengan komposisi data variabel respon yaitu 7514 desa tidak tertinggal dan 207 desa tertinggal.

A. Regresi Logistik pada Semua Variabel

Pada penelitian ini data yang digunakan terdiri dari data menggunakan semua variabel dan variabel signifikan. Kedua data akan dianalisis menggunakan metode klasifikasi Regresi Logistik.

i. Deteksi Multikolinearitas pada Semua Variabel

Deteksi asumsi multikolinearitas harus dilakukan untuk melihat apakah ada hubungan korelasi yang tinggi antar variabel prediktor. Kasus multikolinearitas perlu diatasi karena akan berdampak pada variansi dari hasil estimasi dan standar *error*. Deteksi multikolinearitas dapat dilihat dari nilai *variance inflation factor* yang dapat dilihat pada tabel berikut ini :

Tabel 4.2 Deteksi Data Mutikolinearitas (VIF) pada Data *Imbalanced* Semua Variabel

Variabel	VIF
X_1	1,008
X_2	**4,102**
X_3	**3,689**
X_4	**4,002**
X_5	1,027
X_6	1,035
X_7	**3.021**
X_8	**3,321**

Tabel 4.2 menunjukkan bahwa tidak ada kasus multikolinearitas pada data *imbalanced* dari indikator-indikator status desa tertinggal Provinsi Jawa Timur rahun 2014 karena nilai *variance inflation factor* (*VIF*) yaitu kurang dari 5.

ii. Ketepatan Klasifikasi Regresi Logistik pada Semua Variabel

Setelah mendeteksi kasus multikolinearitas, maka dapat dilanjukan pengklasifikasian menggunakan Regresi Logistik.

Tabel 4.3 Hasil Ketepatan Klasifikasi Regresi Logistik pada Data Imbalanced Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,519	0,515	0,023
Rata-rata G-Mean	0,198	0,113	0,152
Rata-rata Akurasi Total	0,971	0,972	0,003
Rata-rata Sensitivitas	0,041	0,034	0,051
Rata-rata Spesitifitas	0,998	0,997	0,002

Dari Tabel 4.3 dapat diketahui rata-rata ketepatan klasifikasi tertinggi ada pada nilai AUC, yaitu 0,515 sedangkan nilai rata-rata G-Mean bernilai rendah sebesar 0,113. Rata-rata nilai total akurasi yang dimiliki adalah 0,972 sehingga menunjukkan klasifikasi status desa tertinggal di Jawa Timur pada data *imbalanced* menggunakan metode analisis Regresi Logistik menghasilkan nilai rata-rata akurasi total yang sangat tinggi, namun pada data *imbalanced* nilai akurasi tidak dapat dijadikan tolak ukur untuk ketepatan klasifikasi. Selanjutnya, nilai sensitivitas sebesar 0,034 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah, namun berbanding terbalik dengan ketepatan klasifikasi untuk data negatif atau kelas mayoritas, yaitu 0,997. Berdasarkan hasil ketepatan klasifikasi dari Regresi Logistik dan Regresi Logistik Ridge dapat diketahui bahwa data *imbalanced* menghasilkan performansi yang kurang baik, namun keragaman yang dihasilkan dari rata-rata ketepatan klasifikasi menggunakan *10-fold CV* rendah sehingga performasi yang dihasilkan stabil.

B. Regresi Logistik pada Variabel Signifikan

Setelah melakukan analisis menggunakan semua variabel, maka langkah selanjutnya adalah menguji variabel signifikan sehingga dapat dilakukan pengujian parameter terlebih dahulu.

i. Pengujian Parameter

Pengujian parameter secara serentak dilakukan dengan menggunakan *Likelihood Ratio Test* menggunakan $\alpha = 0{,}10$. Hasil uji serentak menunjukkan bahwa H_0 ditolak karena nilai $G > \chi^2_{(8;0,10)}$, yaitu 1694,4 > 13,36 sehingga minimal ada satu variabel yang berpengaruh signifikan. Berikut hasil dari pengujian parsial pada Tabel 4.4 :

Tabel 4.4 Nilai Koefisien, t-hitung, dan P-value Hasil Uji Serentak pada Data Imbalanced Semua Variabel

Variabel	Koefisien	*Standar Error*	Statistik Hitung	*P-Value*
Konstan	0,780	0,866	0,901	0,368
X_1	0,067	0,066	1,021	0,307
X_2	0,215	0,196	1,099	0,272
X_3	-0,023	0,092	-0,257	0,798
X_4	**-5,419**	**0,861**	**-6,295**	**0,000**
X_5	-0,0003	0,006	-0,567	0,571
X_6	**0,138**	**0,012**	**11,142**	**0,000**
X_7	0,018	0,029	0,597	0,550
X_8	-0,002	0,003	-0,874	0,382

Berdasarkan Tabel 4.4, diketahui ada dua variabel signifikan karena memiliki p-*value*<(α=0,10) yaitu variabel rasio keluarga pengguna listrik terhadap total penduduk (X_4) dan jarak tempuh per kilometer ke kantor camat (X_6). Pada model tersebut masih terdapat beberapa variabel yang tidak signifikan pada α=0,10.

Selanjutnya, untuk menemukan variabel signifikan digunakan metode *Backward*. *Backward elimination* digunakan untuk memilih variabel yang berpengaruh signifikan dengan cara mengeluarkan variabel yang paling tidak signifikan secara bertahap. Hasil akhir yang didapat dari *backward elimination* terdapat pada Tabel 4.6.

Tabel 4.5 Nilai Koefisien, Thitung, dan P-value Hasil Backward Elimination Pada Data Imbalanced Semua Variabel

Variabel	Koefisien	*Standar Error*	Statistik Hitung	*P-Value*
Konstan	0,839	0,864	0,972	0,331
X_4	-5,414	0,859	-6,299	0,000
X_6	0,138	0,012	11,227	0,000

Berdasarkan Tabel 4.5, diketahui ada dua variabel signifikan yaitu variabel rasio banyaknya toko/warung kelontong terhadap total penduduk dan jarak tempuh per kilometer ke kantor camat. Selanjutnya dapat dibuat model Regresi Logistik menggunakan dua variabel signifikan.

$$ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = 0{,}839 - 5{,}414\,X_4 + 0{,}138\,X_6$$

Model yang didapat memberikan informasi bahwa jika variabel rasio banyaknya toko/warung kelontong terhadap jumlah penduduk bertambah satu satuan, maka peluang status desa menjadi desa tertinggal akan berkurang menjadi sebesar $\exp(-5{,}414) = 0{,}005$ kali atau *odds ratio* status desa menjadi desa tidak tertinggal sebesar $0{,}005^{-1} = 200$ kali. Informasi lain yang dapat diketahui dari model adalah jika jarak tempuh per kilometer ke kantor camat bertambah 1 km, maka peluang status desa menjadi tertinggal akan naik, dimana *odds ratio* status desa tertinggal menjadi sebesar $\exp(0{,}138) = 1{,}15$ kali atau *odds ratio* status desa menjadi desa tidak tertinggal sebesar 0,869 kali.

ii. Deteksi Multikolinearitas pada Data *Imbalanced*

Deteksi asumsi multikolinearitas harus dilakukan untuk melihat apakah ada hubungan korelasi yang tinggi antar variabel prediktor. Kasus multikolinearitas perlu diatasi karena akan berdampak pada variansi dan standar *error* sehingga perlu dilakukan deteksi untuk mengetahui adanya korelasi yang tinggi atau rendah antar variabel prediktor dengan melihat nilai *VIF*. Jika kasus multikolinearitas terdeteksi, maka akan dilakukan penanganan dengan metode tertentu. Deteksi multikolinearitas dapat dilihat dari nilai *variance inflation factor* yang dapat dilihat pada subbab selanjutnya.

Tabel 4.6 Deteksi Data Multikolinearitas (VIF) pada Data Imbalanced Variabel Signifikan

Variabel	VIF
X_4	1,027
X_6	1,027

Tabel 4.6 menunjukkan bahwa tidak adanya kasus multikolinearitas pada data *imbalanced* dari indikator-indikator status desa tertinggal Provinsi Jawa Timur rahun 2014 dengan menggunakan variabel yang signifikan.

iii. Ketepatan Klasifikasi Regresi Logistik pada Data *Imbalanced*

Klasifikasi Regresi Logistik dilakukan untuk melihat ketepatan klasifikasi dari data *imbalanced* yang masih memiliki permasalahan multikolinearitas. Data akan dibagi menjadi data *training* dan data *testing* menggunakan *stratified 10-fold cross validation.* Berikut nilai ketepatan klasifikasi Regresi Logistik status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.7 :

Tabel 4.7 Hasil Ketepatan Klasifikasi Regresi Logistik pada Data Imbalanced Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,516	0,514	0,023
Rata-rata G-Mean	0,181	0,104	0,141
Rata-rata Akurasi Total	0,973	0,972	0,002
Rata-rata Sensitivitas	0,033	0,029	0,046
Rata-rata Spesitifitas	0,998	0,999	0,002

Berdasarkan Tabel 4.7 didapatkan rata-rata AUC, G-Mean, dan akurasi yang rendah, yaitu 0,514, 0,104, dan 0,972. Selanjutnya, nilai senstivitas sebesar 0,029 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah dan tidak sebanding dengan ketepatan klasifikasi untuk data negatif atau kelas mayoritas, yaitu 0,999. Berdasarkan Tabel 4.7 dapat dikatakan ketepatan klasifikasi dengan metode klasifikasi Regresi Logistik pada data *imbalanced* sangat rendah karena banyak kesalahan pengklasifikasian pada data kelas minoritas atau dengan kata lain kelas data menjadi tidak stabil karena data akan

lebih condong ke bagian data yang memiliki komposisi data yang lebih besar, yaitu desa tidak tertinggal.

4.2.2 Ketepatan Klasifikasi Regresi Logistik Ridge pada Data *Imbalanced*

Pada penelitian ini menggunakan data semua variabel agar selanjutnya dapat dibandingkan efektifitas performansi *classifier* dengan data yang hanya menggunakan variabel signifikan.

A. Regresi Logistik Ridge pada Semua Variabel

Berdasarkan hasil pengujian asumsi terdapat kasus multikolinearitas pada data *imbalanced.* Selanjutnya, dengan menggunakan klasifikasi Regresi Logistik Ridge model dari data *training* digunakan untuk mengklasifikasikan data *testing* dan dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.8.

Tabel 4.8 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge pada Data Imbalanced Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,509	0,507	0,016
Rata-rata G-Mean	0,135	0,053	0,113
Rata-rata Akurasi Total	0,973	0,973	0,001
Rata-rata Sensitivitas	0,019	0,014	0,032
Rata-rata Spesitifitas	0,999	0,999	0,001

Dari Tabel 4.8 dapat diketahui rata-rata nilai AUC lebih tinggi dibandingkan nilai rata-rata G-Mean, yaitu 0,507. Rata-rata nilai akurasi total yang dimiliki adalah 0,973 sehingga menunjukkan klasifikasi status desa tertinggal di Jawa Timur pada data *imbalanced* menggunakan metode analisis Regresi Logistik Ridge menghasilkan nilai akurasi yang sangat tinggi, namun pada data *imbalanced* ketepatan klasifikasi tidak dapat diukur dari nilai akurasi. Selanjutnya, nilai senstivitas sebesar 0,014 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah, namun berbanding terbalik dengan ketepatan klasifikasi untuk data negatif atau kelas mayoritas, yaitu 0,999. Berdasarkan Tabel 4.8 dapat dikatakan ketepatan klasifikasi

dengan metode klasifikasi Regresi Logistik pada data *imbalanced* rendah karena karena data akan lebih condong ke bagian data yang memiliki komposisi data yang lebih besar, yaitu desa tidak tertinggal dan tingkat variabilitas hasil ketepatan klasifikasi juga rendah.

B. Regresi Logistik Ridge pada Variabel Signifikan

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan. Berikut hasil ketepatan klasifikasi dari Regresi Logistik Ridge menggunakan variabel signifikan :

Berdasarkan hasil pengujian asumsi terdapat kasus multikolinearitas pada data *imbalanced.* Selanjutnya, dengan menggunakan klasifikasi Regresi Logistik Ridge model dari data *training* digunakan untuk mengklasifikasikan data *testing* dan dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.9.

Tabel 4.9 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge pada Data Imbalanced Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,516	0,516	0,03
Rata-rata G-Mean	0,183	0,11	0,155
Rata-rata Akurasi Total	0,972	0,972	0,002
Rata-rata Sensitivitas	0,034	0,034	0,06
Rata-rata Spesitifitas	0,998	0,998	0,002

Dari Tabel 4.9 dapat diketahui rata-rata nilai AUC lebih tinggi dibandingkan nilai rata-rata G-Mean, yaitu 0,516. Rata-rata nilai akurasi total yang dimiliki adalah 0,972 sehingga menunjukkan klasifikasi status desa tertinggal di Jawa Timur pada data *imbalanced* menggunakan metode analisis Regresi Logistik Ridge menghasilkan nilai akurasi yang sangat tinggi, namun pada data *imbalanced* ketepatan klasifikasi tidak dapat diukur dari nilai akurasi. Selanjutnya, nilai senstivitas sebesar 0,034 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah, namun berbanding terbalik dengan ketepatan

klasifikasi untuk data negatif atau kelas mayoritas, yaitu 0,998. Berdasarkan Tabel 4.9 dapat dikatakan ketepatan klasifikasi dengan metode klasifikasi Regresi Logistik pada data *imbalanced* rendah karena data diklasifikasikan pada kelas mayoritas serta nilai standar deviasi menunjukkan hasil ketepatan klasifikasi yang stabil.

4.2.3 Analisis Diskriminan Kernel pada Data *Imbalanced*

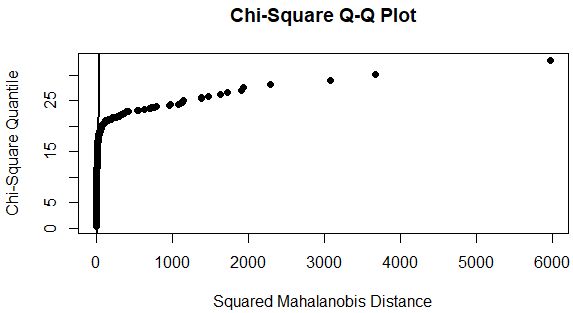
Langkah awal yang harus dilakukan dalam menentukan Analisis Diskriminan Kernel dapat digunakan atau tidak adalah melakukan pengujian apakah asumsi telah dipenuhi atau belum. Asumsi yang harus dipenuhi dalam analisis diskriminan adalah matriks varians kovarians yang bersifat homogen, variabel indikator status desa tertinggal berdistribusi normal multivariat. Apabila data menunjukkan tidak berdistribusi normal multivariat, maka Analisis Diskriminan Kernel dapat digunakan sebagai metode klasifikasi status ketertinggalan desa di Provinsi Jawa Timur tahun 2014.

A. Analisis Diskriminan Kernel pada Semua Variabel

Pada penelitian ini menggunakan data semua variabel agar selanjutnya dapat dibandingkan efektifitas performansi *classifier* dengan data yang hanya menggunakan variabel signifikan. Berikut langkah-langkah pada Analisis Disriminan Kernel dengan menggunakan semua variabel :

i. Hasil Uji Distribusi Normal Multivariat pada Semua Variabel

M*ardia's test* dilakukan untuk menguji apakah data berdistribusi normal multivariat atau tidak. Hasil dari uji normalitas menghasilkan *p-value* yang bernilai 0. *P-value* yang bernilai kurang dari $\alpha = 0{,}05$ menunjukkan bahwa data indikator status desa tertinggal tidak berdistribusi normal multivariat. Oleh karena itu, asumsi distribusi normal multivariate tidak dapat terpenuhi. Gambar 4.10 menunjukkan *qq-plot* dari data indikator status desa tertinggal Provinsi Jawa Timur yang belum seimbang.



Gambar 4.10 Chi-Squared QQ-Plot Data Imbalanced Semua Variabel

Titik-titik hitam pada Gambar 4.10 merepresentasikan data indikator status desa tertinggal Provinsi Jawa Timur secara multivariat. Banyaknya titik-titik yang tersebar jauh dari garis hitam mengindikasikan bahwa data tidak berdistribusi normal multivariat.

ii. Hasil Uji Homogenitas pada Semua Variabel

Perhitungan statistik uji dilakukan berdasarkan persamaan (2.34) yang dijelaskan pada Bab II Tinjauan Pustaka mengenai uji homogenitas. Nilai *p-value* pada data *imbalanced* bernilai kecil dari $\alpha = 0{,}05$ yaitu 0 sehingga menyebabkan H_0 ditolak. Artinya, matriks varians kovarians pada data indikator status desa tertinggal Provinsi Jawa Timur tahun 2014 tidak homogen sehingga asumsi homogenitas tidak terpenuhi.

iii. Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Semua Variabel

Berdasarkan hasil pengujian asumsi diskriminan linear, terdapat dua asumsi yang belum terpenuhi, yaitu uji homogenitas dan uji normal multivariat. Uji homogenitas menunjukkan matriks varians kovarians tidak homogen dan data *imbalanced* tidak berdistribusi normal multivariat. Hal ini menunjukkan metode klasifikasi analisis diskriminan linear tidak tepat untuk mengklasifikasikan data indikator status desa tertinggal di Provinsi Jawa Timur tahun 2014 sehingga metode klasifikasi yang digunakan adalah klasifikasi Analisis Diskriminan Kernel.

Selanjutnya dengan menggunakan model dari data *training* untuk mengklasifikasikan data *testing*, dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.10.

Tabel 4.10 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Data Imbalanced Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,637	0,622	0,036
Rata-rata G-Mean	0,538	0,504	0,079
Rata-rata Akurasi Total	0,960	0,959	0,006
Rata-rata Sensitivitas	0,296	0,266	0,072
Rata-rata Spesitifitas	0,978	0,979	0,005

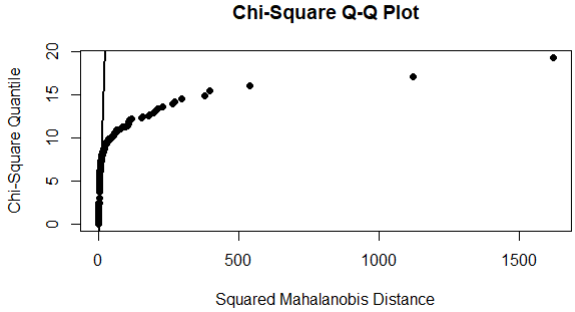
Nilai rata-rata ketepatan klasifikasi *G-Mean* dan AUC secara berurutan yaitu 0.504 dan 0.622 cenderung kecil sedangkan rata-rata akurasi total menunjukkan nilai yang sangat besar yaitu 0,959. Hal ini berarti bahwa klasifikasi status desa tertinggal di Jawa Timur dengan metode Analisis Diskriminan Kernel menghasilkan klasifikasi cukup baik jika dilihat dari nilai AUC. Namun, memiliki nilai senstivitas sebesar 0,266 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah. Selain itu, nilai spesifisitas menunjukkan ketepatan klasifikasi untuk data negatif atau kelas mayoritas sangat besar, yaitu 0.979. Berdasarkan nilai standar deviasi dapat disimpulkan bahwa keragaman dari ketepatan klasifikasi sangat kecil.

B. Analisis Diskriminan Kernel pada Variabel Signifikan

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan.

i. Hasil Uji Distribusi Normal Multivariat pada Variabel Signifikan

Hasil dari uji normalitas menghasilkan *p-value* yang bernilai 0. *P-value* yang bernilai kurang dari $\alpha = 0{,}05$ menunjukkan bahwa data indikator status desa tertinggal tidak berdistribusi normal multivariat. Oleh karena itu, asumsi distribusi normal multivariate tidak dapat terpenuhi.



Gambar 4.11 Chi-Squared QQ-Plot Data Imbalanced Variabel Signifikan

Titik-titik hitam pada Gambar 4.11 merepresentasikan data indikator status desa tertinggal Provinsi Jawa Timur secara multivariat. Banyaknya titik-titik yang tersebar jauh dari garis hitam mengindikasikan bahwa data tidak berdistribusi normal multivariat.

ii. Hasil Uji Homogenitas pada pada Variabel Signifikan

Nilai *p-value* pada data *imbalanced* bernilai kecil dari α = 0,05 yaitu 0 sehingga menyebabkan H_0 ditolak sehingga asumsi homogenitas tidak terpenuhi pada data indikator status desa tertinggal Provinsi Jawa Timur tahun 2014.

iii. Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Variabel Signifikan

Berikut hasil rata-rata ketepatanklasifikasi dengan menggunakan model dari data *training* untuk mengklasifikasikan data *testing*, dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.11.

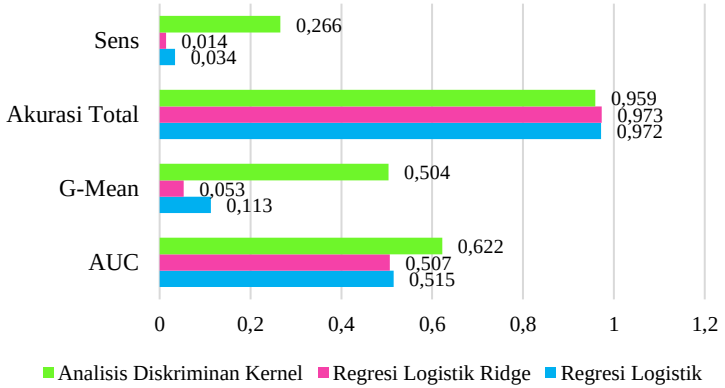
Tabel 4.11 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,613	0,608	0,043
Rata-rata G-Mean	0,495	0,475	0,104
Rata-rata Akurasi Total	0,954	0,954	0,006
Rata-rata Sensitivitas	0.253	0.242	0,086
Rata-rata Spesitifitas	0.974	0.974	0,006

Nilai rata-rata ketepatan klasifikasi AUC dan G-Mean memiliki nilai yang tidak jauh berbeda, namun rata-rata akurasi memiliki nilai yang sangat besar. Lalu, rata-rata nilai senstivitas sebesar 0,242 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas sangat rendah. Selain itu, nilai spesifisitas menunjukkan ketepatan klasifikasi untuk data negatif atau kelas mayoritas sangat besar, yaitu 0,974 dan rata-rata hasil ketepatan klasifikasi memiliki tingkat variabilitas yang rendah karena standar deviasi yang kecil. Berdasarkan nilai AUC pada Tabel 4.11 dapat dikatakan ketepatan klasifikasi dengan menggunakan metode klasifikasi Analisis Diskriminan Kernel cukup baik.

4.2.4 Analisis Gabungan Pada Data *Imbalanced* Semua Variabel dan Variabel Signifikan

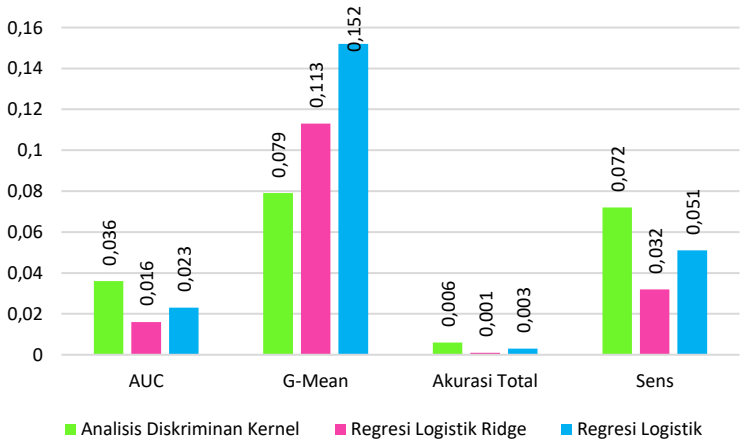
Sebelumnya telah dilakukan klasifikasi status ketertinggalan desa dengan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan semua variabel maupun menggunakan variabel yang signifikan. Berikut pada Gambar 4.12 adalah perbandingan ketiga metode dengan semua variabel.



Gambar 4.12 Perbandingan Performansi pada Data Imbalanced Semua Variabel

Gambar 4.12 menunjukkan bahwa metode Analisis Diskriminan Kernel memiliki performansi tertinggi kecuali pada

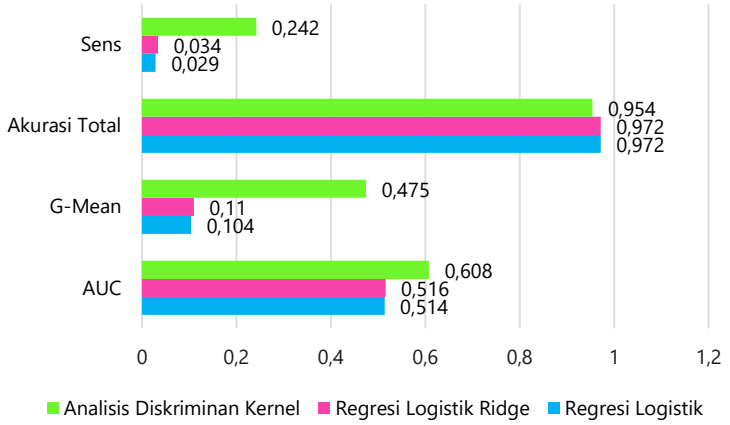
rata-rata akurasi. Namun, rata-rata akurasi pada data *imbalanced* bersifat semu sehingga tidak dapat dijadikan tolak ukur untuk ketepatan klasifikasi karena *classifier* cenderung mengklasifikasikan data pada kelas mayoritas berbeda dengan AUC dan G-Mean yang *robust* terhadap data *imbalanced*. Sementara itu, performansi dari klasifikasi Regresi Logistik lebih bagus dari metode Regresi Logistik, meskipun selisih dari ketepatan klasifikasi yang dihasilkan tidak jauh berbeda. Selanjutnya, dapat dilihat perbandingan nilai standar deviasi pada Gambar 4.13.



Gambar 4.13 Perbandingan Standar Deviasi Data Imbalanced SemuaVariabel

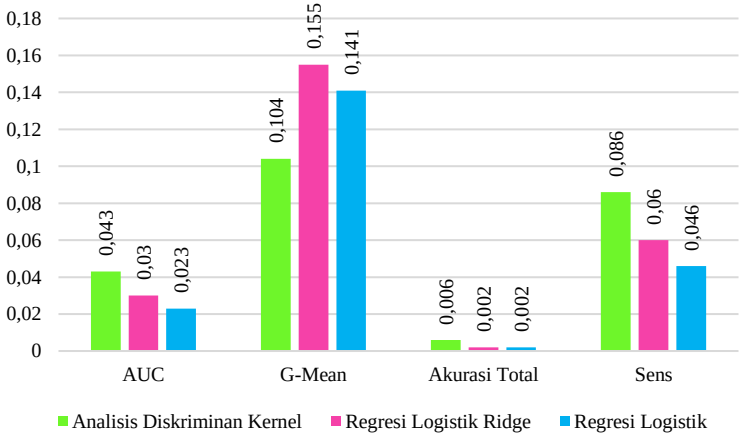
Berdasarkan Gambar 4.13 dapat dilihat bahwa nilai standar deviasi dari masing-masing ketepatan klasifikasi pada Regresi Logistik Ridge cenderung paling rendah. Namun, secara keseluruhan menunjukkan bahwa standar deviasi dari ketiga *classifier* bernilai rendah sehingga variabilitas dari hasil *10 fold – cross validation* kecil, maka ketepatan klasikasi yang dihasilkan dapat dikatakan stabil. Setelah itu, dapat dilihat perbandingan dari

efektivitas performansi pada data *imbalanced* variabel signifikan pada Gambar 4.14.



Gambar 4.14 Perbandingan Ketepatan Klasifikasi Data Imbalanced Variabel Signifikan

Dari Gambar 4.14 dapat dilihat nilai rata-rata ketepatan klasifikasi cenderung tinggi dengan menggunakan metode Analisis Diskriminan Kernel. Nilai rata-rata akurasi pada data *imbalanced* tidak dapat dijadikan tolak ukur untuk menilai efektifitas performansi suatu *classifier* karena hasilnya berkemungkinan bias dan nilai rata-rata akurasi terendah merupakan hasil Analisis Diskriminan Kernel sehingga tidak dapat dikatakan metode ini lebih buruk dari *classifier* lainnya. Namun, ketepatan klasifikasi yang diperoleh dengan klasifikasi Analisis Diskriminan Kernel mengalami penurunan ketika menggunakan variabel signifikan sedangkan ketepatan klasifikasi yang diperoleh dari Regresi Logistik Ridge mengalami kenaikan. Selain itu, nilai rata-rata AUC ketiga *classifier* tidak memiliki selisih yang terlalu signifikan, namun perbedaan selisih yang jauh ada pada nilai rata-rata G-Mean dan sensitivitas. Selanjutnya, dapat dilihat perbandingan nilai standar deviasi pada Gambar 4.15.



Gambar 4.15 Perbandingan Standar Deviasi Data Imbalanced Variabel Signifikan

Berdasarkan Gambar 4.15 dapat dilihat bahwa nilai standar deviasi dari masing-masing ketepatan klasifikasi pada Analisis Diskriminan Kernel cenderung lebih kecil dibandingkan dengan kedua *classifier* lainnya sehingga ketepatan klasifikasinya dapat dikatakan lebih stabil karena memiliki variabilitas data yang rendah.

4.3 Analisis Klasifikasi Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel pada Data *Balanced*

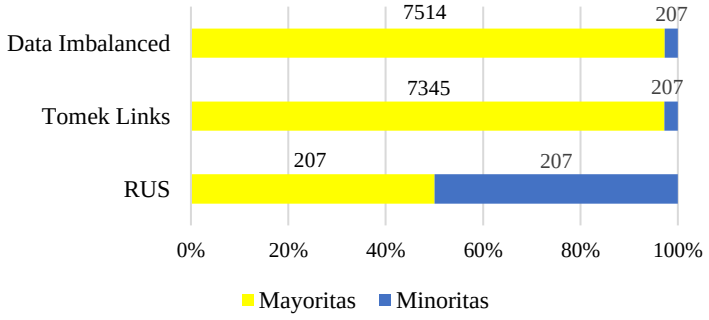
Setelah menganalisis data *imbalanced* menggunakan 3 metode *classifier*, analisis selanjutnya adalah melakukan *resampling* data *imblanced* agar menjadi seimbang atau *balanced*. Sama seperti data *imbalanced*, data status desa tertinggal di Jawa Timur tahun 2014 yang telah seimbang akan dianalisis menggunakan 3 metode *classifier* yang telah dipilih.

4.3.1 Metode *Combine Undersampling*

Kasus status desa tertinggal di Provinsi Jawa Timur merupakan salah satu contoh dari data *imbalanced* karena variabel

respon pada penelitian ini memiliki komposisi jumlah desa tertinggal yang jauh lebih sedikit dibandingkan jumlah desa tidak tertinggal. Data yang *imbalanced* akan berdampak pada hasil ketepatan klasifikasi yang cenderung lebih buruk dibandingkan data *balanced*. Oleh karena itu, dilakukan *resampling* data dengan metode *combine undersampling* untuk meningkatkan ketepatan akurasi disetiap metode *classifier* yang digunakan.

Metode *combine undersampling* merupakan perpaduan antara *Tomek Links* dengan *Random Undersampling*. Penggunaan kedua metode dilakukan secara berurutan. Langkah awal adalah menggunakan *Tomek Links* pada data asli atau *imbalanced*, dimana metode ini akan menghapus data *noise* atau data *borderline* pada data mayoritas. Setelah melakukan pembersihan data dengan *Tomek Links*, komposisi data asli berubah menjadi 7345 untuk desa tidak tertinggal sehingga data masih menunjukkan komposisi yang tidak seimbang. Langkah selanjutnya adalah data hasil dari *Tomek Links* akan di*resampling* menggunakan *Random Undersampling*, dimana metode tersebut akan melakukan penghapusan secara *random* pada data mayoritas sehingga data menjadi *balanced* dan setiap status desa tertinggal maupun tidak tertinggal berjumlah 207 desa. Berikut dapat dilihat perbandingan komposisi data sebelum dan sesudah dilakukan *resampling* data pada Gambar 4.16 :



Gambar 4.16 Perbandingan Komposisi Data Sebelum dan Sesudah Dilakukan Resampling

4.3.2 Klasifikasi Regresi Logistik pada Data *Balanced*

Regresi Logistik merupakan salah satu metode yang digunakan untuk pengklasifikasian status desa tertinggal di Jawa Timur berdasarknan data PODES tahun 2014, dimana pada metode ini data yang seimbang akan menghasilkan ketepatan klasifikasi yang lebih tinggi dibandingkan data *imbalanced*. Pada analisis ini digunakan 8 variabel prediktor dengan komposisi data variabel respon yang telah seimbang yaitu 207 desa tidak tertinggal dan 207 desa tertinggal.

A. Regresi Logistik pada Semua Variabel

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan. Langkah-langkah yang dilakukan sama seperti klasifikasi Regresi Logistik pada data *imbalanced*.

i. Deteksi Multikolinearitas pada Semua Variabel

Deteksi asumsi multikolinearitas pada data *balanced* juga harus dilakukan untuk melihat apakah ada hubungan korelasi yang tinggi antar variabel prediktor setelah dilakukan *Combine Undersampling* pada data *imbalanced*. Deteksi multikolineritas dapat dilihat dari nilai *variance inflation factor* yang dapat dilihat pada tabel berikut ini :

Tabel 4.12 Deteksi Data Mutikolinearitas (VIF) pada Data Balanced Semua Variabel

Variabel	VIF
X_1	1,027
X_2	**1.657,59**
X_3	**1.299,58**
X_4	**1.234,77**
X_5	1,07
X_6	1,07
X_7	**242,14**
X_8	**5,66**

Tabel 4.12 menunjukkan bahwa terdapat 5 variabel yang memiliki nilai VIF diatas 5, yaitu rasio banyaknya praktik bidan terhadap total penduduk (X_2), rasio banyaknya poskesdes terhadap

total penduduk (X_3), rasio banyaknya toko kelontong tehadap total penduduk (X_4), rasio banyaknya penderita gizi buruk terhadap total penduduk (X_7), rasio banyaknya pendapatan asli desa tehadap total penduduk (X_8).

ii. Ketepatan Klasifikasi Regresi Logistik pada Semua Variabel

Klasifikasi Regresi Logistik dilakukan untuk melihat ketepatan klasifikasi dari data *imbalanced* yang masih memiliki permasalahan multikolinearitas. Data akan dibagi menjadi data *training* dan data *testing* menggunakan *stratified 10-fold cross validation.* Nilai ketepatan klasifikasi Regresi Logistik status desa tertinggal di Jawa timur tahun 2014 dapat dilihat pada Tabel 4.13.

Tabel 4.13 Hasil Ketepatan Klasifikasi Regresi Logistik pada Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,736	0,737	0,073
Rata-rata G-Mean	0,735	0,728	0,085
Rata-rata Akurasi Total	0,736	0,736	0,074
Rata-rata Sensitivitas	0,691	0,692	0,136
Rata-rata Spesitifitas	0,781	0,783	0,121

Berdasarkan Tabel 4.13 didapatkan nilai *G-Mean* dan AUC secara berurutan yaitu 0,728 dan 0,737. Selanjutnya nilai akurasi total yang didapatkan sebesar 0,736 sehingga rata-rata ketepatan klasifikasi dapat dikatakan baik. Selanjutnya, nilai senstivitas sebesar 0,692 sedangkan ketepatan klasifikasi untuk data negatif atau kelas mayoritas lebih bagus, yaitu 0,783. Berdasarkan Tabel 4.13 dapat dikatakan rata-rata ketepatan klasifikasi dengan menggunakan metode klasifikasi Regresi Logistik pada data *balanced* baik karena memiliki nilai AUC dalam rentang 0,70-0,80 dan memiliki tingkat variabilitas yang rendah.

B. Regresi Logistik pada Variabel Signifikan

Setelah melakukan analisis menggunakan semua variabel, maka langkah selanjutnya adalah menguji variabel signifikan sehingga perlu melakukan deteksi multikolinearitas pada data menggunakan variabel signifikan.

i. Pengujian Parameter

Pengujian parameter secara serentak dilakukan dengan menggunakan *Likelihood Ratio Test* dengan α = 0,10. Hasil uji serentak menunjukkan bahwa H_0 ditolak karena nilai $G > \chi^2_{(8;0,10)}$, yaitu 399,997 > 13,36 sehingga minimal ada satu variabel yang berpengaruh signifikan. Berikut Tabel 4.14 hasil dari pengujian parsial :

Tabel 4.14 Nilai Koefisien, Thitung, Dan Pvalue Hasil Uji Serentak Data Balanced Semua Variabel

Variabel	Koefisien	*Standar Error*	Statistik Hitung	*P-Value*
Konstan	3,342	2,756	1,213	0,225
X_1	**0,506**	**0,195**	**2,595**	**0,009**
X_2	-0,891	4,048	-0,220	0,826
X_3	6,767	4,204	1,610	0,108
X_4	**-0,655**	**0,237**	**-2,761**	**0,006**
X_5	**-5,345**	**2,781**	**-1,922**	**0,055**
X_6	**0,232**	**0,032**	**7,338**	**0,000**
X_7	1,476	0,994	1,484	0,138
X_8	0,004	0,003	1,167	0,243

Berdasarkan Tabel 4.14 diatas, diketahui ada empat variabel signifikan yaitu variabel rasio banyaknya SD/MI terhadap total murid SD/MI, rasio banyaknya keluarga pengguna listrik terhadap total rumah tangga, rasio banyaknya toko/warung kelontong terhadap total penduduk dan jarak tempuh per kilometer ke kantor camat. Selanjutnya, menemukan variabel signifikan menggunakan metode *Backward*. dengan cara mengeluarkan variabel yang paling tidak signifikan secara bertahap. Hasil akhir yang didapat dari *backward elimination* adalah sebagai berikut :

Tabel 4.15 Nilai Koefisien, Thitung, dan Pvalue Hasil Backward Elimination

Variabel	Koefisien	*Standar Error*	Statistik Hitung	*P-Value*
Konstan	3,811	2,797	1,362	0,173
X_1	0,457	0,189	2,412	0,016
X_5	-6,067	2,816	-2,156	0,031
X_6	0,233	0,031	7,551	0,000

Berdasarkan Tabel 4.15, diketahui ada tiga variabel signifikan yaitu rasio banyaknya SD/MI terhadap total murid SD/MI, variabel rasio banyaknya keluarga pengguna listrik terhadap total penduduk dan jarak tempuh per kilometer ke kantor camat. Selanjutnya dapat dibuat model Regresi Logistik menggunakan variabel-variabel signifikan.

$$ln\left(\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right) = 3{,}811 + 0{,}457\,X_1 - 6{,}067\,X_5 + 0{,}233\,X_6$$

Model yang didapat memberikan informasi bahwa jika variabel rasio banyaknya murid SD/MI terhadap jumlah siswa SD/MI bertambah satu satuan, maka peluang status desa menjadi desa tertinggal akan naik sebesar $\exp(0{,}457) = 1{,}58$ kali atau *odds ratio* status desa menjadi desa tidak tertinggal sebesar $1{,}58^{-1} = 0{,}633$ kali. Berdasarkan data yang dianalisis, hal ini dikarenakan nilai rasio banyaknya murid SD/MI terhadap jumlah siswa SD/MI pada desa tertinggal lebih besar dari desa tidak tertinggal. Berikut perbandingan nilai rasio dari masing-masing status desa :

Tabel 4.16 Perbandingan Nilai Rasio pada Variabel X_1

Desa Tertinggal	Desa Tidak Tertinggal
11,765	5,455
4,795	5,455
3,922	3,333
3,333	2,618
3,185	2,439
⋮	⋮
0,335	0,255
0,327	0,243
0,308	0,196
0,231	0,177
0	0

Tabel 4.16 menunjukkan bahwa data nilai rasio banyaknya murid SD/MI terhadap jumlah siswa SD/MI pada desa tertinggal secara keseluruhan bernilai lebih besar sehingga berdampak pada model yang dihasilkan dari klasifikasi Regresi Logistik.

Informasi lainnya adalah jika variabel rasio banyaknya keluarga pengguna listrik terhadap total penduduk bertambah satu satuan, maka peluang status desa menjadi desa tertinggal berkurang sebesar $\exp(-6{,}067) = 0{,}002$ kali atau *odds ratio* status desa menjadi desa tidak tertinggal sebesar $0{,}002^{-1} = 500$ kali. Informasi lain yang dapat diketahui dari model adalah jika jarak tempuh per kilometer ke kantor camat bertambah 1 km, maka peluang status desa menjadi tertinggal akan naik, dimana *odds ratio* status desa tertinggal menjadi sebesar $\exp(0{,}233) = 1{,}26$ kali atau *odds ratio* status desa menjadi desa tidak tertinggal sebesar 0,794 kali.

ii. Deteksi Multikolinearitas pada Variabel Signifikan

Deteksi asumsi multikolinearitas harus dilakukan untuk melihat apakah ada hubungan korelasi yang tinggi antar variabel prediktor. Kasus multikolinearitas perlu diatasi karena variansi dan statndar error dari hasil estimasi menjadi besar. Deteksi multikolinearitas dapat dilihat dari nilai *variance inflation factor* yang dapat dilihat pada tabel berikut ini :

Tabel 4.17 Deteksi Data Mutikolinearitas (VIF) pada variabel signifikan

Variabel	VIF
X_2	1,011
X_5	1,052
X_6	1,060

Tabel 4.17 menunjukkan bahwa tidak adanya kasus multikolinearitas pada data *balanced* dari indikator-indikator status desa tertinggal Provinsi Jawa Timur tahun 2014 dengan menggunakan variabel yang signifikan.

iii. Ketepatan Klasifikasi Regresi Logistik pada Variabel Signifikan

Klasifikasi Regresi Logistik dilakukan untuk melihat ketepatan klasifikasi dari data *imbalanced* yang masih memiliki permasalahan multikolinearitas. Data akan dibagi menjadi data *training* dan data *testing* menggunakan *stratified 10-fold cross validation.* Berikut nilai ketepatan klasifikasi Regresi Logistik status desa tertinggal di Jawa Timur tahun 2014 pada Tabel 4.18 :

Tabel 4.18 Hasil Ketepatan Klasifikasi Regresi Logistik pada Data Balanced Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,732	0,730	0,070
Rata-rata G-Mean	0,729	0,720	0,080
Rata-rata Akurasi Total	0,732	0,729	0,128
Rata-rata Sensitivitas	0,666	0,672	0,114
Rata-rata Spesitifitas	0,799	0,787	0,071

Berdasarkan Tabel 4.18 didapatkan nilai *G-Mean* dan AUC secara berurutan yaitu 0,730 dan 0,720. Selain itu, rata-rata akurasi yang tidak jauh berbeda dengan AUC dan G-Mean sebesar 0,729. Selanjutnya, nilai senstivitas sebesar 0,672 menunjukkan bahwa rata-rata ketepatan klasifikasi pada data positif atau kelas minoritas cukup baik dan ketepatan klasifikasi untuk data negatif atau kelas mayoritas dikatakan baik karena memiliki nilai sebesar 0,787. Berdasarkan Tabel 4.18 dapat dikatakan rata-rata ketepatan klasifikasi dengan metode klasifikasi Regresi Logistik pada data *balanced* dengan menggunakan variabel signifikan baik. Selain itu, nilai standar deviasi menunjukkan tingkat keragaman dari rata-rata ketepatan klasifikasi secara keseluruhan bernilai kecil.

4.3.3 Ketepatan Klasifikasi Regresi Logistik Ridge pada Data *Balanced*

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan. Berikut langkah-langkah menggunakan klasifikasi Regresi Logistik Ridge:

A. Regresi Logistik Ridge pada Semua Variabel

Berdasarkan hasil pengujian asumsi multikolinearitas didapatkan lima variabel yang terdeteksi memiliki korelasi yang tinggi antar variabel prediktor. Adanya kasus multikolinearitas akan berdampak pada variansi dari hasil estimasi yang memiliki nilai nilai *error* yang besar dan interval kepercayaan menjadi lebar sehingga multikolinearitas harus diatasi. Selanjutnya, dengan menggunakan klasifikasi Regresi Logistik Ridge kasus multikolinearitas dapat teratasi. Model dari data *training*

digunakan untuk mengklasifikasikan data *testing* dan dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Provinsi Jawa timur tahun 2014 pada Tabel 4.19.

Tabel 4.19 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge pada Data *Balanced* Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,688	0,692	0,084
Rata-rata G-Mean	0,641	0,636	0,142
Rata-rata Akurasi Total	0,689	0,688	0,087
Rata-rata Sensitifitas	0,582	0,594	0,276
Rata-rata Spesitifitas	0,795	0,790	0,188

Tabel 4.19 menunjukkan bahwa nilai rata- rata *G-Mean* dan AUC secara berurutan yaitu 0,692 dan 0,636 sedangkan rata-rata akurasi pada data *balanced* sebesar 0,688 sehingga klasifikasi status desa tertinggal di Jawa Timur tahun 2014 pada data *balanced* menggunkan semua variabel menghasilkan ketepatan klasifikasi yang cukup baik dengan menggunakan metode analisis Regresi Logistik Ridge. Nilai senstivitas sebesar 0,594 menunjukkan bahwa ketepatan akurasi pada kelas minoritas kurang baik baik, namun akurasi ketepatan untuk kelas mayoritas lebih tinggi, yaitu 0,790. Selanjutnya, nilai. Berdasarkan Tabel 4.19 dapat dikatakan ketepatan klasifikasi dengan metode klasifikasi Regresi Logistik Ridge pada data *balanced* secara keseluruhan dapat dikatakan cukup baik. Dari nilai standar deviasi dapat diketahuai bahwa tingkat keragaman dari rata-rata ketepatan klasifikasi kecil.

B. Regresi Logistik Ridge pada Variabel Signifikan

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan.

Berdasarkan hasil pengujian asumsi terdapat kasus multikolinearitas pada data *imbalanced.* Selanjutnya, dengan menggunakan klasifikasi Regresi Logistik Ridge model dari data *training* digunakan untuk mengklasifikasikan data *testing* dan dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.20.

Tabel 4.20 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge pada Data *Balanced* Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,734	0,735	0,066
Rata-rata G-Mean	0,730	0,725	0,013
Rata-rata Akurasi Total	0,734	0,734	0,068
Rata-rata Sensitivitas	0,663	0,667	0,131
Rata-rata Spesitifitas	0,805	0,802	0,105

Dari Tabel 4.20 dapat diketahui rata-rata nilai AUC, G-Mean, dan akurasi total tidak jauh berbeda. Hal ini berarti klasifikasi status desa tertinggal di Jawa Timur tahun 2014 pada data *balanced* menggunakan metode analisis Regresi Logistik Ridge dapat dikatakan baik. Selanjutnya, nilai senstivitas sebesar 0,667 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas cukup baik, namun rata-rata ketepatan klasifikasi untuk data negatif atau kelas mayoritas lebih tinggi, yaitu 0,802. Keragaman dari rata-rata ketepatan klasifikasi menunjukkan nilai yang kecil sehingga nilai setiap *fold* tidak jauh berbeda atau sudah stabil.

4.3.4 Analisis Diskriminan Kernel pada Data *Balanced*

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan. Berikut tahapan langkah yang digunakan :

A. Analisis Diskriminan Kernel pada Semua Variabel

Langkah – langkah yang dilakukan dalam menentukan analisis diskriminan dapat digunakan atau tidak sama seperti pada pengujian asumsi untuk analisis diskriminan pada data *imbalanced*. Apabila data menunjukkan tidak berdistrubusi normal multivariat , maka Analisis Diskriminan Kernel dapat digunakan sebagai metode klasifikasi status ketertinggalan desa di Provinsi Jawa Timur tahun 2014 pada data yang telah *balanced*.

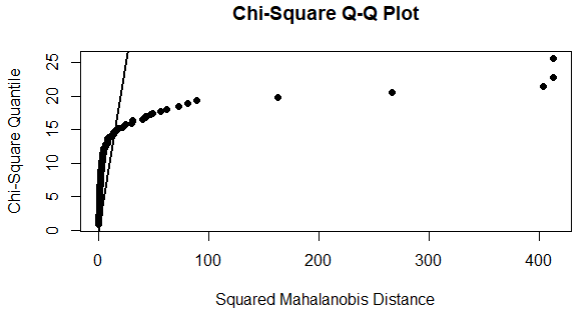
i. Hasil Uji Homogenitas pada Semua Variabel

Hasil uji homogenitas menunjukkan bahwa nilai *p-value* bernilai kecil dari $\alpha = 0,05$ yaitu 0 sehingga menyebabkan H_0

ditolak. Artinya, matriks varians kovarians pada data indikator status desa tertinggal yang telah *balanced* belum memenuhi asumsi homogenitas.

ii. Hasil Uji Distribusi Normal Multivariat pada Semua Variabel

M*ardia's test* dilakukan untuk menguji apakah data berdistribusi normal multivariat atau tidak. Hasil dari uji normalitas menghasilkan *p-value* yang bernilai 0. *P-value* yang bernilai kurang dari α = 0,05 menunjukkan bahwa data indikator status desa tertinggal tidak berdistribusi normal multivariat. Oleh karena itu, asumsi distribusi normal multivariat tidak dapat terpenuhi. Gambar 4.17 menunjukkan *qq-plot* dari data indikator status desa tertinggal Provinsi Jawa Timur tahun 2014 yang telah *balanced*.



Gambar 4.17 Chi-Squared QQ-Plot Data Balanced Semua Variabel

Titik-titik hitam pada Gambar 4.17 merepresentasikan data indikator status desa tertinggal Provinsi Jawa Timur secara multivariat. Banyaknya titik-titik yang tersebar jauh dari garis hitam mengindikasikan bahwa data tidak berdistribusi normal multivariat.

iii. Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Semua Variabel

Berdasarkan hasil pengujian asumsi diskriminan linear, terdapat dua asumsi yang belum terpenuhi, yaitu uji homogenitas

dan uji normal multivariat . Uji homogenitas menunjukkan matriks varians kovarians tidak homogen dan data *balanced* tidak berdistribusi normal multivariat. Selanjutnya dengan menggunakan model dari data *training* untuk mengklasifikasikan data *testing*, dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.21.

Tabel 4.21 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Data *Balanced* Semua Variabel

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,804	0,780	0,010
Rata-rata G-Mean	0,803	0,776	0,086
Rata-rata Akurasi Total	0,804	0,780	0,083
Rata-rata Sensitifitas	0,826	0.807	0,120
Rata-rata Spesitifitas	0,782	0,754	0,108

Tabel 4.21 menunjukkan bahwa data *balanced* memiliki nilai rata -rata *G-Mean*, AUC, akurasi total secara berurutan adalah 0,780 dan 0,776 dan akurasi total sebesar 0,780. Selanjutnya, nilai senstivitas sebesar 0,807 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas dapat dikatakan baik. Selanjutnya, ketepatan klasifikasi untuk data negatif atau kelas mayoritas, yaitu 0,754. Berdasarkan nilai AUC, *G-Mean*, dan akurasi dapat dikatakan rata-rata ketepatan klasifikasi dengan metode klasifikasi Analisis Diskriminan Kernel pada data *balanced* sudah bagus dan stabil karena memiliki nilai standar deviasi yang kecil.

B. Analisis Diskriminan Kernel pada Variabel Signifikan

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan. Berikut langkah-langkah dalam klasifikasi Analisis Diskriminan Kernel dengan variabel signifikan :

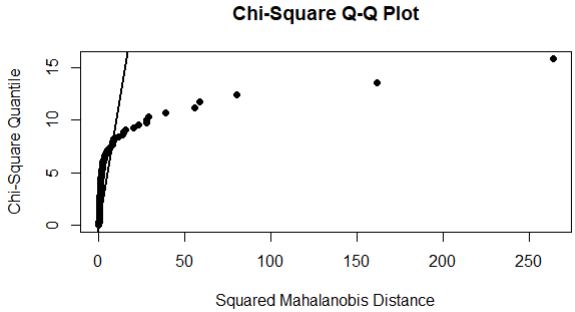
i. Hasil Uji Homogenitas pada Variabel Signifikan

Nilai *p-value* pada data *balanced* bernilai kecil dari $\alpha = 0,05$ yaitu 0 sehingga menyebabkan H_0 ditolak. Artinya, matriks varians

kovarians pada data indikator status desa tertinggal Provinsi Jawa Timur tahun 2014 tidak homogen sehingga asumsi homogenitas tidak terpenuhi.

ii. Hasil Uji Distribusi Normal Multivariat pada Variabel Signifikan

M*ardia's test* dilakukan untuk menguji apakah data berdistribusi normal multivariat atau tidak. Hasil dari uji normalitas menghasilkan *p-value* yang bernilai 0. *P-value* yang bernilai kurang dari α = 0,05 menunjukkan bahwa data indikator status desa tertinggal tidak berdistribusi normal multivariat. Oleh karena itu, asumsi distribusi normal multivariate tidak dapat terpenuhi. Gambar 4.18 menunjukkan *qq-plot* dari data indikator status desa tertinggal Provinsi Jawa Timur yang belum seimbang.



Gambar 4.18 Chi-Squared QQ-Plot Data Balanced Variabel Signifikan

Titik-titik hitam pada Gambar 4.18 merepresentasikan data indikator status desa tertinggal Provinsi Jawa Timur secara multivariat. Banyaknya titik-titik yang tersebar jauh dari garis hitam mengindikasikan bahwa data tidak berdistribusi normal multivariat.

iii. Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Variabel Signifikan

Berdasarkan hasil pengujian asumsi diskriminan linear, terdapat dua asumsi yang belum terpenuhi, yaitu uji homogenitas dan uji normal multivariat. Uji homogenitas menunjukkan matriks varians kovarians tidak homogen dan data *balanced* tidak

berdistribusi normal multivariat. Hal ini menunjukkan metode klasifikasi analisis diskriminan linear tidak tepat untuk mengklasifikasikan data indikator status desa tertinggal di Provinsi Jawa Timur tahun 2014 sehingga metode klasifikasi yang digunakan adalah klasifikasi Analisis Diskriminan Kernel. Selanjutnya dengan menggunakan model dari data *training* untuk mengklasifikasikan data *testing*, dapat dilihat nilai ketepatan klasifikasi status desa tertinggal di Jawa timur tahun 2014 pada Tabel 4.22.

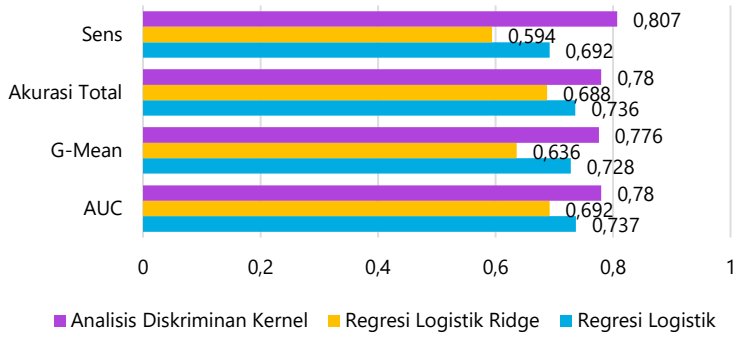
Tabel 4.22 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel pada Data Balanced Variabel Signifikan

	Training	***Testing***	***Statdev (Testing)***
Rata-rata AUC	0,712	0,699	0,077
Rata-rata G-Mean	0,712	0,690	0,087
Rata-rata Akurasi Total	0,712	0,698	0,078
Rata-rata Sensitivitas	0,691	0,682	0,143
Rata-rata Spesitifitas	0,734	0,715	0,121

Nilai rata-rata AUC, G-Mean, dan akurasi total tidak jauh berbeda, yaitu, 0,699, 0,690, dan 0,698. Hal ini berarti bahwa klasifikasi status desa tertinggal di Jawa Timur dengan metode Analisis Diskriminan Kernel menghasilkan klasifikasi cukup baik. Selain itu, rata-rata nilai senstivitas sebesar 0,682 menunjukkan bahwa ketepatan klasifikasi pada data positif atau kelas minoritas cukup baik sedangkan nilai spesifisitas menunjukkan ketepatan klasifikasi untuk data negatif atau kelas mayoritas lebih tinggi, yaitu 0,715. Berdasarkan nilai rata-rata ketepatan klasifikasi pada Tabel 4.21 dapat disimpulkan bahwa metode klasifikasi Analisis Diskriminan Kernel cukup baik untuk digunakan sebagai *classifier* dan menunjukkan variabiltas yang kecil sehingga hasil dari *10 - fold CV* sudah stabil. Akan tetapi, jika dibandingkan dengan hasil dari ketepatan klasifikasi menggunakan Analisis Diskriminan Kernel pada semua variabel data status desa tertinggal di Provinsi Jawa Timur tahun 2014 terjadinya penurunan performansi yang cukup jauh.

4.3.5 Analisis Gabungan Data *Balanced* Semua Variabel dan Variabel Signifikan

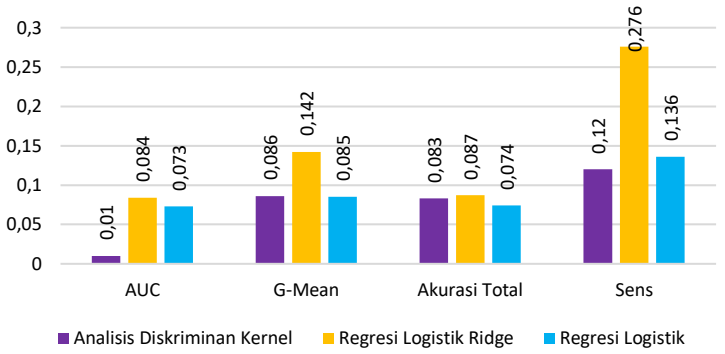
Pada data yang telah *balanced* dengan metode *combine undersampling*, sebelumnya telah dianalisis menggunakan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel baik pada semua variabel dan variabel signifikan. Berikut adalah perbandingan ketiga metode *classifier* dengan semua variabel :



Gambar 4.19 Perbandingan Ketepatan Klasifikasi Data Balanced Semua Variabel

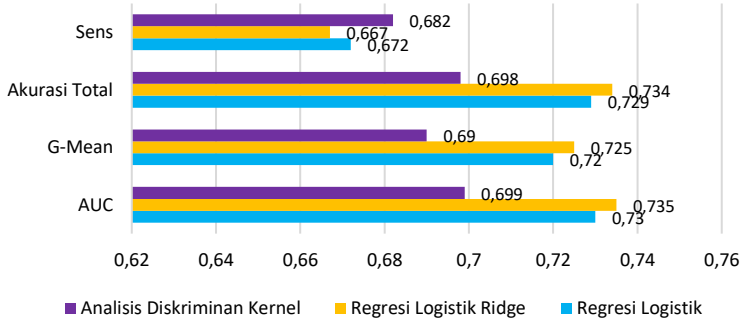
Gambar 4.19 menunjukkan bahwa pada data *balanced* efektivitas performansi tertinggi dihasilkan dari metode Analisis Diskriminan Kernel. Hal ini sama seperti hasil ketepatan klasifikasi pada data *imbalanced* yang menunjukkan Analisis Diskriminan Kernel lebih baik untuk penelitian status desa tertinggal di Provinsi Jawa Timur tahun 2014 dengan menggunakan semua variabel prediktor. Perubahan yang signifikan dapat dilihat dari nilai rata-rata sensitivitas, yaitu dengan adanya metode *balancing* mampu meningkatkan ketepatan klasifikasi pada kelas data minoritas. Selain itu, nilai ketepatan klasifikasi yang dihasilkan dari metode Regresi Logistik Ridge cenderung paling rendah, namun metode tersebut mampu mengakomodasi masalah multikolinearitas

sehingga hal ini dapat dijadikan metode yang dipertimbangkan. Tingkat keragaman ketepatan klasifikasi pada data *balanced* dapat dilihat dari nilai standar deviasi seperti pada Gambar 4.20.



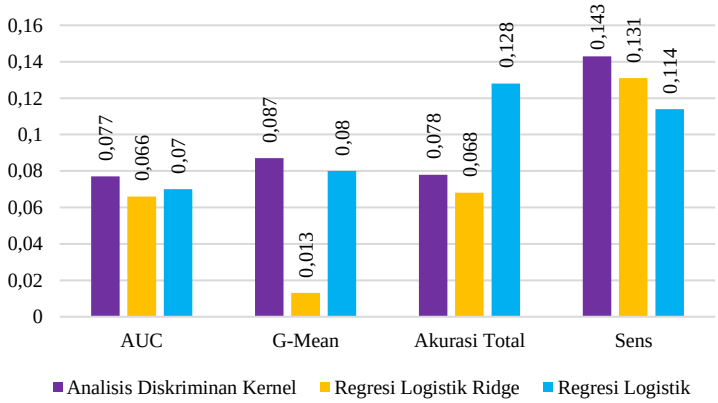
Gambar 4.20 Perbandingan Standar Deviasi Data Balanced Semua Variabel

Dari Gambar 4.20 dapat diketahui bahwa nilai standar deviasi dari masing-masing ketepatan klasifikasi menunjukkan Analisis Diskriminan Kernel cenderung memiliki keragaman yang rendah sehingga hal ini berdampak pada nilai rata-rata AUC yang lebih tinggi dari metode lainnya.. Sementara itu, perbandingan ketepatan klasifikasi dengan variabel signifikan dapat dilihat pada Gambar 4.21.



Gambar 4.21 Perbandingan Ketepatan Klasifikasi Data Balanced Variabel Signifikan

Gambar 4.21 menunjukkan bahwa pada data *balanced* dengan menggunakan variabel signifikan memiliki efektivitas performansi tertinggi dihasilkan dari metode Regresi Logistik Ridge sedangkan nilai ketepatan klasifikasi yang dihasilkan dari metode Analisis Diskriminan Kernel cenderung rendah . Tingkat keragaman ketepatan klasifikasi pada data *balanced* dapat dilihat dari nilai standar deviasi seperti pada Gambar 4.22.



Gambar 4.22 Perbandingan Standar Deviasi Data Balanced Variabel Signifikan

Gambar 4.22 menunjukkan bahwa nilai standar deviasi dari masing-masing ketepatan klasifikasi pada Regresi Logistik Ridge cenderung memiliki keragaman yang rendah sehingga hal ini berdampak pada nilai rata-rata AUC, G-Mean, dan akurasi yang lebih tinggi dari metode lainnya.

4.4 Efektivitas Metode *Combine Undersampling*

Dari evaluasi ketepatan model klasifikasi dapat diketahui performansi dari metode *classifier* yang digunakan. Bukan hanya itu, perbandingan data *balanced* dan *imbalanced* serta data yang menggunakan semua variabel atau variabel signifikan akan berdampak pada hasil dari rata-rata ketepatan klasifikasi dari metode *classifier*.

4.4.1 Efektivitas Metode *Combine Undersampling* pada Semua Variabel

Pada penelitian ini menggunakan data semua variabel agar nantinya dapat dibandingkan efektifitas performansinya dengan data yang hanya menggunakan variabel signifikan.

Dalam menentukan metode terbaik untuk memprediksi status desa tertinggal di Jawa Timur dapat dilihat dari perbandingan dari nilai akurasi, G-Mean, dan AUC pada data asli maupun data *balanced*. Berikut perbandingan dari ketiga metode *classifier* :

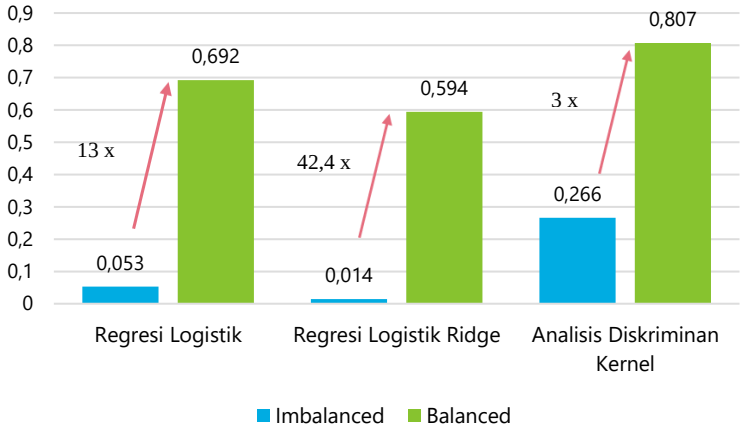
Tabel 4.23 Perbandingan Metode Klasifikasi Terbaik pada Semua Variabel

		AUC	G-Mean	Sens.	Stdv. AUC	Stdv. G - Mean	Stdv Sens.
Regresi Logistik	1	0,515	0,113	0,034	0,023	0,152	0,051
	2	0,737	0,728	0,692	0,073	0,141	0,136
Regresi Logistik Ridge	1	0,507	0,053	0,019	0,016	0,113	0,032
	2	0,692	0,636	0,594	0,084	0,155	0,276
Analisis Diskriminan Kernel	1	0,622	0,504	0,266	0,036	0,079	0,072
	2	0,780	0,776	0,807	0.010	0,104	0,120

Keterangan : 1. *Imbalanced*
2. *Balanced*

Tabel 4.23 menunjukkan adanya perbedaan yang signifikan terhadap nilai ketepatan klasifikasi antara data *imbalanced* dengan data *balanced*. Pada data *imbalanced*, ketepatan klasifikasi pada nilai akurasi tidak dapat dijadikan tolak ukur yang tepat untuk melihat efektifitas dari hasil klasifikasi karena nilai spesitifitas dan sensitivitas yang berbeda jauh sehingga berdampak pada nilai akurasinya. Metode *classifier* terbaik pada data *imbalanced* dilihat dari nilai rata-rata AUC, G-Mean, dan sensitivitas sebesar 62,2%, 50,4%, dan 26,6% dengan menggunakan metode klasifikasi Analisis Diskriminan Kernel sedangkan metode Regresi Logistik Ridge menunjukkan nilai ketepatan klasifikasi yang sangat rendah, yaitu 50,7%, 5,3%, dan 1,9%.

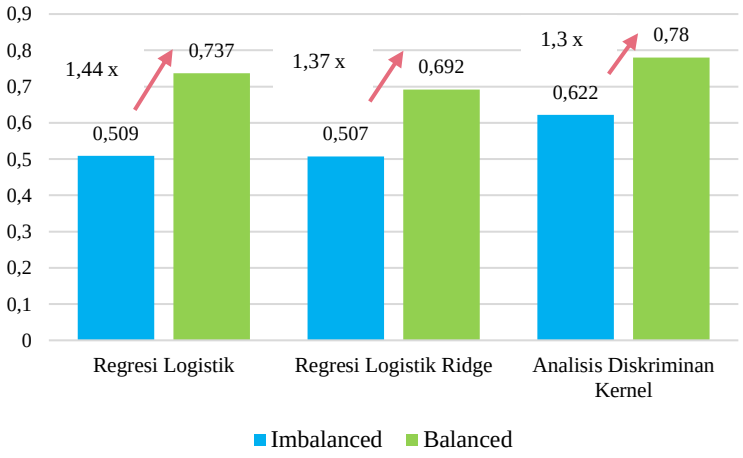
Selanjutnya, metode Combine Undersampling digunakan untuk mengatasi data tidak seimbang pada data status desa tertinggal di Provinsi Jawa Timur tahun 2014 mampu meningkatkan ketepatan klasifikasi dari ketiga metode *classifier* yang digunakan. *Classifier* terbaik pada data *balanced* adalah Analisis Diskriminan Kernel dengan nilai akurasi sebesar 78% sedangkan ketepatan klasifikasi pada kelas data mayoritas dan minoritas tidak memiliki selisih yang terlalu jauh. Selain itu, nilai rata-rata AUC,G-Mean, akurasi, dan sensitivitas secara berurutan adalah 78%, 77,6%, 78% dan 80,7%. Jadi, dapat disimpulkan bahwa analisis kasus data status desa tertinggal Provinsi Jawa Timur tahun 2014 menunjukkan penerapan *Combine Undersampling* pada klasifikasi Analisis Diskriminan Kernel dengan menggunakan semua variabel merupakan metode yang efektif untuk penyelesaian permasalahan ini. Visualisasi dari perbandingan nilai rata-rata sensitivitas antara data *imbalanced* dengan data *balanced* dapat dilihat pada Gambar 4.23.



Gambar 4.23 Perbandingan Nilai Rata-Rata Sensitivitas Ketiga Classifier pada Semua Variabel

Dari Gambar 4.23 diketahui bahwa peningkatan performansi klasifikasi tertinggi pada rata-rata sensitivitas ada pada klasifikasi

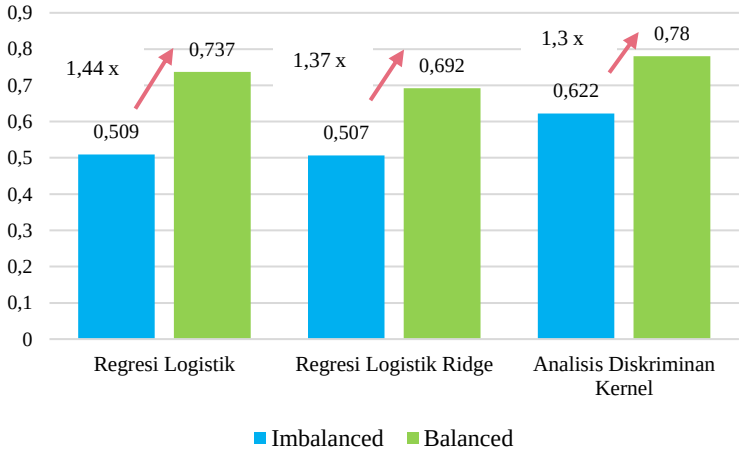
Regresi Logistik Ridge, dimana penerapan metode *Combine Undersampling* mampu meningkatkan ketepatan klasifikasi sebesar 42,4 kali dari data *imbalanced*. Uriuan kedua adalah metode klasifikasi Regresi Logistik dengan meningkatkan performansi hingga 13 kali dan Analisis Diskriminan Kernel dengan peningkatan paling rendah, yaitu 3 kali. Selanjutnya, dapat ditunjukkan perbandingan nilai rata-rata AUC pada data *imbalanced* dengan tiga metode *classifier*.



Gambar 4.24 Perbandingan Nilai Rata-Rata AUC Ketiga Classifier pada Semua Variabel

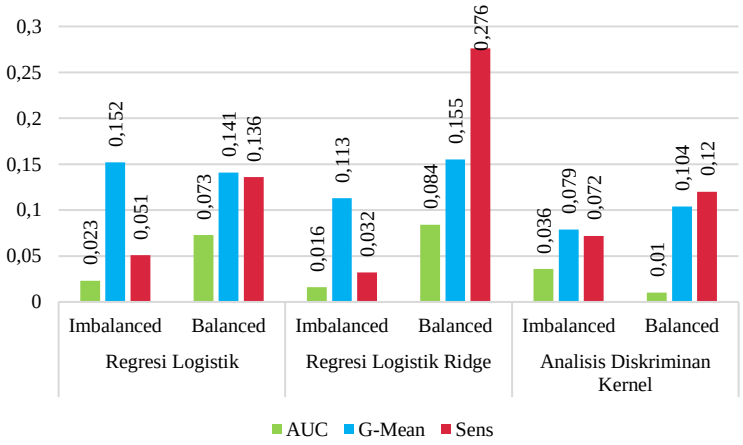
Gambar 4.24 menunjukkan bahwa kenaikan dari nilai rata-rata AUC terbesar yaitu 1,44 kali dengan menggunakan metode Regresi Logistik. Kemudian disusul oleh klasifikasi Regresi Logistik Ridge sebesar 1,37 kali dan terendah dengan menggunakan Analisis Diskriminan Kernel sebesar 1,3 kali. Hal ini berarti penanganan data *imbalanced* dengan *Combine Undersampling* mampu meningkatkan nilai AUC secara signifikan pada Regresi Logistik, walaupun metode ini tidak menghasilkan nilai rata-rata AUC tertinggi seperti Analisis Diskriminan Linier.

Kemudian visualisasi dari perbandingan nilai rata-rata G-Mean dilihat pada Gambar 4.25.



Gambar 4.25 Perbandingan Nilai Rata-Rata G-Mean Ketiga Classifier pada Semua Variabel

Dari perbandingan nilai rata-rata G-Mean dapat diketahui bahwa penerapan *Combine Undersampling* pada klasifikasi Regresi Logistik Ridge mampu meningkatkan ketepatan klasifikasi sebesar 12 kali dari data *imbalanced* dan ini merupakan peningkatan yang tertinggi dari metode lainnya sedangkan Regresi Logistik dapat meningkatkan sebesar 6,4 kali. Selanjutnya, peningkatan performansi terendah tetap dari hasil dari klasifikasi Analisis Diskriminan Kernel, yaitu 1,5 kali. Hal ini berarti bahwa peningkatan ketepatan klasifikasi secara signifikan pada data imbalanced akan sangat berdampak ketika menggunakan metode Regresi Logistik dan Regresi Logistik Ridge sehingga adanya metode resampling sangat baik untuk diterapkan, khususnya pada data status desa tertinggal di Provinsi Jawa Timur tahun 2014. Perbandingan dari nilai rata-rata AUC dan sensitivitas dapat divisualisasikan pada Gambar 4.26.



Gambar 4.26 Perbandingan Standar Deviasi dengan Semua Variabel

Dari Gambar 4.26 diketahui bahwa standar deviasi dari rata-rata AUC, G-Mean, dan sensitivitas pada data *balanced* untuk setiap metode *classifier* lebih tinggi dibandingkan data *imbalanced* sehingga dapat dikatakan hasil ketepatan klasifikasi pada data imbalanced lebih stabil dibandingkan data *balanced*. Namun, secara keseluruhan standar deviasi dari rata-rata AUC, G-Mean, dan sensitivitas bernilai kecil, maka variabilitas dari hasil *10 fold cross validation* relatif stabil.

4.4.2 Efektivitas Metode *Combine Undersampling* pada Variabel Signifikan

Setelah melakukan analisis menggunakan semua variabel, maka langkah selanjutnya adalah menguji variabel signifikan sehingga dapat dilakukan pengujian parameter terlebih dahulu.

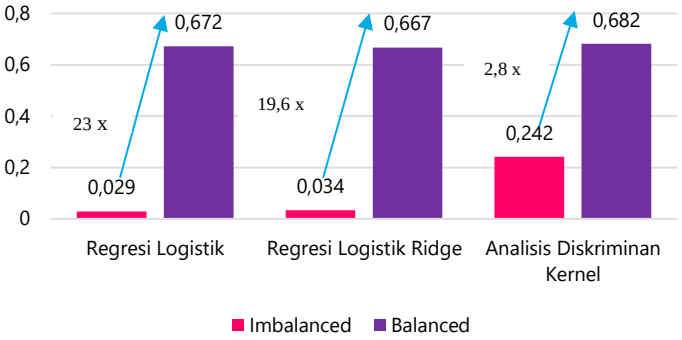
Dalam menentukan metode terbaik untuk memprediksi status desa tertinggal di Jawa Timur dapat dilihat dari perbandingan dari nilai akurasi, G-Mean, dan AUC pada data asli maupun data *balanced*. Berikut perbandingan dari ketiga metode *classifier* :

Tabel 4.24 Perbandingan Metode Klasifikasi Terbaik pada Data Balanced Variabel Signifikan

		AUC	G-Mean	Sens	Stdv. AUC	Stdv. G-Mean	Stdv. Sens
Regresi Logistik	1	0,514	0,104	0,029	0,023	0,085	0,046
	2	0,730	0,720	0,672	0,070	0,080	0,114
Regresi Logistik Ridge	1	0,516	0,11	0,034	0,03	0,142	0,06
	2	0,735	0,725	0,667	0,066	0,013	0,131
Analisis Diskriminan Kernel	1	0,608	0,475	0,242	0,043	0,086	0,086
	2	0,699	0,690	0,682	0.077	0,087	0,143

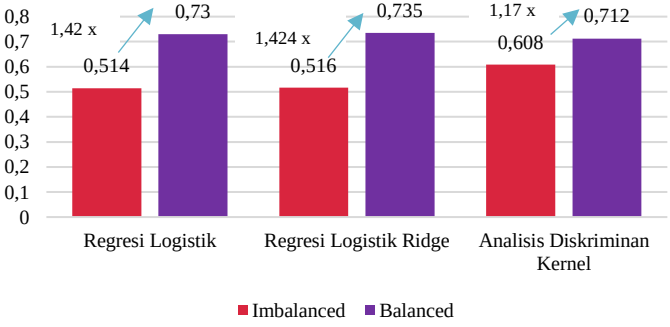
Keterangan : 1. *Imbalanced*
2. *Balanced*

Tabel 4.24 menunjukkan *classifier* terbaik pada data *balanced* yang menggunakan adalah Regresi Logistik Ridge, dimana metode ini juga mampu mengatasi kasus multikolinearitas dan memiliki nilai akurasi sebesar 73,4% sedangkan ketepatan klasifikasi pada kelas data mayoritas dan minoritas tidak memiliki selisih yang terlalu jauh. Selain itu, nilai G-Mean dan AUC secara berurutan adalah 73,5% dan 72,5%. Jika dibandingkan ketika menggunakan data dengan semua variabel performansi dari Analisis Diskriminan Kernel lebih tinggi dari metode lainnya. Tetapi, ketika menggunakan variabel signifikan, ketepatan klasifikasi yang dihasilkan paling rendah dibandingkan klasifikasi Regresi Logistik Ridge dan Regresi Logistk sehingga permasalahan ini merupakan salah satu kelemahan dari metode Analisis Diskriminan Kernel dalam mengklasifikasikan data status desa tertinggal di Jawa Timur tahun 2014. Jadi, dapat disimpulkan bahwa analisis status desa tertinggal Provinsi Jawa Timur tahun 2014 menunjukkan penerapan *Combine Undersampling* pada klasifikasi Regresi Logistik Ridge dengan menggunakan variabel signifikan merupakan metode yang efektif untuk penyelesaian permasalahan ini. Visualisasi dari perbandingan nilai rata-rata sensitivitas menggunakan variabel signifikan dari ketiga *clasissifier* dapat dilihat pada Gambar 4.27.



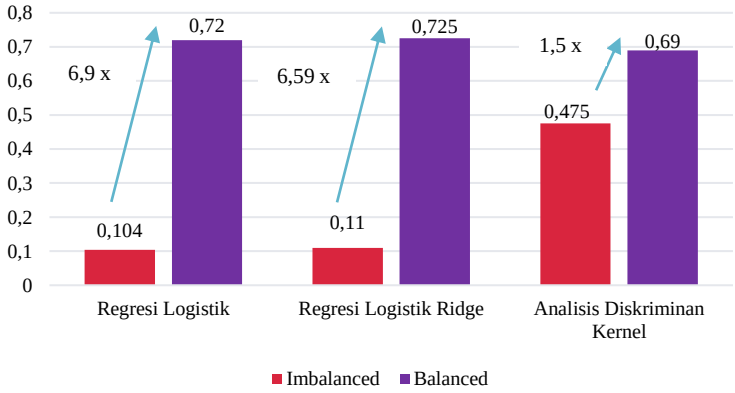
Gambar 4.27 Perbandingan Nilai Sensitivitas Ketiga Classifier pada Variabel Signifikan

Gambar 4.27 menunjukkan bahwa penerapan *Combine Undersampling* pada data *imbalanced* mampu meningkatkan ketepatan klasifikasi secara signifikan dan tertinggi pada Regresi Logistik sebesar 23 kali. Urutan kedua dengan peningkatan sebesar 19,6 kali pada klasifikasi Regresi Logistik Ridge dan terakhir klasifikasi Analisis Diskriminan Kernel sebesar 2,8 kali. Visualisasi dari perbandingan rata-rata nilai AUC pada Gambar 4.28.



Gambar 4.28 Perbandingan Nilai AUC Ketiga *Classifier* pada Variabel Signifikan

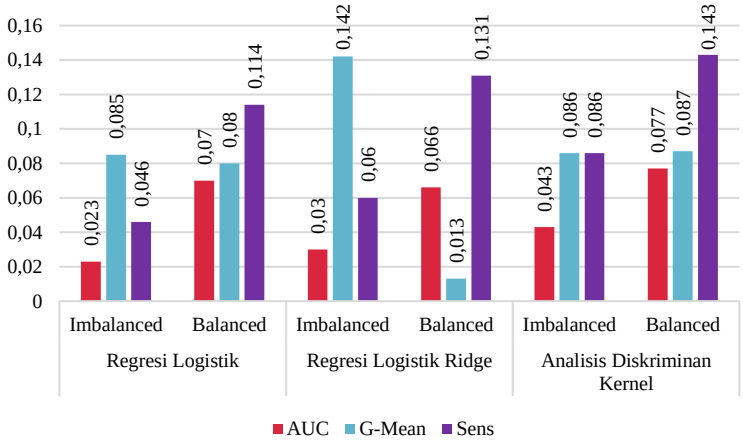
Berdasarkan Gambar 4.28 dapat diketahui bahwa peningkatan nilai AUC terbesar adalah 1,424 kali dengan menggunakan klasifikasi Regresi Logistik Ridge sedangkan Regresi Logistik sebesar 1,42 kali dan Analisis Diskriminan Kernel sebesar 1,17 kali. Selain itu, dapat diperhatikan bahwa jika dibandingkan dengan nilai rata-rata dari G-Mean dan sensitivitas, rata-rata AUC tidak mengalami peningkatan yang teralu besar. Selanjutnya, dapat dilihat perbandingan nilai rata-rata G-Mean pada Gambar 4.29.



Gambar 4.29 Perbandingan Nilai Rata-Rata G-Mean Ketiga Classifier pada Variabel Signifikan

Gambar 4.29 menunjukkan bahwa klasifikasi Regresi Logistik mengalami peningkatan terbesar yaitu 6,9 kali dari data yang belum di*resampling*. Disusul klasifikasi Regresi Logistik yang mampu meningkatkan nilai rata-rata G-Mean, yaitu 6,59 kali dan terakhir klasifikasi Analisis Diskriminan Kernel sebesar 1,5 kali. Hal ini menunjukkan bahwa penerapan *Combine Under sampling* (*Tomek Links + Random Undersampling*) pada data yang menggunakan variabel signifikan memiliki dampak yang besar untuk metode klasifikasi Regresi Logistik dan Regresi Logistik Ridge. Kemudian, dapat dilihat perbandingan dari nilai rata-rata

AUC, G-Mean, dan sensitivitas dapat divisualisasikan pada Gambar 4.30.



Gambar 4.30 Perbandingan Standar Deviasi pada Variabel Signifikan

Gambar 4.30 menunjukkan bahwa standar deviasi dari AUC dan sensitivitas mengalami kenaikan pada ketiga metode *classifier* kecuali G-Mean yang mengalami penurunan . Jika dibandingkan, nilai standar deviasi dari data *imbalanced* lebih stabil dibandingkan data *balanced*. Akan tetapi, secara keseluruhan hasil ketepatan klasifikasi sudah dapat dikatakan stabil karena nilai standar deviasi yang bernilai rendah.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan, diperoleh kesimpulan sebagai berikut :

1. Analisis statistika deskriptif dari seluruh desa di Provinsi Jawa Timur menunjukkan bahwa sebagian besar nilai tertinggi pada masing-masing variabel di desa tidak tertinggal didominasi oleh desa di Kabupaten Sumenep sedangkan di desa tertinggal variabel tertinggi terdapat di desa dari beberapa kabupaten yang berbeda. Terdapat 6 variabel yang memiliki nilai minimum 0, hal ini menunjukkan masih ada beberapa desa dari status desa tidak tertinggal ataupun desa tertinggal tidak memiliki jumlah SD/MI, tempat praktik bidan, poskesdes, toko kelontong, penderita gizi buruk, dan pendapatan asli desa.
2. Pada data *imbalanced*, hasil rata-rata ketepatan klasifikasi tertinggi menggunakan semua variabel memiliki nilai AUC, G-Mean, dan sensitivitas secara berurutan adalah 62,2% ,50,4%, dan 26,6% dengan menggunakan metode Analisis Diskriminan Kernel. Sama halnya dengan menggunakan variabel signifikan hasil ketepatan klasifikasi menunjukkan rata-rata AUC, G-Mean, sensitivitas adalah 60,8%, 47,5%, 24,2%. Rendahnya ketepatan klasifikasi pada data *imbalanced* disebabkan kesalahan pengklasifikasian pada data kelas minor.
3. Hasil ketepatan klasifikasi 3 metode berdasarkan nilai rata-rata AUC, G-Mean, akurasi, dan sensitivitas menunjukkan metode klasifikasi Analisis Diskriminan Kernel dengan semua variabel, menghasilkan ketepatan klasifikasi tertinggi dengan nilai rata-rata AUC, G-Mean, akurasi total, dan sensitivitas secara berurutan sebesar 78,0%, 77,6%, 78,0%, dan 80,7%. Selain itu, klasifikasi Regresi Logistik Ridge memiliki rata-rata ketepatan klasifikasi tertinggi saat digunakan pada variabel signifikan dengan nilai rata-rata AUC, G-Mean, akurasi total, dan sensitivitas sebesar 73,5%, 72,5%, 73,4%, dan 66,7% . Metode

classifier terbaik untuk penelitian klasifikasi data status desa tertinggal di Provinsi Jawa Timur tahun 2014 adalah *Combine Undersampling* pada Analisis Diskriminan Kernel.

4. Penerapan *Combine Undersampling* pada data *imbalanced* mampu menaikkan ketepatan klasifikasi pada ketiga *classifier* yang digunakan. Peningkatan ketepatan klasifikasi tertinggi diperoleh dari nilai rata-rata sensitivitas hasil klasifikasi Regresi Logistik Ridge dengan data yang menggunakan semua variabel sebesar 42,4 kali.

5.2 Saran

Berdasarkan kesimpulan yang diperoleh, dapat dirumuskan saran sebagai pertimbangan penelitian selanjutnya adalah sebagai berikut :

1. Menerapkan jenis-jenis kernel lainnya sebagai pembanding dari hasil kernel *Gaussian RBF*.
2. Menerapkan metode *Combine Undersampling* pada data campuran dan *multiclass*.
3. Mengaplikasikan metode *resampling* dan klasifikasi lainnya pada data *imbalanced* sebagai metode pembanding sehingga didapatkan metode yang memiliki ketepatan klasifikasi yang lebih baik dari metode yang digunakan pada penelitian ini.

DAFTAR PUSTAKA

Adisasmita, R. (2005). *Dasar-dasar Ekonomi Wilayah*. Yogyakarta: Graha Ilmu.

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley & Sons.

Agresti, A. (2002). *Categorical Data Analysis* (3rd ed.). New York: John Wiley & Sons, Inc.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Wiley-Interscience.

Aitchison, J., & Aiken, C. G. (1976). Multivariate Binary Discrimination by Kernel Method. *Biometrika*, 413-420.

Arar, O. F., & Ayan, K. (2015). Software defect prediction using cost-sensitive neural network. *Applied Soft Computing*, 1-15. doi:http://doi.org/10.1016/j.asoc.2015

Azminuddin, A. I., Suhartono, V., & Himawan, H. (2017). Model Multi-Class SVM Menggunakan Strategi 1V1 Untuk Klsifikasi Wall-Following Robot Navigation Data. *Teknologi Informasi*.

Bappenas. (2015). *Indeks Pembangunan Desa 2014*. Diambil kembali dari Tantangan Pemenuhan Standar Pelayanan Minimum Desa: https://www.bappenas.go.id/files/5514/4704/6044/Buku_Indeks_Pembangunan_Desa_2014.pdf

Batista, G. E., Bazzan , A. L., & Monard, M. C. (2003). Balancing Training Data for Automated Annotation of Keywords : a Case Study.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Asessment Over Imbalanced Data Sets. *Journal of Information Engineering and Applications, 3*(10).

BPS. (2010). *Badan Pusat Statistika*. Diambil kembali dari Hasil Sensus Penduduk 2010 Data Agregat per Provinsi: http://kesga.kemkes.go.id/images/pedoman/SP2010_Jumlah%20Penduduk%20Per%20Provinsi.pdf

BPS. (2015). *Indeks Pembangunan Desa 2014 "Tantangan Pemenuhan Standar Pelayanan Minimum Desa"*. Jakarta: Bappenas.

Curt, H. (1995). "The devil's in the detail : techniques : Tools, and applications for database mining and knowledge discovery-Part". *Intelligent Software Strategies* , pp. 1-15.

Devi, D., Biswas, S. K., & Purkayastha, B. (2017). Redundancy-driven modified Tomek-link based undersampling. *Pattern Recognition Letters*, 3-12.

Dimulyo, S. (2009). Penggunaan Geographically Weighted Regression-Kriging Untuk Klasifikasi Desa Teringgal. *Seminar Nasional Aplikasi Teknologi Informasi 2009 (SNATI 2009)*.

Djuraidah, A., & Aunuddin. (2004). Analisis Diskriminan Kernel Untuk Pengelompokkan Warna. *Forum Statistika dan Komputasi*, 101-106.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (Third ed.). Canada: John Wiley & Sons.

Elhassan, T., Aljurf, M., Al-Mohanna, F., & Shoukri, M. (2016). Classification of Imbalanced Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Journal of Informatics and Data Mining, 1*(2).

Haibo, H., Member, & Edwardo, G. A. (2009). Learning from Imbalance Data. *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 21*, pp. 1041-4347.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2006). *Multivariate Data Analysis.* Pearson Education Pretince Hall. Inc.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Technique Third Edition.* United States of America: Elsevier Inc.

Hardle, W. (1990). Smoothing Techniques with Implementation in Statistics. *Spinger-Verlag*.

Haerdle, W. K., Prastyo, D. D., & Hafner, M., (2014), *Support Vector Machines with Evolutionary Model Selection for Default Prediction*, in J. Racine, L. Su, and A. Ullah (Eds.), The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Oxford University Press, New York, pp. 346-373.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning . *Springer Verlag*.

Hocking, R. R. (2003). *Methods and Applications of Linear Models 2nd Edition.* New Jersey: John Wiley & Sons.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression Second Edition.* New York, United State of America : John Wiley & Sons,Inc.

Hsu, C., Chang, C., & Lin, C. (2004). A Practical Guide to Support Vector Classification. *Information Engineering Taiwan University*.

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Data Cleaning for Classification Using Misclassfication Analysis.

Johnson, R. A., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Pearson Education, Inc.

Khattree, R., & Naik, D. N. (2000). *Multivariate Data Reduction and Discrimination with SAS Software.* Cary: NC: SAS Intitute. Inc.

Kubat, M., & Matwin, S. (1997). Addresing the Curse of Imbalanced Training Set : One Sided Selection. *14th International Conference on Machine Learning*, 179-186.

Li, Y., Gong, S., & Liddell, H. (2003). Recognising Trajectories of Facial Identities Using Kernel Duscriminant Analysis. *Image and Vision Computing*, 1077-1086.

Liu, S., Kawamoto, T., Morita, O., Yoshinari, K., & Honda, H. (2017). Discriminating Between Adaptive And Carcigonic liver Hyperthropy In Rat Studies Using Logistic Ridge

Regression Analysis of Toxicogenomic Data. The Mode of Action And Predictive Models.

Lorena, A. C., Batista, G. E., Carvalho, A. D., & Monard, M. C. (2002). Splice Junction Recognition Using Machine Learning Techniques.

Maalouf , & Siddiqi. (2014). "Weighted Logistic Regression for Large-Scale Imbalanced And Rare Events Data". *Journal of Knowledge -Based Systems, 59*, pp.141-148.

Maalouf. (2011). Logistic Regression In Data Analysis : An Overview. *Data Analysis Techniques and Strategies, 3*(3), 281-299.

Maalouf, M., & Trafalis, T. B. (2010). Robust weighted kernel logistic regression in imbalanced and rare. *Computational Statistics and Data Analysis 2011, 55*, 168-183.

Meilianawati, P., Sumarminingsih, E., & Wardhani, N. W. (2013). Pendekatan Model Proportional Odds Dan Analisis Diskriminan Kernel Pada Regresi Respon Ordinal.

Mika, S., Ratsch, G., Jason, W., Scholkopf, B., & Muller, K. R. (1999). *Fisher Discriminant Analysis with Kernel.* University of London :Egham.

Morton, R., Hebel, J., & McCarter, R. (2008). *A Study Guide to Epidemiology and Biostatistics, 5th Edition.* Sudbury: Jones and Bartlett Publishers, Inc.

Nugroho, A., Witarto, A., & Handoko, D. (2003). Support Vector Machine, Teori dan Aplikasinya dalam Bioinformatika. *Proceeding of Indonesian Scientific.* Japan: Ilmu Komputer.com.

Park, B., Oh, S., & Pedrycz, W. (2013). The design of polynomial function-based Neural Network predictors for detection of software defects. *Information Sciences, 229*, 40-57. doi:http://doi.org/10.1016/j.ins.2011.01.026

Park, Y., & Sklansky, J. (1990). Fast Tree Classifiers. *Pattern Recognition, 1990. Proceedings., 10th International Conference on*. doi:10.1109/ICPR.1990.118192

Putra, D. M., & Ratnasari, V. (2015). Pemodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur Dengan Menggunakan Metode Regresi Ridge. *Jurnal Sains dan Seni ITS, IV*.

Rahmawati, R., Djuraidah, A., & Aidi, M. N. (2010). Penggunaan Geographically Weighted Regression (GWR) Dengan Pembobot Gauss Kernel Untuk Klasifikasi Desa Miskin. *Prosiding Seminar Nasional Matematika, 5*, pp. 43-48.

Rancher, A. (2002). *Methods of Multivariate Analysis* (2nd ed.). Canada: John Wiley & Sons, Inc.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of A Density Function. 832-837.

Ryan, T. P. (1997). *Modern Regression Methods.* New York: John Wiley & Sons.

Sain, H., & Purnami, S. W. (2015). Combine Sampling Support Vector Machine for Imbalanced Data Classification. *72*, 59-66.

Sambodo, H. P., Purnami, S. W., & Rahayu, S. P. (2014). *Ketepatan Klasifikasi Status Ketertinggalan Desa Dengan Pendekatan Reduce Support Vector Machine (RSVM) Di Provinsi Jawa Timur.* Surabaya: Tesis, Statistika FMIPA ITS.

Seber, G. A. (1984). *Smoothing Observations.* New York: John Willey & Sons.

Sharma, S. (1996). *Applied Multivariates Techniques.* Canada: John Wiley & Sons, Inc.

Solberg, A., & Solberg , R. (1996). A Large-Scale Evaluation of Features for Automatic Detection of Oil Spills in ERS SAR Images. *In International Geoscience and Remote Sensing Sypmosium*, pp. 1484-1486.

Sukrajap, M. A., & Harahap, D. H. (2017). Analisis Pengaruh Pelatihan Self-Leadership Dan Motivasi Berprestasi Terhadap Kinerja Kepala Desa Di Kabupaten Gunung Kidul Yogyakarta. *Jurnal Maksipreneur, VII*(1), pp. 94-106.

Sulasih, A., Purnami, S. W., & Rahayu, S. P. (2016). *Rare Event Weihted Logistic Regression Untuk Klasifikasi Imbalanced Data (Studi Kasus: Klasifikasi Desa Tertinggal Di Provinsi Jawa Timur).* Surabaya: Thesis, Statistika FMIPA ITS.

Sunyoto, Setiawan, & Zain, I. (2009). *Regresi Logistik Ridge Pada Keberhasilan Siswa SMA Negeri 1 Kediri Diterima di Perguruan Tinggi.* Surabaya: Tesis, Statistika FMIPA-ITS.

Tomek, I. (1997). Two Modifications of CNN . *IEEE Transactions of Systems Man and Communications* , 769-772.

Vago, H., & Kemeny, S. (2006). Logistic Ridge For Clinical Data Analysis (A Case Study). *Applied Ecology And Environmental Research*, 171-179.

Yan, X., & Su, X. G. (2009). *Linear Regression Analysis : Theory and Computing* . Singapore: World Scientific.

Yang, J., Jin, Z., Yang, J., Zhang, D., & Frangi, A. F. (2004). Essence of kernel Fisher discriminant : KPCA plus LDA. *Pattern Recognition, 37*(10), 2097-2100.

Yu, D., Hu, J., Tang, Z., Shen, H., Yang, J., & Yang, J. (2013). Neurocomputing Improving protein-ATP Binding Residues Prediction by Boosting SVMs with Random Under-sampling. *Neurocomputing, 104*, 180-190.

Yu, X., Zhou, M., Chen, X., Deng, L., & Wang, L. (2017). Using Class Imbalance Learning for Cross-Company Defect Prediction.

Yuchun, T., Ya-Qing, Z., Chawla, N. V., & Sven, K. (2002). SVMs Modeling for Highly Imbalanced Classfication. *Journal of Latex Class Files, 1*(11).

Zhang, Y., Wu, L., & Wang, S. (2011). Magnetic Resonance Brain Image Classification By An Improved Artificial Bee Colony Algorithm. *Progress In Electromagnetics Research, 116*, 67-79.

LAMPIRAN

Lampiran 1. Data Imbalanced Rasio Indikator Desa Tertinggal di Provinsi Jawa Timur Tahun 2014.

Desa	X_1	X_2	X_3	⋯	X_8	Status
1	1,261	0,03	0,03	⋯	0,473	0
2	2,632	0	0,073	⋯	0,583	0
3	1,863	0,029	0	⋯	0,345	0
4	2,564	0,035	0,035	⋯	0,212	0
5	1,645	0,021	0	⋯	0,295	0
6	1,304	0	0	⋯	0,542	0
7	1,961	0,033	0	⋯	0,955	0
8	1,242	0,049	0	⋯	1,234	0
9	0,99	0,032	0,032	⋯	0,956	0
10	1,245	0,028	0	⋯	1,452	0
11	0,893	0,021	0	⋯	1,358	0
12	1,068	0,044	0,022	⋯	2,368	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮
7710	0,847	0,018	0,018	⋯	11,724	0
7711	0,774	0,038	0	⋯	0,458	0
7712	0,323	0,018	0,009	⋯	2,074	0
7713	1,023	0,018	0	⋯	20,165	0
7714	0,922	0,016	0,016	⋯	7,987	0
7715	0,865	0,017	0,017	⋯	1,153	0
7716	0,858	0,03	0,015	⋯	1,215	0
7717	0,41	0,057	0,019	⋯	3,868	0
7718	0,592	0,043	0,011	⋯	21,099	0
7719	0,743	0,026	0,026	⋯	0,156	0
7720	0,423	0,01	0,01	⋯	4,643	0
7721	0,23	0,022	0,022	⋯	4,946	0

Lampiran 2. Data Balanced Rasio Indikator Desa Tertinggal di Provinsi Jawa Timur Tahun 2014.

Desa	X_1	X_2	X_3	⋯	X_8	Status
1	1,316	0,036	0,036	⋯	0,761	0
2	1,111	0,024	0,024	⋯	1,178	0
3	1,141	0	0,013	⋯	0,564	0
4	1,361	0,014	0,014	⋯	0,63	0
5	0,752	0	0,053	⋯	2,315	0
6	0,926	0	0,045	⋯	2,528	1
7	1,119	0,02	0,02	⋯	8,35	0
8	0,581	0	0,044	⋯	2,404	1
9	0,565	0,032	0,032	⋯	3,175	0
10	0,549	0,032	0	⋯	5,152	0
11	1,408	0,052	0,052	⋯	10,388	0
12	0,576	0,038	0,038	⋯	3,376	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮
403	1,333	0,022	0,022	⋯	0,224	0
404	0,781	0,035	0,035	⋯	1,046	0
405	1,083	0,05	0,05	⋯	0,248	1
406	1,095	0	0,031	⋯	0,187	1
407	0,93	0,084	0,084	⋯	0,335	1
408	0,752	0,043	0.043	⋯	0,174	1
409	1,695	0	0	⋯	0,559	1
410	0,862	0,137	0,137	⋯	0,412	1
411	0,935	0	0,144	⋯	0,432	1
412	1,015	0	0	⋯	0,323	1
413	0,803	0	0,081	⋯	0,244	1
414	1,333	0	0	⋯	1,094	1

Lampiran 3. Hasil Uji Homogenitas *Box's M Test*

a. Data Imbalaced dengan semua Variabel

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data)[, -9]
Chi-Sq (approx.) = 6625, df = 36, p-value < 2.2e-16
```

b. Data Imbalanced Variabel Signifikan

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data)[, -3]
Chi-Sq (approx.) = 2772.6, df = 3, p-value < 2.2e-16
```

c. Data Balanced Semua Variabel

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data)[, -9]
Chi-Sq (approx.) = 4344.7, df = 36, p-value < 2.2e-16
```

d. Data Balanced Semua Variabel

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data)[, -4]
Chi-Sq (approx.) = 428.38, df = 6, p-value < 2.2e-16
```

Lampiran 4. Hasil Uji Distribusi Normal Multivariat

a. Data Imbalanced dengan Semua Variabel

```
Mardia's Multivariate Normality Test
g1p            : 8328.206
chi.skew       : 10717013
p.value.skew   : 0
g2p            : 12575.38
z.kurtosis     : 43400.66
p.value.kurt   : 0
chi.small.skew : 10722103
p.value.small  : 0
Result         : Data are not multivariate normal.
```

b. Data Imbalanced Semua Variabel

```
Mardia's Multivariate Normality Test
  g1p            : 341.2051
chi.skew       : 439074
p.value.skew   : 0
g2p            : 686.0759
z.kurtosis     : 7447.751
p.value.kurt   : 0
chi.small.skew : 439358.4
p.value.small  : 0
Result         : Data are not multivariate normal.
```

c. Data Balanced Semua Variabel

$`multivariateNormality`

	Test	Statistic	p value	Result
1	Mardia Skewness	97249.25	0	NO
2	Mardia Kurtosis	1194.37	0	NO
3	MVN	<NA>	<NA>	NO

d. Data Balanced Variabel Signifikan

$`multivariateNormality`

	Test	Statistic	p value	Result
1	Mardia Skewness	12507.62	0	NO
2	Mardia Kurtosis	491.07	0	NO
3	MVN	<NA>	<NA>	NO

Lampiran 5. Output Regresi Logistik

a. Data Imbalanced Semua Variabel

```
Call:
glm(formula = Status ~ ., family = binomial(link =
"logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8054  -0.2245  -0.1872  -0.1642   3.5193

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.779744   0.865812   0.901    0.368
SD           0.066917   0.065513   1.021    0.307
Poskesdes    0.215291   0.195844   1.099    0.272
Bidan       -0.023465   0.091466  -0.257    0.798
Listrik     -5.419086   0.860916  -6.295 3.08e-10 ***
Toko        -0.003358   0.005921  -0.567    0.571
Jarak        0.137598   0.012349  11.142  < 2e-16 ***
Gizi.Buruk   0.017900   0.029964   0.597    0.550
PAD         -0.002283   0.002612  -0.874    0.382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1721.0  on 6948  degrees of freedom
Residual deviance: 1523.5  on 6940  degrees of freedom
AIC: 1541.5

Number of Fisher Scoring iterations: 8
```

Lampiran 5. Output Regresi Logistik (Lanjutan)

b. Data Imbalanced Variabel Signifikan

```
Call:
glm(formula = Status ~ ., family = binomial(link =
"logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8220  -0.2185  -0.1885  -0.1644   2.9826

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.83992    0.86395   0.972    0.331
Listrik     -5.41432    0.85961  -6.299    3e-10 ***
Jarak        0.13829    0.01232  11.227   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1721.0  on 6948  degrees of freedom
Residual deviance: 1526.9  on 6946  degrees of freedom
AIC: 1532.9

Number of Fisher Scoring iterations: 6
```

c. Data Balanced Semua Variabel

```
glm(formula = Status ~ ., family = binomial(link =
"logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6223  -0.7965   0.0000   0.8471   2.0832
```

Lampiran 5. Output Regresi Logistik (Lanjutan)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.342261   2.755939   1.213  0.22523
SD           0.505781   0.194888   2.595  0.00945 **
Bidan       -0.890592   4.048071  -0.220  0.82587
Poskesdes    6.767307   4.204437   1.610  0.10749
Toko        -0.654844   0.237195  -2.761  0.00577 **
Listrik     -5.344518   2.781023  -1.922  0.05463 .
Jarak        0.232118   0.031632   7.338 2.17e-13 ***
Gizi.Buruk   1.475452   0.994149   1.484  0.13777
PAD          0.003714   0.003182   1.167  0.24324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 517.09  on 372  degrees of freedom
Residual deviance: 385.59  on 364  degrees of freedom
AIC: 403.59

Number of Fisher Scoring iterations: 12
```

d. Data Balanced Variabel Signifikan

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.81052    2.79688   1.362   0.1731
SD           0.45696    0.18946   2.412   0.0159 *
Listrik     -6.06962    2.81586  -2.156   0.0311 *
Jarak        0.23263    0.03081   7.551 4.33e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 517.09  on 372  degrees of freedom
```

Lampiran 6. Output Regresi Logistik Ridge

a. Data Imbalanced Semua Variabel

```
Call:
logisticRidge(formula = Status ~ ., data = as.data.frame(t
rain),
    lambda = "automatic")

Coefficients:
```

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	1.33E-01	NA	NA	NA	NA	
SD	7.52E-02	5.03E+00	2.62E+00	1.92	0.0549	.
Poskesdes	1.83E-02	1.72E+00	1.98E+00	0.868	0.3856	
Bidan	1.51E-03	2.01E-01	2.16E+00	0.093	0.9257	
Listrik	-4.31E+00	-1.35E+01	2.22E+00	-6.101	1.05E-09	***
Toko	-3.02E-04	-1.17E+00	1.81E+00	-0.648	0.5171	
Jarak	7.74E-02	2.91E+01	2.67E+00	10.894	< 2e-16	***
Gizi.Buruk	2.48E-03	1.35E+00	2.22E+00	0.609	0.5424	
PAD	-7.90E-05	-9.82E-01	1.88E+00	-0.523	0.601	

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Ridge paramter: 0.01732251, chosen automatically, computed
using 3 PCs

Degrees of freedom: model 4.412 , variance 4.031
```

Lampiran 6. Output Regresi Logistik Ridge (Lanjutan)

b. Data Imbalanced Variabel Signifikan

```
Call:
logisticRidge(formula = Status ~ ., data = as.data.
frame(train),
    lambda = "automatic")

Coefficients:
```

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	1.03E+00	NA	NA	NA	NA	
Listrik	-5.34E+00	-1.72E+01	2.46E+00	-6.999	2.58E-12	***
Jarak	1.04E-01	4.26E+01	3.86E+00	11.035	< 2e-16	***

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.
’ 0.1 ‘ ’ 1

Ridge paramter: 0.006460343, chosen automatically,
computed using 2 PCs

Degrees of freedom: model 2.651 , variance 3.837
```

c. Data Balanced Semua Variabel

```
Call:
logisticRidge(formula = Status ~ ., data = as.data.frame(t
rain),
    lambda = "automatic")
Coefficients:
```

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	3.00E-02	NA	NA	NA	NA	
SD	4.19E-03	6.96E-02	2.42E-02	2.882	0.00395	**
Bidan	-1.67E-04	-4.48E-03	2.33E-02	-0.192	8.48E-01	
Poskesdes	-1.14E-04	-3.06E-03	2.33E-02	-0.131	0.89575	

Lampiran 6. Output Regresi Logistik Ridge (Lanjutan)

Toko	-1.53E-05	-1.66E-02	2.34E-02	-0.709	0.47863	
Listrik	-4.72E-02	-8.99E-02	2.41E-02	-3.723	1.97E-04	***
Jarak	7.47E-04	1.47E-01	2.41E-02	6.096	1.09E-09	***
Gizi.Buruk	2.92E-04	2.42E-02	2.39E-02	1.012	0.31139	
PAD	-4.03E-06	-7.71E-03	2.35E-02	-0.328	0.74276	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge paramter: 10.22287, chosen automatically, computed using 1 PCs

Degrees of freedom: model 1.095 , variance 1.003

d. Data Balanced Variabel Signifikan

Call:
logisticRidge(formula = Status ~ ., data = as.data.frame(train),
lambda = "automatic")

Coefficients:

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	3.60E+00	NA	NA	NA	NA	
SD	4.44E-01	7.38E+00	3.01E+00	2.453	0.0142	*
Listrik	-5.69E+00	-1.08E+01	4.68E+00	-2.317	2.05E-02	*
Jarak	2.10E-01	4.13E+01	5.20E+00	7.948	2.00E-15	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Ridge paramter: 0.001576135, chosen automatically, computed using 3 PCs

Degrees of freedom: model 3.802 , variance 9.

Lampiran 7. Uji Asumsi Analisis Diskriminan Kernel

a. Syntax Uji Homogenitas

```
library(biotools)
data=read.csv("E:/Tugas Akhir/Data/Data V.3/data
semua.csv",header=T,sep=";")
head(data)

Uji_Homogen<-boxM(data.matrix(data)[,-9], data[,9])
Uji_Homogen
```

b. Syntax Uji Normal Multivariat

```
library(MVN)
data=read.csv("E:/Tugas Akhir/Data/Data V.3/data
semua.csv",header=T,sep=";")
head(data)

uji=mvn(data.matrix(data[,-9]), subset = NULL, mvnTest =
c("mardia", "hz", "royston", "dh",
                          "energy"), covariance = TRUE, tol = 1e-25,
alpha = 0.5, scale = FALSE,
   desc = TRUE, transform = "none", R = 1000, univariateTest =
c("SW",
                                                "CVM", "Lillie", "SF",
"AD"), univariatePlot = "none",
   multivariatePlot = "qq", multivariateOutlierMethod = "none",
   showOutliers = FALSE, showNewData = FALSE)
uji
```

Lampiran 8. Syntax Combine Undersampling

a. Tomek Links

```
library(unbalanced)
data<-read.csv("E:/Tugas Akhir/Data/Data V.3/data semua.csv",
header=TRUE, sep=";")
head(data)

set.seed(1234)
print(table(data$Status))

Y=as.factor(data$Status)
X=data[,-9]

databaru=ubTomek(X, Y, verbose = TRUE)
newdata=cbind(databaru$X, databaru$Y)
print(table(databaru$Y))

write.csv(newdata, file = "E:/Tugas
Akhir/Data/Hasil/R_TomekLinks.csv")
```

b. Random Undersampling

```
library(unbalanced)
data<-read.csv("E:/Tugas Akhir/Data/Hasil/R_TomekLinks.csv",
header=TRUE, sep=";")
head(data)
set.seed(1234)
print(table(data$Status))

Y=as.factor(data$Status)
X=data[,-9]

databaru=ubUnder(X, Y, perc= 50, method =  "percPos")
newdata=cbind(databaru$X, databaru$Y)
print(table(databaru$Y))

write.csv(newdata, file = "E:/Tugas Akhir/Data/Hasil/R_RUS.csv")
```

Lampiran 9. Syntax Regresi Logistik

```
library(data.table)
library(caret)
library(MASS)
library(MXM)
library(glmnet)
library(e1071)
#DATA
data<-read.csv("E:/Tugas Akhir/Data/Hasil/R_RUS.csv",
header=TRUE, sep=";")
head(data)
Y<-as.factor(data$Status)
for(i in 1:r)
{
 train=data[-fold[[i]],]
 test=data[fold[[i]],]
 model = glm(Status~., data=train, family = binomial(link='logit'))
 predtrain=round(predict(model, train[,-10], type = "response"))
 predtest=round(predict(model, test[-10], type="response"))
 tabel1=table(train$Status, predtrain)
 tabel2=table(test$Status, predtest)
 TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
 SensTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
 SpesTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))
 TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
 SensTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
 SpesTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))
 AUCTrain[i]=1/2*(SensTrain[i]+SpesTrain[i])
 AUCTest[i]=1/2*(SensTest[i]+SpesTest[i])
 GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
 GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
mean(TotalAccuracyTrain)
mean(TotalAccuracyTest)
mean(SensTrain)
mean(SensTest)
mean(SpesTrain)
mean(SpesTest)
```

Lampiran 9. Syntax Regresi Logistik (Lanjutan)

```
mean(GmeanTrain)
mean(GmeanTest)
#export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                      TotalAccuracyTest,
SensTrain,
                      SensTest,
                      SpesTrain,
                      SpesTest,
                      Gmean Train,
                      Gmean Test)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                     mean(TotalAccuracyTest),
                     mean(SensTrain),
                     mean(SensTest),
                     mean(SpesTrain),
                     mean(SpesTest),
                     mean(Gmean Train),
                     mean(Gmean Test))
write.csv(hasiltotal, file = "E:/Tugas
Akhir/Data/Hasil/Reglog_R_RUS_hasil_total.csv")
write.csv(hasilmean,file="E:/Tugas
Akhir/Data/Hasil/Reglog_R_RUS_hasil_total_mean.csv")
summary(model)
```

Lampiran 10. Syntax Regresi Logistik Ridge

```
library(data.table)
library(caret)
library(MASS)
library(MXM)
library(ridge)
library(e1071)

#DATA
data<-read.csv("E:/Tugas Akhir/Data/Hasil/R_RUS.csv",
header=TRUE, sep=";")
head(data)
Y<-as.factor(data$Status)
#CROSS VALIDASI
r=10
fold=generatefolds(Y, nfolds=r, stratified=TRUE,seed = 12345)
TotalAccuracyTrain=rep(0,r)
SensTrain=rep(0,r)
SpesTrain=rep(0,r)
TotalAccuracyTest=rep(0,r)
SensTest=rep(0,r)
SpesTest=rep(0,r)
AUCTrain=rep(0,r)
AUCTest=rep(0,r)
GmeanTrain=rep(0,r)
GmeanTest=rep(0,r)
#MODEL REGRESI LOGISTIK RIDGE
for(i in 1:r)
{
  train=data[-fold[[1]],]
  test=data[fold[[1]],]
  model = logisticRidge(Status~., data=as.data.frame(train), lambda
="automatic" )
  predtrain=round(predict(model, train[,-9], type = "response"))
  predtest=round(predict(model, test[-9], type="response"))
  tabel1=table(train$Status, predtrain)
  tabel2=table(test$Status, predtest)
  TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
```

Lampiran 10. Syntax Regresi Logistik Ridge (Lanjutan)

```
  SpesTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SensTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))
  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  SpesTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
  SensTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))
  GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
  GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
mean(TotalAccuracyTrain)
mean(TotalAccuracyTest)
mean(SensTrain)
mean(SensTest)
mean(SpesTrain)
mean(SpesTest)
mean(GmeanTrain)
mean(GmeanTest)
#export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                      TotalAccuracyTest,
                      SensTrain,
                      SensTest,
                      SpesTrain,
                      SpesTest,
                      GmeanTrain,
                      GmeanTest)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                     mean(TotalAccuracyTest),
                     mean(SensTrain),
                     mean(SensTest),
                     mean(SpesTrain),
                     mean(SpesTest),
                      mean(GmeanTrain),
                     mean(GmeanTest))
write.csv(hasiltotal, file = "E:/Tugas
Akhir/Data/Hasil/ReglogRidge_R_RUS_hasil_total.csv")
write.csv(hasilmean,file="E:/Tugas
Akhir/Data/Hasil/ReglogRidgeRUS_R_RUS_hasil_total_mean.csv")
summary(model)
```

Lampiran 11. Syntax Analisis Diskriminan Kernel

```
library(data.table)
library(caret)
library(kernlab)
library(MASS)
library(kfda)
library(MXM)

#DATA
data<-read.csv("E:/Tugas Akhir/Data/Hasil/R_RUS.csv",
header=TRUE, sep=";")
head(data)
Y<-as.factor(data$Status)
#CROSS VALIDASI
r=10
fold=generatefolds(Y, nfolds=r, stratified=TRUE,seed = 12345)
TotalAccuracyTrain=rep(0,r)
SensTrain=rep(0,r)
SpesTrain=rep(0,r)
TotalAccuracyTest=rep(0,r)
SensTest=rep(0,r)
SpesTest=rep(0,r)
AUCTrain=rep(0,r)
AUCTest=rep(0,r)
GmeanTrain=rep(0,r)
GmeanTest=rep(0,r)
#MODEL ANDISKER
for(i in 1:r)
{
  train=data[-fold[[1]],]
  test=data[fold[[1]],]
  model = kfda(trainData=train,kernel.name="rbfdot")
  predtrain1=kfda.predict(model, train)
  predtrain2=predtrain1$class
  predtrain3=as.vector(predtrain2)
  predtrain=as.numeric(predtrain3)
```

Lampiran 11. Syntax Analisis Diskriminan Kernel (Lanjutan)

```
  predtest1=kfda.predict(model, test)
  predtest2=predtest1$class
  predtest3=as.vector(predtest2)
predtest=as.numeric(predtest3)
  write.csv(predtest ,"E:/Tugas Akhir/Hasil/dataRUS2_test.csv")

  tabel1=table(train$Status, predtrain)
  tabel2=table(test$Status, predtest)

  TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
  SensTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SpesTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))

  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  SensTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
  SpesTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))

  AUCTrain[i]=1/2*(SensTrain[i]+SpesTrain[i])
  AUCTest[i]=1/2*(SensTest[i]+SpesTest[i])

  GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
  GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
mean(TotalAccuracyTrain)
mean(TotalAccuracyTest)
mean(SensTrain)
mean(SensTest)
mean(SpesTrain)
mean(SpesTest)
mean(AUCTrain)
mean(AUCTest)
mean(GmeanTrain)
mean(GmeanTest)
summary(model)
  #export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                      TotalAccuracyTest,
```

Lampiran 11. Syntax Analisis Diskriminan Kernel (Lanjutan)

```
                    SensTrain,
                    SensTest,
                    SpesTrain,
                    SpesTest,
                    AUCTrain,
                    AUCTest,
                    GmeanTrain,
                    GmeanTest)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                    mean(TotalAccuracyTest),
                    mean(SensTrain),
                    mean(SensTest),
                    mean(SpesTrain),
                    mean(SpesTest),
                    mean(AUCTrain),
                    mean(AUCTest),
                    mean(GmeanTrain),
                    mean(GmeanTest))
write.csv(hasiltotal, file = "E:/Tugas
Akhir/Data/Hasil/Andisker_R_RUS_hasil_total.csv")
write.csv(hasilmean,file="E:/Tugas
Akhir/Data/Hasil/Andisker_RUS_R_RUS_hasil_total_mean.csv")
summary(model)
```

Lampiran 12. Syntax Regresi Logistik Backward

```
library(data.table)
library(caret)
library(MASS)
library(MXM)
library(glmnet)
library(e1071)
#COBA
data<-read.csv("E:/Tugas Akhir/Data/R_RUS.csv", header=TRUE,
sep=";")
head(data)
Y<-as.factor(data$Status)

#CROSS VALIDASI
r=10
fold=generatefolds(Y, nfolds=r, stratified=TRUE,seed = 12345)
TotalAccuracyTrain=rep(0,r)
SensTrain=rep(0,r)
SpesTrain=rep(0,r)
TotalAccuracyTest=rep(0,r)
SensTest=rep(0,r)
SpesTest=rep(0,r)
AUCTrain=rep(0,r)
AUCTest=rep(0,r)
GmeanTrain=rep(0,r)
GmeanTest=rep(0,r)

#MODEL REGRESI LOGISTIK
for(i in 1:r)
{
  train=data[-fold[[i]],]
  test=data[fold[[i]],]
  model = glm(Status~., data=train, family = binomial(link='logit'))
  model= step(model,direction="backward",trace = FALSE)

  predtrain=round(predict(model, train[,-9], type = "response"))
  predtest=round(predict(model, test[-9], type="response"))

  tabel1=table(train$Status, predtrain)
```

Lampiran 12. Syntax Regresi Logistik Backward (Lanjutan)

```
  tabel2=table(test$Status, predtest)

  TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
  SpesTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SensTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))

  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  predtest=round(predict(model, test[-9], type="response"))

  tabel1=table(train$Status, predtrain)
  tabel2=table(test$Status, predtest)

 TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
  SpesTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SensTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))

  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  SpesTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
  SensTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))

  AUCTrain[i]=1/2*(SensTrain[i]+SpesTrain[i])
  AUCTest[i]=1/2*(SensTest[i]+SpesTest[i])

  GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
  GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
mean(TotalAccuracyTrain)
mean(TotalAccuracyTest)
mean(SensTrain)
mean(SensTest)
mean(SpesTrain)
mean(SpesTest)
mean(AUCTrain)
mean(AUCTest)
mean(GmeanTrain)
mean(GmeanTest)
summary(model)
```

Lampiran 12. Syntax Analisis Diskriminan Kernel (Lanjutan)

```
                #export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                TotalAccuracyTest,
                SensTrain,
                SensTest,
                SpesTrain,
                SpesTest,
                AUCTrain,
                AUCTest,
                GmeanTrain,
                GmeanTest)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                mean(TotalAccuracyTest),
                mean(SensTrain),
                mean(SensTest),
                mean(SpesTrain),
                mean(SpesTest),
                mean(AUCTrain),
                mean(AUCTest),
                mean(GmeanTrain),
                mean(GmeanTest))
hasilmean
write.csv(hasiltotal, file = "E:/Tugas
Akhir/Data/Hasil/R_RUSReglogback_hasil_total.csv")
write.csv(hasilmean,file="E:/Tugas
Akhir/Data/Hasil/R_RUSReglogback_hasil_total_mean.csv")
summary(model)
```

Lampiran 13. Surat Pernyataan Permintaan Data

BADAN PUSAT STATISTIK
PROVINSI JAWA TIMUR

SENSUS EKONOMI 2016

SURAT KETERANGAN

Yang bertanda tangan dibawah ini :

N a m a	:	Thomas Wunang Tjahjo, M.Sc, M.Eng.
N I P	:	19700329 1992 11 1 001
Jabatan	:	Kepala Bidang Integrasi Pengolahan dan Diseminasi Statistik

Dengan ini menerangkan bahwa :

N a m a	:	Rahma Shintia
Fakultas/Program Studi	:	Fakultas Matematika, Komputasi dan Sains Data / Statistika
N.R.P	:	06211440000054
Alamat Rumah	:	Perumahan Dosen Institut Teknologi Sepuluh November Blok. U No.149, Kec. Sukolilo, Surabaya
Akademi / Universitas	:	Institut Teknologi Sepuluh Nopember (ITS) Telp (031) 594 3352, (031) 599 4251-55 Fax (031) 592 2940

Benar-benar telah mencari data di Kantor Badan Pusat Statistik (BPS) Provinsi Jawa Timur dalam rangka menyusun Tugas Akhir / Skripsi dengan judul :

" Penerapan Undersampling Tomek Links Untuk Klasifikasi Regresi Logistik Ridge dan Analisis Diskriminan Kernel pada Imbalanced Data (Studi Kasus : Desa Tertinggal di Jawa Timur 2014) "

Demikian surat keterangan ini dibuat dan agar dipergunakan sebagaimana mestinya

Surabaya, 2 Mei 2018

An. Kepala BPS Provinsi Jawa Timur

Kepala Bidang IPDS

Thomas Wunang Tjahjo, M.Sc, M.Eng.

Jalan Raya Kendangsari Industri No. 43 - 44, Surabaya - 60292
Telp. 031 - 8439343 Fax. 031 - 8494007. Homepage: http://jatim.bps.go.id E-mail: bps3500@bps.go.id

(*Halaman ini sengaja dikosongkan*)

BIODATA PENULIS

Penulis dengan nama lengkap Rahma Shintia dilahirkan di Kabupaten Agam pada 29 Agustus 1996. Penulis menempuh pendidikan formal di SDN 14 Tanjung Alam, SMPN 6 Bukittinggi, dan SMAN 1 Bukittinggi. Kemudian penulis diterima sebagai Mahasiswa Departemen Statistika ITS melalui jalur SNMPTN pada tahun 2014. Selama masa perkuliahan, penulis aktif di berbagai kepanitiaan salah satunya adalah panitia sie acara *Statistics Competition* (STATION) 2016 yang merupakan olimpiade statistika bagi murid SMA dan sederajat. Selain itu, penulis juga aktif dalam organisasi yang menaungi Departemen Statistika yaitu sebagai staff keilmiahan HIMASTA-ITS 2015/2016, staff Divisi PERS HIMASTA-ITS 2015/2016, Sekretaris keilmiahan HIMASTA-ITS 2016/2017, dan staff ahli divisi PERS HIMASTA-ITS 2016/2017. Dibidang akademik, penulis diberi kesempatan untuk menjadi semifinalis lomba karya tulis ilmiah dari Jurusan Statistika Universitas Padjajaran pada PADJADJARAN STATISTICS OLYMPIAD (RASIO) 2017, finalis lomba karya tulis ilmiah dari Jurusan Teknik Informatika Universitas Udayana dan pada tahun 2018, penulis mampu menjadi Juara II *Indonesian Research Competition dalam 3rd ISCO 2017*, dan mendapat penghargaan *best gold paper* pada *Regional Conference on Student Activism* (RECONSA) 2018 di Universiti Teknologi Petronas. Selain itu, penulis juga berkesempatan mendapatkan pendanaan penelitian berupa PKM Pengabdian Masyarakat pada tahun 2016. Apabila pembaca ingin memberi kritik dan saran serta diskusi lebih lanjut mengenai Tugas Akhir ini, dapat menghubungi penulis melalui email rahmashintia298@gmail.com atau nomor telepon 081335785296.