TUGAS AKHIR – SS141501

PENERAPAN METODE *COMBINE SAMPLING* PADA KLASIFIKASI *IMBALANCED* DATA BINER STATUS KETERTINGGALAN DESA DI JAWA TIMUR

DEWI LUTFIA PRATIWI
NRP 062114 40000 054

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 METODE *COMBINE SAMPLING* PADA KLASIFIKASI *IMBALANCED* DATA BINER STATUS KETERTINGGALAN DESA DI JAWA TIMUR

DEWI LUTFIA PRATIWI
NRP 062114 4000 054

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

THE APPLICATION OF COMBINE SAMPLING METHOD TO CLASSIFICATION OF IMBALANCED BINARY DATA IN UNDERDEVELOPED VILLAGES STATUS OF EAST JAVA

DEWI LUTFIA PRATIWI
SN 062114 4000 054

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 :
Dewi Lutfia Pratiwi
NRP. 062114 4000 0054

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 METODE *COMBINE SAMPLING* PADA KLASIFIKASI *IMBALANCED* DATA BINER STATUS KETERTINGGALAN DESA DI JAWA TIMUR

Nama Mahasiswa : Dewi Lutfia Pratiwi
NRP : 062114 40000 054
Departemen : Statistika
Dosen Pembimbing : Dr. Santi Puteri Rahayu

Abstrak

Permasalahan kesenjangan pembangunan antar daerah di Indonesia masih perlu diperhatikan, seperti masih adanya desa tertinggal di beberapa provinsi di Indonesia dimana salah satunya berada di Jawa Timur. Penelitian ini bertujuan meng-klasifikasikan desa tertinggal di Jawa Timur berdasarkan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi, sehingga klasifikasi desa tertinggal dapat dilakukan dengan tepat. Salah satu masalah dalam klasifikasi data adalah komposisi data yang tidak seimbang (imbalanced data), dimana dapat diatasi menggunakan penggabungan metode oversampling SMOTE dan metode undersampling Tomek Links. Perbandingan hasil klasifikasi antara data imbalanced dengan data balanced menunjukkan bahwa metode Regresi Logistik, Regresi Logistik Ridge, maupun Analisis Diskriminan Kernel memiliki nilai AUC, G-mean dan sensitivitas yang meningkat setelah dilakukan balancing data, dimana peningkatan tertinggi pada sensitivitas sebesar 23 kali dan ketiga metode klasifikasi memiliki hasil ketepatan klasifikasi yang comparable. Tetapi jika dibandingkan, metode Regresi Logistik Ridge memiliki AUC, G-mean dan akurasi total yang lebih tinggi pada data balanced dengan memasukkan semua variabel yaitu 78%, 77,91% dan 78,1%. Sehingga klasifikasi status ketertinggalan desa baik diklasifikasikan dengan metode Regresi Logistik Ridge.

Kata Kunci : Combine Sampling, Desa Tertinggal, Imbalanced Data, Klasifikasi

(*Halaman ini sengaja dikosongkan*)

THE APPLICATION OF COMBINE SAMPLING METHOD TO CLASSIFICATION OF IMBALANCED BINARY DATA IN UNDERDEVELOPED VILLAGES STATUS OF EAST JAVA

Name : Dewi Lutfia Pratiwi
Student Number : 062114 40000 054
Department : Statistics
Supervisor : Dr. Santi Puteri Rahayu

Abstract

The problems of development gap between regions in Indonesia still need to be considered, as there are still underdeveloped villages in several provinces in Indonesia where one of them is in East Java. The aim of this research is to classify backward villages in East Java based on 5 districts that have the highest percentage of underdeveloped villages, so that the classification of underdeveloped villages can be done appropriately. One of the problems in data classification is the unbalanced data composition (imbalanced data), which can be solved using combine between oversampling SMOTE and undersampling Tomek Links. The comparison of the classification results between the imbalanced data and the balanced data indicates that the Logistic Regression, Ridge Logistic Regression, and Kernel Discriminant Analysis have AUC, G-mean and sensitivity have an increased value after balancing data, where sensitivity has the highest increase that is 23 times and all classifier method have comparable classification accuracy result. But when compared, the Ridge Logistic Regression method has higher AUC, G-mean and accuracy on the balanced data by including all the variables that is 78%, 77,9% and 78,1%. So the classification of the underdeveloped status of the village is well classified by the Ridge Logistic Regression method.

.***Keywords : Classification, Combine Sampling, Imbalanced Data, Underdeveloped Villages***

(*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 Metode *Combine Sampling* pada Klasifiasi *Imbalanced* Data Biner Status Ketertinggalan Desa di Jawa Timur"** 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. M. Sjahid Akbar, S.Si., M.Si.dan Dr.rer.pol. Heri Kuswanto, S.Si., M.Si. selaku dosen penguji yang telah banyak memberi masukan kepada penulis.
4. Dr. Purhadi, M.Sc. 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 Rahma Shintia dan Canggih Shoffi Imanwardhani yang selama ini telah berjuang bersama dan saling memberikan semangat.
7. Sahabat-sahabat penulis, Bekti Indasari, Maulina Firdaus, Intan Maharani, Siti Halimah, Muhammad Lukman, dan Taufik Afiif yang selama ini telah membantu, mendukung, dan mendengarkan keluh kesah penulis selama masa perkuliahan berlangsung.

8. Teman-teman Statistika ITS angkatan 2014, Respect, yang selalu memberikan dukungan kepada penulis selama ini.
9. 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

BAB I

PENDAHULUAN

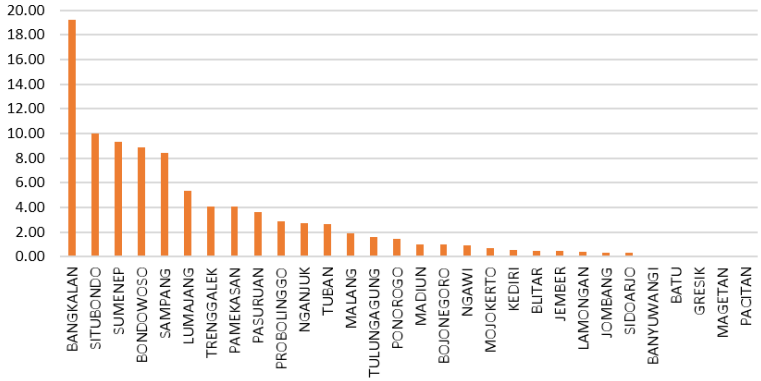
1.1 Latar Belakang

Kesenjangan antar wilayah dan ketertinggalan suatu daerah masih menjadi isu yang harus diatasi sejalan dengan tujuan pembangunan nasional. Penyebab terjadinya kesenjangan tersebut sangatlah beragam mulai dari perbedaan ketersediaan sumber daya alam, letak geografis, kualitas sumber daya manusia, kemajuan ekonomi, hingga pada aspek sosial budaya. Kesenjangan pembangunan tersebut ditunjukkan dengan masih adanya daerah-daerah yang tingkat perkembangannya masih tertinggal dibandingkan daerah lainnya dengan kata lain keberadaan daerah tertinggal sebagai indikator adanya kesenjangan dalam pembangunan.

Tahun 2015 pemerintah membentuk kementrian desa, pembangunan daerah tertinggal, dan transmigrasi melalui Peraturan Presiden Nomor 12 Tahun 2015 yang berfokus pada pembangunan desa. Fokus pemerintah terhadap pembangunan desa ini menarik untuk diperhatikan karena pembangunan yang tepat sasaran merupakan hal yang mutlak diperlukan. Desa tertinggal yaitu desa yang mempunyai ketersediaan dan akses terhadap pelayanan dasar, infrastruktur, aksesibilitas/transportasi, pelayanan umum, dan penyelenggaraan pemerintahan yang masih minim (Bappenas & Badan Pusat Statistik , 2015). Pemerintah hingga saat ini terus berupaya meningkatkan pemerataan pembangunan pada daerah-daerah tertinggal di Indonesia. Dalam RPJMN 2015-2019, target Kementrian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi adalah mengurangi jumlah desa tertinggal hingga 5000 desa dan meningkatkan jumlah desa mandiri sedikitnya 2000 desa.

Menurut data PODES (Potensi Desa) tahun 2014 dari 73.709 desa di Indonesia, terdapat 36.838 atau 49,98% desa yang berstatus tertinggal dan 2.047 atau 2,78% yang berstatus sangat tertinggal. Jumlah desa tertinggal paling banyak di Pulau Jawa-Bali terdapat

di Provinsi Jawa Timur yaitu sebanyak 208 unit desa (Bappenas & Badan Pusat Statistik , 2015). Jawa Timur memiliki wilayah terluas di antara 6 provinsi di Pulau Jawa, dan memiliki jumlah penduduk terbanyak kedua di Indonesia setelah Jawa Barat. Jika dibandingkan dengan provinsi lain di pulau Jawa, Jawa Timur masih memiliki daerah tertinggal. Menurut Perpres Nomor 131 tahun 2015, dari 122 kabupaten tertinggal terdapat 4 kabupaten diantaranya berasal dari Jawa Timur yaitu Kabupaten Bondowoso, Kabupaten Situbondo, Kabupaten Bangkalan dan Kabupaten Sampang (Kemendesa, 2016). Apabila dilihat dari data PODES tahun 2014, persentase kabupaten yang memiliki daerah tertinggal di Jawa Timur dapat dilihat dari Gambar 1.1 berikut (BPS, 2015).



Gambar 1. 1 Persentase Desa Tertinggal Tiap Kabupaten di Jawa Timur

Gambar 1.1 menunjukkan bahwa jika dilihat dari 5 kabupaten yang memiliki persentase desa tertinggal tertinggi, terdapat Kabupaten Bangkalan di urutan pertama, kemudian Kabupaten Situbondo, Kabupaten Sumenep, Kabupaten Bondowoso, dan Kabupaten Sampang.

Berdasarkan hal tersebut, diperlukan program pembangunan daerah tertinggal yang difokuskan pada percepatan pembangunan daerah yang kondisi sosial, budaya, ekonomi, keuangan daerah, aksesibilitas, serta ketersediaan infrastruktur masih tertinggal dibanding dengan daerah lainnya sehingga pembangunan yang

dilakukan pemerintah akan tepat sasaran. Oleh karena itu, diperlukan ketepatan klasifikasi desa tertinggal di Jawa Timur berdasarkan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi untuk membantu menyelesaikan permasalahan-permasalahan yang ada, karena dimungkinkan terjadi pemekaran di beberapa desa, sehingga klasifikasi desa tertinggal masih belum tepat dilakukan.

BPS pernah melakukan pengklasifikasian status ketertinggalan desa pada tahun 1993, 1994, dan 1995. Kemudian pada tahun 2005, BPS melakukan pengklasifikasian desa tertinggal kembali dalam rangka penyaluran bantuan pemerintah yaitu kompensasi BBM yang dilakukan sampai saat ini. Data yang digunakan pada tahun 1993 adalah data PODES dengan faktor penentu daerah perkotaan sebanyak 25 variabel dan sebanyak 27 variabel untuk daerah pedesaan. Selanjutnya pada tahun 1994 data yang digunakan adalah data PODES dengan faktor penentu untuk daerah perkotaan sebanyak 17 variabel dan untuk daerah pedesaan sebanyak 18 variabel. Sedangkan pada tahun 2005 data yang digunakan adalah data PODES dan SUSENAS, dimana hasil yang diperoleh adalah model yang digunakan sebagai penentu status ketertinggalan desa. Tetapi hasil klasifikasi yang dilakukan BPS memiliki bias pada desa pertanian karena belum mencakup desa-desa hasil pemekaran. Perlu disadari bahwa dominasi rumah tangga miskin hanya terdapat pada 51% desa tertinggal (Agusta, 2007). Pada tahun 2014, klasifikasi status ketertinggalan desa dilakukan oleh BPS menggunakan metode PCA (*Principal Component Analysis*) dengan menghitung IPD (Indeks Pembangunan Desa) berdasarkan faktor-faktor yang mempengaruhi ketertinggalan desa dari hasil PCA, dimana dihasilkan 42 indikator yang terdapat pada 12 variabel (BPS, 2015). Klasifikasi desa tersebut terdiri dari tiga kategori yaitu desa tertinggal, desa berkembang, dan desa mandiri. Secara teknis, desa dikatakan tertinggal ketika memiliki IPD kurang dari sama dengan 50, untuk desa berkembang memiliki IPD lebih dari 50 namun kurang dari atau sama dengan 75, sedangkan desa mandiri

memiliki IPD lebih dari 75 (Bappenas & Badan Pusat Statistik, 2015). Tetapi pada penelitian ini status ketertinggalan desa dikategorikan menjadi 2 yaitu desa tertinggal dan desa tidak tertinggal (desa berkembang dan desa mandiri). Penelitian mengenai desa tertinggal sudah pernah dilakukan beberapa peneliti. Sambodo, Purnami, dan Rahayu (2013) melakukan penelitian tentang ketepatan klasifikasi status ketertinggalan desa dengan pendekatan *Reduce Support Vector Machine* (RSVM) di Provinsi Jawa Timur menggunakan jumlah data sebanyak 8502 desa dengan rasio 1:1 dan menghasilkan akurasi sebesar 50,25%. Kemudian Sulasih dan Purnami (2016) melakukan penelitian tentang *Rare Event Weighted Logistic Regression* (RE-WLR) untuk klasifikasi *imbalanced* data dengan studi kasus klasifikasi desa tertinggal di Provinsi Jawa Timur. Jumlah data yang digunakan adalah sebanyak 7721 desa dengan rasio 1:36 dan akurasi tertinggi pada $\lambda = 2$ yaitu sebesar 98,06 %.

Metode statistika yang dapat digunakan untuk mengklasifikasikan desa tertinggal di Jawa Timur salah satunya adalah Regresi Logistik. Regresi Logistik merupakan salah satu metode klasifikasi klasik yang bertujuan untuk mengetahui hubungan variabel respon yang bersifat kualitatif dengan variabel prediktor yang bersifat kualitatif ataupun kuantitatif dan digunakan pada respon dataset yang proporsional (Agresti, 2002). Ada beberapa kelebihan dari Regresi Logistik yaitu ketika T*runcated Regularised Iteratively Reweighted Least Square* (TR-IRLS) diimplementasikan pada Regresi Logistik, maka hasilnya akan efektif untuk klasifikasi data berskala besar dan akurasinya sebanding dengan SVM (Maalouf & Trafalis, 2011). Selain itu, Regresi Logistik hanya memerlukan pemecahan masalah secara *unconstrained optimation* dan secara alami memberikan probabilitas keanggotaan klasifikasi, dimana dengan menggunakan algoritma yang tepat maka proses perhitungannya akan lebih cepat dibandingkan metode lain seperti SVM (*Support Vector Machine*) yang memerlukan pemecahan *constrained quadratic optimation* (Maalouf & Siddiqi, 2014). Masalah dalam Regresi Logistik salah

satunya adalah adanya korelasi tinggi antar variabel (multikolinearitas). Multikolinearitas pada regresi logistik menyebabkan estimasi parameter, estimasi standar error, dan p-*value* hasil uji hipotesis menjadi tidak reliabel dan menyebabkan kesimpulan yang tidak akurat (Rossi, 2009). Oleh karena itu, digunakan Regresi Logistik Ridge untuk mengakomodasi kasus multikolinearitas tersebut, dimana dilakukan permodelan dengan menambahkan suatu bilangan positif kecil yang disebut *ridge parameter* pada estimasi parameter. Sebelumnya telah dilakukan penelitian oleh Maumere dan Ratnasari (2015) tentang Permodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur dengan menggunakan metode Regresi Logistik Ridge dimana jumlah datanya sebanyak 38 kabupaten/kota dengan rasio data 1:5 dan diperoeh akurasi tinggi yaitu sebesar 97,37% . Selain itu Sunyoto, Setiawan, dan Zain (2009) pernah menggunakan metode Regresi Logistik Ridge untuk menentukan keberhasilan siswa SMA Negeri 1 Kediri yang diterima di Perguruan Tinggi Negeri.

Salah satu masalah dalam klasifikasi data adalah komposisi data yang tidak seimbang (*imbalanced data*). Pada klasifikasi data dengan dua kelas, salah satu kelas memiliki jumlah sampel lebih besar dari kelas lainnya. Kelas data yang banyak disebut kelas mayoritas atau kelas negatif, sedangkan kelas data yang sedikit disebut kelas minoritas atau kelas positif. Dapat diketahui bahwa menurut data PODES tahun 2014, rasio jumlah desa tertinggal dibandingkan desa tidak tertinggal pada 5 kabupaten yang memiliki persentase desa tertinggal tertinggi adalah sebesar 1:9. Hal ini menunjukkan bahwa terjadi ketidakseimbangan data yang cukup tinggi. Permasalahan yang sering terjadi pada data *imbalanced*, *classifier* cenderung memprediksi kelas yang memiliki komposisi data lebih besar. Akibatnya dihasilkan hasil akurasi prediksi yang baik pada kelas data training yang mayoritas, sedangkan akan dihasilkan akurasi prediksi yang buruk pada data training yang minoritas (Japkowicz & Stephen, 2002 ; Sain & Purnami, 2015). *Sampling-based approaches* merupakan pendekatan sampling yang memodifikasi distribusi data training

sehingga kedua kelas data (negatif maupun positif) dipresentasikan dengan baik dengan data training (Choi, 2010). Pendekatan *sampling* dibedakan menjadi dua yaitu *undersampling* dan *oversampling.* Metode *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan jumlah data kelas minor. Masalah yang muncul dari metode *oversampling* adalah masalah *overfitting* yang menyebabkan aturan klasifikasi menjadi semakin spesifik meskipun akurasi untuk data training semakin baik. Sedangkan metode *undersampling* dilakukan dengan mengurangi jumlah data kelas mayor agar data menjadi seimbang. Kekurangan metode *undersampling* adalah semakin berkurangnya informasi dari data karena banyak data yang dihilangkan, yang banyak informasinya sehingga efektivitas klasifikasi menurun, sedangkan penghapusan data yang tidak relevan, berlebihan, ataupun *noise* dapat mengakibatkan efektivitas klasifikasi meningkat (Chawla, Bowyer, Hall, & Kegelmeyer, 2002 ; Sain & Purnami, 2015).

Salah satu metode *oversampling* adalah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) yang diperkenalkan oleh Chawla, Bowyer, Hall, dan Kegelmeyer (2002) pada klasifikasi *imbalanced data* dengan *decision tree*. SMOTE digunakan untuk menambah jumlah data kelas minoritas dengan cara mereplikasi data secara acak agar seimbang dengan jumlah kelas mayoritas (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Khaulasari, Purnami, & Prastyo, 2016). Sedangkan, salah satu metode *undersampling* adalah *Tomek Links* yang diperkenalkan oleh Tomek tahun 1997. Metode ini bekerja untuk menghapus data kelas negatif (mayoritas) yang merupakan kasus borderline atau yang memiliki kesamaan karakteristik. Selanjutnya, beberapa kasus dalam literatur telah menunjukkan bahwa kombinasi teknik *undersampling* dan *oversampling* umumnya memberikan hasil yang lebih baik daripada teknik tunggal (Batista, Prati, & Monard, 2004). Gaudio *et al.* (2013) memperoleh kesimpulan bahwa metode gabungan SMOTE+*Tomek Links* efektif untuk menangani *higher imbalanced* data. Menurut Sain dan Purnami (2015),

dengan metode SVM 5-*fold cross validation* yang diterapkan pada data medis, penggunaan metode *Combine sampling* (SMOTE+*Tomek Links*) secara umum lebih baik dari metode SMOTE dan *Tomek Links*. Dimana akurasi tertinggi pada data Ecoli 1 dengan jumlah data sebanyak 200 dan rasio data adalah 1 : 18 yaitu sebesar 96,90%. Khaulasari, Purnami, dan Prastyo (2016) menerapkan Sampling (SMOTE+*Tomek Links*) LS-SVM untuk klasifikasi *multi class imbalanced* dengan menggunakan data medis yaitu adalah data *thyroid*, kanker payudara dan kanker serviks. Hasilnya menunjukkan bahwa metode terbaik yang digunakan dalam memprediksi status pasien penderita *thyroid*, kanker payudara dan kanker serviks adalah metode *Combine sampling Least Square Support Vector Machine PSO-GSA*. Berdasarkan kelebihan dan kekurangan dari metode *oversampling* dan *undersampling*, pada penelitian ini akan digunakan *Combine sampling* yaitu penggabungan metode *oversampling* (SMOTE) dan metode *undersampling* (*Tomek Links*) untuk mengatasi data *imbalanced.*

Sebagai metode pembanding hasil dari klasifikasi Regresi Logistik, Regresi Logistik Ridge dengan *Combine sampling* akan dibandingkan efektivitas *Combine sampling* pada Analisis Diskriminan Kernel. Analisis Diskriminan Kernel adalah pengembangan dari analisis diskriminan sebagai metode statistik fundamental untuk biner *classifier*. Selain itu, Analisis Diskriminan Kernel adalah suatu metode pendekatan nonparametrik yang bersifat fleksibel karena tidak harus memenuhi asumsi tertentu pada analisisi diskriminan parametrik, yaitu asumsi normal multivariat dan matriks ragam peragam yang homogen (Hardle, 1990). Menurut Li, Gong, dan Liddell (2001), menyebutkan bahwa metode Analisis Diskriminan Kernel lebih baik dibandingkan dengan metode *Principal Component Analysis* (PCA), Kernel PCA, dan Analisis Diskriminan Linier. Penelitian sebelumnya dilakukan oleh Wahyuningtias dan Otok (2012) mengenai evaluasi ketepatan klasifikasi kelulusan tes keterampilan seleksi nasional masuk Perguruan Tinggi bidang olahraga dengan

analisis diskriminan kernel, dimana diperoleh ketepatan klasifikasi sebesar 96,82%. Selain itu Azkiya, Mukid, dan Ispriyanti (2015) melakukan penelitian tentang klasifikasi nasabah kredit Bank "X" di Provinsi Lampung menggunakan analisis diskriminan kernel. Hasil yang diperoleh menunjukkan bahwa dengan menggunakan analisis diskriminan kernel dengan fungsi kernel normal memiliki ketepatan klasifikasi lebih baik dibandingkan analisis diskriminan kernel dengan fungsi kernel epanechnikov yaitu sebesar 92%.

1.2 Rumusan Masalah

Salah satu masalah dalam klasifikasi data adalah komposisi data yang tidak seimbang (*imbalanced data*). Hal tersebut menyebabkan hasil akurasi prediksi yang baik pada kelas data training mayoritas, sedangkan akan dihasilkan akurasi prediksi yang buruk pada data training minoritas karena *classifier* cenderung memprediksi kelas data mayoritas. Berdasarkan uraian tersebut, permasalahan utama dalam penelitian ini adalah bagaimana kinerja metode *combine sampling* (SMOTE+*Tomek Links*) pada Regresi Logistik dan Regresi Logistik Ridge dalam meningkatkan akurasi klasifikasi data desa tertinggal di Jawa Timur. Hasil klasifikasi dari metode tersebut akan dibandingkan efektivitas *combine sampling* pada Analisis Diskriminan Kernel.

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan yang ingin dicapai pada penelitian ini ada-lah sebagai berikut.

1. Mengetahui karakteristik desa tertingal di Jawa Timur berdasarkan variabel yang diduga mempengaruhi status ketertinggalan desa.
2. Mengetahui ketepatan klasifikasi data *imbalanced* pada status ketertinggalan desa di Jawa Timur menggunakan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.
3. Membandingkan tingkat ketepatan klasifikasi untuk status ketertinggalan desa di Jawa Timur melalui

pendekatan *combine sampling* pada Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.

4. Mengetahui efektivitas hasil penanganan *combine sampling* pada Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel untuk klasifikasi ketertinggalan desa di Jawa Timur.

1.4 Manfaat

Hasil dari penelitian ini diharapkan mampu memberikan hasil klasifikasi yang tepat mengenai status ketertinggalan desa di Jawa Timur berdasarkan kabupaten yang memiliki jumlah desa tertinggal tertinggi, sehingga dapat membantu memberikan masukan bagi pemerintah dalam memprediksi status ketertinggalan desa khususnya pada desa-desa yang telah mengalami pemekaran agar kebijakan-kebijakan yang ditetapkan tepat sasaran. Selain itu bagi pembaca, penelitian ini dapat menambah wawasan dalam penerapan metode c*ombine sampling*, Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.

1.5 Batasan Masalah

Batasan masalah yang digunakan dalam penelitian ini adalah metode c*ombine sampling* yang digunakan adalah SMOTE dan *Tomek Links*. Data yang digunakan adalah data PODES tahun 2014 dengan skala data numerik dan digunakan desa dari 5 kabupaten yang memiliki persentase desa tertinggal tertinggi yaitu Bangkalan, Situbondo, Sumenep, Bondowoso, dan Sampang. Metode klasifikasi yang digunakan adalah Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel. Selain itu, fungsi kernel yang digunakan adalah kernel RBF.

(*Halaman ini sengaja dikosongkan*)

BAB II

TINJAUAN PUSTAKA

Pada bab ini akan dibahas metode-metode yang digunakan untuk menganalisis data yaitu *Combine sampling* yang meliputi SMOTE dan *Tomek Links*, Regresi Logistik, Regresi Ridge, Analisis Diskriminan, Analisis Diskriminan Kernel, dan desa tertinggal.

2.1 *Imbalanced* Data

Imbalanced data adalah 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 Sampling* yaitu kombinasi antara *oversampling* SMOTE dan *undersampling Tomek Links* sebagai berikut.

2.1.1 *Synthetic Minority Oversampling Technique (*SMOTE)

Salah satu metode *oversampling* adalah *Synthetic Minority Oversampling Technique* (SMOTE). Metode ini diperkenalkan oleh Chawla et al (2002) untuk mengatasi *imbalanced data* pada suatu *dataset*. SMOTE digunakan untuk menambah jumlah data

kelas minoritas dengan cara mereplikasi data secara acak agar seimbang dengan jumlah kelas mayoritas. Algoritma SMOTE sendiri yaitu mencari *k nearest neighbor* (ketetanggaan data) untuk setiap data di kelas minoritas, setelah itu dibuat *synthetic* data atau replikasi data sebanyak persentase duplikasi yang diinginkan antara data kelas minoritas dan *k nearest neighbour* yang dipilih secara random. Kemudian akan terbentuk sampel buatan baru (x_{syn}) dengan rumus sebagai berikut.

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \tag{2.1}$$

dimana,

x_{syn} : data hasil replikasi

x_i : data yang akan direplikasi

x_{knn} : data yang memiliki jarak terdekat dari data yang akan direplikasi, nilai ini ditentukan dengan jarak Euclidean yaitu $d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

δ : nilai antara 0 dan 1

Berikut adalah contoh penggunaan SMOTE pada data.

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

Data yang akan direplikasi yaitu data dari kelas minoritas (*Y=1*). Jumlah data minoritas (*Y=1*) sebanyak 3 sedangkan jumlah data mayoritas (*Y=0*) sebanyak 6 sehingga nilai persentase SMOTE yang akan digunakan adalah (6/3) x 100% = 200%. Hal ini menyimpulkan bahwa setiap data minoritas akan direplikasi 1

kali dan tetangga data dari data yang akan di replikasi akan dipilih hanya salah satu yang merupakan tetangga data yang terdekat (x_{knn}), sehingga dengan menggunakan rumus (2.1) maka data setiap data kelas minoritas akan menghasilkan satu data replikasi dengan δ= 0,3 dan hasil sebagai berikut.

Menentukan tetangga terdekat (x_{knn}) diawali dengan perhitungan antara observasi 1 dengan observasi 2 dibandingkan observasi 1 dengan observasi 3.

Observasi 1 dengan observasi 2

$$d\left(\begin{bmatrix}2\\13\end{bmatrix},\begin{bmatrix}3\\15\end{bmatrix}\right)=\sqrt{(2-3)^2+(13-15)^2}=2,236$$

Observasi 1 dengan observasi 3

$$d\left(\begin{bmatrix}2\\13\end{bmatrix},\begin{bmatrix}3\\10\end{bmatrix}\right)=\sqrt{(2-3)^2+(13-10)^2}=3,162$$

Dapat diketahui dari jarak Euclidean bahwa jarak terdekat terletak pada observasi 1 dan 2, sehingga digunakan x_{knn} adalah observasi 2. Berikut adalah perhitungan data sintesis berdasarkan persamaan (2.1).

$y_1=(1),\ x_1=2 \text{ dan } x_2=13$

$y_{knn}=(1),\ x_{knn(1)}=3 \text{ dan } x_{knn(2)}=15$

$$x_{syn(1)}=x_1+\left(x_{knn(1)}-x_1\right)\times\delta$$
$$=x_1\left(1-\delta\right)+\delta x_{knn(1)}=2\left(1-0,3\right)+0,3(3)=2,3$$
$$x_{syn(2)}=x_2+\left(x_{knn(2)}-x_2\right)\times\delta$$
$$=x_2\left(1-\delta\right)+\delta x_{knn(2)}=13\left(1-0,3\right)+0,3(15)=13,6$$

Sehingga diperoleh nilai $y_{syn}=(1),\ x_{syn(1)}=1,3 \text{ dan } x_{syn(2)}=1,6$, begitu seterusnya untuk observasi data minoritas yang lain sampai memenuhi persentase oversampling.

2.1.2 Tomek Links

Tomek Links dapat didefinisikan sebagai berikut, diberikan dua sampel dan milik kelas yang berbeda, dan $d(x,y)$ adalah jarak antara 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 Links*, maka salah satu dari kedua sampel adalah data noise atau kedua contoh adalah *borderline*. *Tomek Links* dapat digunakan sebagai metode *undersampling* yaitu hanya sampel dari kelas negatif yang akan dieliminasi atau sebagai metode pembersihan data yaitu kedua sampel dari kedua kelas yang berbeda akan dihapus.

Berikut adalah contoh penggunaan *Tomek Links* pada data (Sain, 2013).

No.	X_1	X_2	Y
1.	2	2	0
2.	3	6	0
3.	4	2	0
4.	6	5	0
5.	1	2	1
6.	1	4	1
7.	3	1	1
8.	3	3	1
9.	3	4	1

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

Nilai $Y=0$ merupakan sampel dari kelas negatif dan $Y=1$ adalah sampel dari kelas positif, sehingga contoh hasil perhitungannya adalah sebagai berikut.

Dengan menggunakan rumus jarak Euclidean yaitu :

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

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 dideteksi berada dekat dengan antara titik y_5 dan y_1 adalah $y_{18} = (2,(1,5))$ sehingga

$$d(y_5, y_{18}) = \sqrt{(1-2)^2 + (2-1{,}5)^2} = 1{,}12$$

$$d(y_1, y_{18}) = \sqrt{(2-2)^2 + (2-1{,}5)^2} = 0{,}5$$

Diperoleh kesimpulan bahwa $d(y_5, y_{18}) = 1{,}12 > d(y_5, y_1) = 1$ atau $d(y_1, y_{18}) = 0{,}5 < d(y_5, y_1) = 1$, sehingga titik y_5 dan y_1 bukan merupakan kasus *Tomek Links* karena telah memenuhi syarat definisi kasus *Tomek Links*.

2. Misalkan y_8 = (3,3) dan y_{17} = ((2,5),(2,5)) dideteksi merupakan kasus *Tomek Links*.

   $$d(y_8, y_{17}) = \sqrt{(3-2{,}5)^2 + (3-2{,}5)^2} = 0{,}707$$

   Titik lain yang dideteksi berada dekat dengan antara titik y_8 dan y_{17} adalah y_9 = (3,4) sehingga

   $$d(y_8, y_9) = \sqrt{(3-3)^2 + (3-4)^2} = 1$$

   $$d(y_{17}, y_9) = \sqrt{(2{,}5-3)^2 + (2{,}5-4)^2} = 1{,}58$$

   Diperoleh kesimpulan bahwa $d(y_8, y_9) = 1 > d(y_8, y_{17}) = 0{,}707$ atau $d(y_{17}, y_9) = 1{,}58 > d(y_8, y_{17}) = 0{,}707$, sehingga 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 dan begitu seterusnya untuk observasi data mayoritas yang lain sampai data tersebut bersih dari *noise* dan *borderline*.

2.2 Multikolinearitas

Multikolinearitas adalah kondisi dimana terjadi korelasi tinggi antar variabel prediktor. Dalam mendeteksi multikolinearitas pada regresi yang mempunyai lebih dari dua variabel prediktor dapat menggunakan (VIF). Nilai VIF untuk parameter regresi ke-*j* memiliki persamaan sebagai berikut (Hocking, 2003).

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2.2}$$

R_j^2 merupakan koefisien determinasi antara X_j dengan variabel prediktor lainnya pada persamaan regresi dimana $j = 1, 2, ..., p$. Apabila nilai *VIF* lebih dari 5, maka dapat diindikasikan terdapat kasus multikolinieritas. Sedangkan jika nilai *VIF* lebih besar dari 10, maka hal tersebut mengindikasikan bahwa variabel prediktor memiliki kasus multikolinieritas yang serius dan harus diatasi. Hal-hal yang akan terjadi apabila kasus multikolinearitas tidak diatasi adalah variansi estimasi menjadi besar, interval kepercayaan menjadi lebar dikarenakan variansi dan standar error besar. Kemudian pengujian signifikansi secara parsial menjadi tidak signifikan. Serta koefisien determinasi (R^2) tinggi, tetapi hanya sedikit variabel prediktor yang signifikan.

2.3 Regresi Logistik

Regresi logistik adalah salah satu metode statistik yang digunakan untuk memodelkan variabel respon yang bersifat kategorik dengan satu atau lebih variabel prediktor bersifat kategorik atau kontinu (Hosmer, Lemeshow, & Sturdivant, 2013). Regresi logistik berdasarkan skala dibagi menjadi tiga, yaitu regresi logistik biner, multinomial, dan ordinal.

Misalkan $\mathbf{x}_i \in R^{p+1}$ adalah vektor untuk setiap observasi di **X** dengan i=1, …, n. **β** adalah vektor parameter dan **y** adalah vektor respon biner. Variabel respon (Y) bersifat dikotomus atau hanya memiliki dua kategori yaitu 1 menyatakan jika sukses (kelas positif/minoritas) dan 0 jika gagal (kelas negative/mayoritas). Pada dasarnya, regresi logistik dibangun untuk variabel prediktor kontinyu ($x \in R$)

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Regeresi logistik biner memiliki variabel respon mengikuti distribusi Bernoulli (Binomial) dengan peluang sukses sebesar π. Untuk setiap observasi ke *i* dapat ditulis

$$y_i \sim Binomial\ (1, \pi_i)$$

Fungsi probabilitas untuk setiap observasi adalah sebagai berikut: (Hosmer, Lemeshow, & Sturdivant, 2013)

$$f\left(y_i\right) = \left(\pi\left(\mathbf{x}_i\right)\right)^{y_i} \left(1-\pi\left(\mathbf{x}_i\right)\right)^{1-y_i}, y = \{0,1\} \tag{2.3}$$

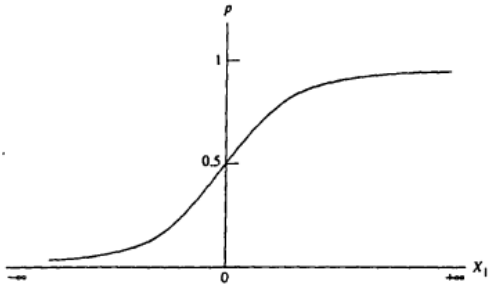
π_i adalah probabilitas (peluang) dari kejadian ke-i.
Jika y_i = 0, maka $f(y_i) = (\pi_i)^0(1-\pi_i)^{1-0} = (1-\pi_i)$
Jika y_i = 1, maka $f(y_i) = (\pi_i)^1(1-\pi_i)^{1-1} = \pi_i$
Menurut Hosmer dan Lemeshow (2000), fungsi logistik yang digunakan untuk memodelkan $\mathbf{x}_i$ dengan nilai ekspektasi y_i nya yaitu

$$\pi\left(\mathbf{x}_i\right) = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)}{1+\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)} = \frac{e^{\mathbf{x}_i{}^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i{}^T \boldsymbol{\beta}}} \tag{2.4}$$

atau

$$\pi\left(\mathbf{x}_i\right) = \frac{1}{1+e^{-\left(\mathbf{x}_i{}^T \boldsymbol{\beta}\right)}} \tag{2.5}$$

Dalam regresi logistik, hubungan antara variabel prediktor dan variabel respon bukanlah suatu fungsi linier (Gambar 2.5).



(Sumber: Buku Sharma,1996)
Gambar 2.1 Kurva Regresi Logistik

Apabila variabel prediktor ada sebanyak p variabel, maka model regresi Logistik dapat dituliskan dalam bentuk logit, yaitu fungsi link dari regresi Logistik.

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

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

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

$$\text{Logit}\left[\pi(\mathbf{x}_i)\right] = \left(\mathbf{x}_i^T\boldsymbol{\beta}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \qquad (2.6)$$

Pengklasifikasian observasi dilakukan dengan terlebih dahulu mengestimasi nilai probabilitas pada persamaan (2.5). Setelah didapat nilai probabilitas, klasifikasi observasi kedalam kelompok berdasarkan nilai probabilitas dengan nilai *cutoff* yang biasanya diasumsikan sebesar 0,5 (Sharma, 1996). Observasi dengan nilai probabilitas lebih besar sama dengan 0,5 diklasifikasikan kedalam kelas sukses atau kelas positif (1), sedangkan jika nilai probabilitas kurang dari *cutoff* diklasifikasikan kedalam kelas gagal atau kelas negatif (0).

a. Estimasi Parameter

Dalam mengestimasi parameter dalam model regresi logistik digunakan metode *Maximum Likelihood Estimator* (MLE). Metode MLE digunakan karena distribusi dari variabel respon telah diketahui. Metode ini mengestimasi parameter β dengan cara memaksimumkan fungsi *likelihood*. Dari Persamaan (2.3) didapatkan fungsi *likelihood*: (Hosmer, Lemeshow, & Sturdivant, 2013)

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

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

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

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

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

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

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

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

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

Melalui Persamaan 2.7 dilakukan penurunan terhadap β menjadi Persamaan 2.8.

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

$$=\sum_{i=1}^{n}\left(y_i\mathbf{x}_i{}^T-\pi_i\mathbf{x}_i{}^T\right)=\mathbf{X}^T(\mathbf{y}-\hat{\boldsymbol{\pi}}) \tag{2.8}$$

Untuk $\frac{\partial \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=0,$ maka $\mathbf{X}^T(\mathbf{y}-\hat{\boldsymbol{\pi}})=0$. Bila $\hat{\mathbf{y}}=\hat{\boldsymbol{\pi}},$ maka didapatkan persamaan

$$\mathbf{X}^T(\mathbf{y}-\hat{\mathbf{y}})=0 \tag{2.9}$$

Persamaan (2.9) didapatkan menggunakan metode Newton-Raphson. Turunan kedua adalah sebagai berikut:

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

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

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

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

$$= -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\pi_i(1-\pi_i) = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

$$\frac{\partial^2 \ln L(\mathbf{X},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \mathbf{H}(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T Var(\pi_i) \qquad (2.10)$$

dimana ***H***(**β**) adalah matriks Hessian. Karena turunan kedua selalu bernilai negative, maka didapat bahwa nilai **β** membuat fungsi *likelihood* bernilai ekstrim maksimum. Namun karena hasil turunan pertama pada persamaan (2.8) tidak mendapatkan hasil yang eksplisit atau rumus untuk mencari nilai **β** tidak didapat, maka akan digunakan Deret Taylor. Apabila dilakukan ekspansi berdasarkan Deret Taylor disekitar nilai β, maka didapatkan persamaan (2.10).

$$\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}} = \frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}}\bigg|_{\hat{\beta}=\hat{\beta}_0} + \frac{\partial^2 L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}^T}\bigg|_{\hat{\beta}=\hat{\beta}_0}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right) \qquad (2.11)$$

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

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

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

Hasil substitusi persamaan (2.6) dan (2.9) ke dalam persamaan (2.11) menghasilkan estimasi parameter $\widehat{\boldsymbol{\beta}}$ ditunjukkan pada persamaan (2.12) (Hosmer & Lemeshow, 2000).

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

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \qquad (2.12)$$

dengan $\mathbf{z}$ merupakan vektor $n \times 1$ dan $\mathbf{W}$ merupakan pembobot dengan fungsi seperti dibawah ini :

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

$$\mathbf{z}_i = \text{Logit}\left[\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\right] + \frac{y_i - \hat{\boldsymbol{\pi}}(\mathbf{x}_i)}{\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\left[1-\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\right]} \qquad (2.13)$$

Matriks kovarian untuk $\hat{\boldsymbol{\beta}}$ ditampilkan pada persamaan sebagai berikut :

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^T diag\left[\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right]\mathbf{X}\right)^{-1} \qquad (2.14)$$

b. Pengujian Signifikansi Parameter

Pengujian signifikansi 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:

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ (model tidak berpengaruh signifikan)

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

Statistik uji yang digunakan adalah :

$$G = -2\ln\left(\frac{\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}}{\prod_{i=1}^{n}\hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{1-y_i}}\right) \tag{2.15}$$

dengan n_0 adalah banyaknya pengamatan yang bernilai Y=0 dan n_1 adalah banyaknya pengamatan bernilai Y=0. Pengambilan keputusan, H_0 akan ditolak apabila $G \geq \chi^2_{(p,\alpha)}$ atau $p - value < \alpha$. Jika pada pengujian serentak menghasilkan kesimpulan tolak H_0, maka pengujian akan dilanjutkan dengan uji 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

$H_0 : \beta_j = 0$ (variabel ke-j tidak berpengaruh signifikan)

$H_1 : \beta_j \neq 0,\ j = 1,2,..., p$ (variabel ke-j berpengaruh signifikan)

Statistik uji :

$$Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{2.16}$$

dengan $SE(\hat{\beta}_j) = \sqrt{\text{var}(\hat{\beta}_j)}$. Daerah kritis: H_0 ditolak, jika nilai $|Z| > Z_{\alpha/2}$ atau *p-value* < α. Artinya, variabel ke-*j* berpengaruh signifikan terhadap pembentukan model.

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, 1998). Model regresi linier dinyatakan dengan persamaan:

$$\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.17}$$

didapatkan error, $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}^T\boldsymbol{\beta}$. Melalui metode *least square* dengan meminimalkan jumlah kuadrat error,

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

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}^T\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}^T\mathbf{Y} + \mathbf{X}\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{\beta} \tag{2.18}$$

yaitu dengan mengusahakan turunan pertama persamaan (2.18) terhadap vektor $\boldsymbol{\beta}$ sama dengan nol (Draper & Smith, 1998).

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.19}$$

Pada Regresi Ridge, estimasi parameter ditambahkan dengan ridge parameter pada elemen diagonal matriks, dimana ridge parameter merupakan bilangan positif kecil, sehingga bias yang terjadi dapat dikendalikan. Nilai koefisien untuk parameter Regresi Ridge dalam bentuk matriks dituliskan pada persamaan (2.20) (Drapper & Smith, 1998).

$$\hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^T\mathbf{X} + \theta\boldsymbol{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.20}$$

yang mana didapat dengan meminimalkan fungsi obyektif

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

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 *least square* pada Regresi Linier. Jika nilai θ lebih dari 0, maka estimasi ridge akan bias terhadap parameter β, tetapi cenderung lebih stabil (Sunyoto, Setiawan, & Zain, 2009).

2.5 Regresi Logistik Ridge

Fungsi obyektif untuk Regresi Ridge yang didapat dari model linier $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ pada persamaan (2.21) adalah:

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

Sedangkan fungsi obyektif Regresi Logistik pada persamaan (2.6) dituliskan:

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

Kemudian, Vago & Kemeny (2006) dengan menerapkan teknik pada Regresi Ridge pada Regresi Logistik, didapatkan fungsi obyektif untuk Regresi Logistik Ridge pada persamaan (2.22).

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

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

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

$$\phi(\hat{\boldsymbol{\beta}}^{\oplus}) = \sum_{i=1}^{n} y_i \ln\left[\pi_i(\mathbf{x}_i)\right] + \sum_{i=1}^{n}(1-y_i)\ln\left[1-\pi_i(\mathbf{x}_i)\right] - \theta\hat{\boldsymbol{\beta}}^{\oplus T}\hat{\boldsymbol{\beta}}^{\oplus} \qquad (2.22)$$

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

Selanjutnya diturunkan secara parsial terhadap $\hat{\beta}^{\oplus}$.

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

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

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

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

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

Kemudian dilakukan penurunan kedua.

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

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

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

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

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

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

$$\frac{\partial^2 \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus} \partial \hat{\boldsymbol{\beta}}^{\oplus T}} = -\mathbf{X}^T \mathbf{W} \mathbf{X} - 2\theta \mathbf{I} \qquad (2.23)$$

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

Estimasi parameter Regresi Logistik Ridge menggunakan metode MLE dengan iterasi *Newton-Raphson* yang akan

digunakan untuk memaksimumkan fungsi obyektif pada persamaan (2.22). Kemudian diekspansikan di sekitar $\beta^{\oplus}$menurut Deret *Taylor* dan didapatkan persamaan (2.24).

$$\left.\frac{\partial \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0^{\oplus}} = -\left.\frac{\partial^2 \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus}\partial \hat{\boldsymbol{\beta}}^{\oplus T}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0^{\oplus}} (\hat{\boldsymbol{\beta}}^{\oplus} - \hat{\boldsymbol{\beta}}_0{}^{\oplus}) \qquad (2.24)$$

Hasil penurunan di substitusikan ke dalam persamaan (2.24) menghasilkan estimasi parameter Regresi Logistik Ridge pada persamaan (2.25) (Vago & Kemeny, 2006).

$$\hat{\boldsymbol{\beta}}^{\oplus} = (\mathbf{X}^T\mathbf{W}\mathbf{X} + 2\theta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \qquad (2.25)$$

dengan $\beta^{\oplus}$ adalah parameter ridge untuk Regresi Logistik Ridge yang merupakan bilangan positif kecil. z merupakan vektor berukuran $n \times 1$, dengan $z_i = \text{Logit}[\hat{p}_i(\mathrm{x}_i)] + \dfrac{y_i - \hat{p}_i(\mathrm{x}_i)}{\hat{\pi}_i(\mathrm{x}_i)[1-\hat{p}_i(\mathrm{x}_i)]}$.

Dengan menambahkan *ridge parameter* untuk Regresi Logistik Ridge pada elemen diagonal matriks kovarian dari Regresi Logistik, maka variansi Regresi Logistik Ridge dapat dihitung dengan formula pada persamaan (2.26).

$$Var(\hat{\boldsymbol{\beta}}^{\oplus}) = \left(\mathbf{X}^T \text{ diag}[\hat{\pi}_i(1-\hat{\pi}_i)]\mathbf{X} + \theta^{\oplus}\mathbf{I}\right)^{-1}. \qquad (2.26)$$

2.6 Analisis Diskriminan

Analisis diskriminan merupakan metode statistik multivariat untuk mengelompokkan atau mengklasifikasikan sejumlah objek ke dalam beberapa kelompok, berdasarkan beberapa variabel sedemikian hingga setiap obyek menjadi anggota dari salah satu kelompok, tidak ada obyek yang menjadi anggota lebih dari pada 1 kelompok. Analisis diskriminan akan menghasilkan variabel independen yang benar-benar membedakan antar kelompok (Johnson & Winchern, Applied Multivariate Statistical Analysis, 1992). Klasifikasi pada analisis diskriminan bersifat *mutually exclusive*, yaitu jika suatu pengamatan telah masuk pada salah satu kelompok maka tidak dapat menjadi anggota dari kelompok yang lain (Hair, Anderson, Babin, & Black, 2006). Analisis Diskriminan Fisher dibangun dari pendekatan ECM (*Expected Cost of*

Missclassification) sehingga diperlukan adanya asumsi distribusi normal multivariat. ECM terbangun dari fungsi distribusi normal *p-variat*, sehingga asumsi yang harus dipenuhi dalam Analisis Diskriminan adalah asumsi homogenitas dan asumsi distribusi normal multivariat.

2.6.1 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.

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

dan (2.27)

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

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

$$z_1 = \frac{(\mathrm{p}+1)(\mathrm{n}+1)(\mathrm{n}+3)}{6[(\mathrm{n}+1)(\mathrm{p}+1)-6]}b_{1,\mathrm{p}} \qquad (2.28)$$

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.

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \tag{2.29}$$

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.

2.6.2 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 & Wichern, 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 (2.30):

$$\chi^2 = -2(1-\mathrm{c}_i)\left[\frac{1}{2}\sum_{i=1}^{2} v_i \ln|\mathbf{S}_i| - \frac{1}{2}\ln|\mathbf{S}_{pool}|\sum_{i=1}^{2} v_i\right] \tag{2.30}$$

dengan

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

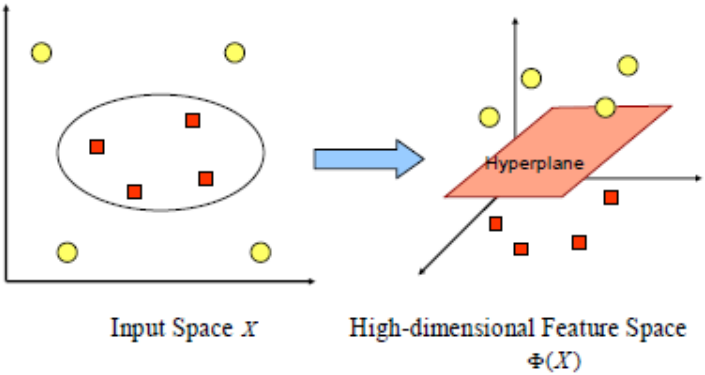
$$c_1 = \left[\sum\nolimits_{i=1}^{2} \frac{1}{v_i} - \frac{1}{\sum_{i=1}^{2} v_i}\right]\left[\frac{2p^2+3p-1}{6(\mathrm{p}+1)}\right]$$

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

2.7 Analisis Diskriminan Kernel

Analisis Diskriminan Kernel adalah pendekatan analisis diskriminan nonlinier berdasarkan pada teknik kernel yang

dikembangkan untuk model yang memiliki pola nonlonier pada bentuk maupun teksturnya (Li, Gong, & Liddell, 2001). Analisis diskriminan merupakan salah satu teknik dalam analisis multivariat dengan metode dependensi (dimana hubungan antar variabel sudah bisa dibedakan mana variabel terikat dan mana variabel bebas) (Hair, Black, Babin, & Anderson, 2006). Dalam penggunaannya analisis diskriminan kernel tidak terikat asumsi apapun. Dalam metode kernel, suatu data *x* di *input space* dipetakan ke kernel *space F* dengan dimensi yang lebih tinggi seperti Gambar (2.2).

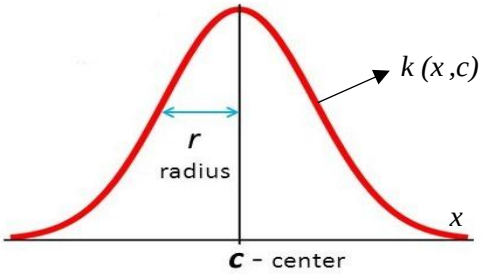


Gambar 2.2 Pemetaan Data ke Ruang Vektor yang Lebih Tinggi

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. Pada penelitian ini akan menggunakan fungsi kernel dengan pendekatan kernel Gaussian RBF dengan persamaan (2.31).

$$k(\mathbf{x},\mathrm{c}) = \exp\left(-\frac{(x-c)^2}{r^2}\right) \tag{2.31}$$

Kernel Gaussian RBF dapat divisualisasikan pada Gambar 2.3 berikut.



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 (2.32) (Mika, Ratsch, Jason, Scholkopf, & Muller, 1999):

$$J(\boldsymbol{\omega}) = \frac{\omega^T S_B^{\Phi} \omega}{\omega^T S_W^{\Phi} \omega} \tag{2.32}$$

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 (2.32) dan fungsi perluasan dari $\boldsymbol{\omega}$ pada persamaan (2.33).

$$\boldsymbol{\omega} = \sum_{i=1}^{l} \alpha_i \Phi(x_i) \tag{2.33}$$

Persamaan 2.33 dan persamaan m_i^{Φ} menghasilkan persamaan (2.34).

$$\boldsymbol{\omega}^T m_i^{\Phi} = \frac{1}{l_i} \sum_{j=1}^{l} \sum_{k=1}^{l_i} \alpha_j \, \mathrm{k}(x_j, x_k^i) = \alpha^T M_i \tag{2.34}$$

dengan $(M_i)_j = \frac{1}{l_i}\sum_{j=1}^{l}\sum_{k=1}^{l_i} \mathrm{k}(x_j, x_k^i)$. Melalui persamaan (2.34) diperoleh persamaan baru dari $\boldsymbol{\omega}^T S_B^{\Phi}\boldsymbol{\omega}$ sebagai berikut:

$$\boldsymbol{\omega}^T S_B^{\Phi}\boldsymbol{\omega} = \alpha^T M\alpha \tag{2.35}$$

Dengan $M = (M_1 - M_2)(M_1 - M_2)^T$. Persamaan $\boldsymbol{\omega}^T S_W^{\Phi}\boldsymbol{\omega}$ juga berubah menjadi sebagai berikut:

$$\boldsymbol{\omega}^T S_W^{\Phi}\boldsymbol{\omega} = \alpha^T N\alpha \tag{2.36}$$

dimana $N = \sum_{j=1,2} K_j(\boldsymbol{I} - 1_{lj})K_j^T$. 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 (2.37).

$$J(\alpha) = \frac{\alpha^T M\alpha}{\alpha^T N\alpha} \tag{2.37}$$

Pola baru dari x akan diproyeksikan kedalam dengan fungsi sebagai berikut:

$$(\boldsymbol{\omega}, \Phi(x)) = \sum_{i=1}^{l} \alpha k(\boldsymbol{x}_i, \boldsymbol{x}) \tag{2.38}$$

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 :

$$\hat{f}_t(\mathbf{x}) = \frac{1}{n_t}\sum_{i=1}^{n_t} K_t(\mathbf{x} - \mathbf{x}_i) \tag{2.39}$$

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).

2.8 Evaluasi Performansi Ketepatan Klasifikasi

Evaluasi performansi suatu sistem klasifikasi merupakan hal yang penting. Performansi sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. Semakin tinggi akurasi klasifikasi berarti performansi teknik klasifikasi juga semakin tinggi. Ketepatan klasifikasi untuk kelas dikotomus dapat dihitung dengan menggunakan *confusion matrix* (tabel klasifikasi). Tabel klasifikasi dapat dilihat pada Tabel 2 berikut.

Tabel 2. 1 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 2.1, 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).

$$\text{rata - rata akurasi total} = \frac{\sum_{i=1}^{k} \frac{TP_i + TN_i}{TP_i + TN_i + FN_i + FP_i}}{k}; i = 1, 2, \ldots, k \quad (2.40)$$

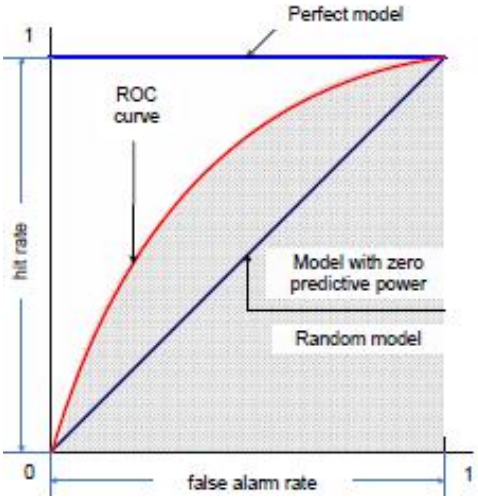
$$\text{rata - rata sensitivitas} = \frac{\sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i}}{k}; i = 1, 2, \ldots, k \quad (2.41)$$

$$\text{rata} - \text{rata spesifisitas} = \frac{\sum_{i=1}^{k} \frac{TN_i}{TN_i + FP_i}}{k}; i = 1, 2, \ldots, k \quad (2.42)$$

Perhitungan ketepatan klasifikasi bisa menggunakan *Geometric mean* (*G-mean*). Nilai ini akan memaksimalkan keakuratan masing-masing kelas dengan keseimbangan yang baik (Barandela, Sanchez, Garcia, & Rangel, 2003).

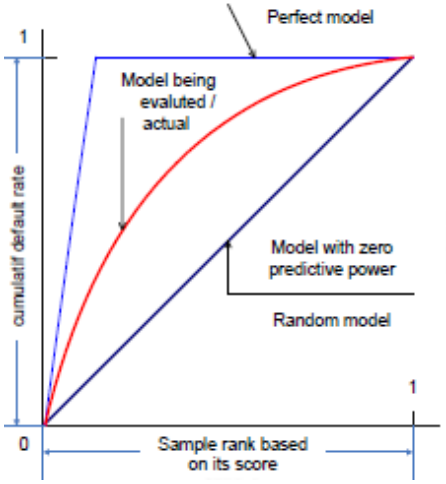
$$\text{Rata - rata } G\text{ - mean} = \sqrt{\frac{\sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i}}{k} \times \frac{\sum_{i=1}^{k} \frac{TN_i}{TN_i + FP_i}}{k}}; i = 1, 2, \ldots, k \quad (2.43)$$

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 karena *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. 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 (2.44).

$$AR = \frac{\int_0^1 Y_{actual}\, F\, dF - \frac{1}{2}}{\int_0^1 Y_{perfect}\, F\, dF - \frac{1}{2}}. \quad (2.44)$$

Selanjutnya, dari hubungan antara *AR* dan *AUC* didapatkan persamaan (2.45)

$$AR = 2AUC - 1 \quad (2.45)$$

sehingga didapatkan rumus rata-rata *AUC* pada persamaan (2.46).

$$\text{Rata-rata } AUC = \frac{\sum_{i=1}^{k} \frac{1}{2}\left(AR_k + 1\right)}{k}; i = 1,2,\ldots,k \quad (2.46)$$

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

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

$$\text{Rata - rata } AUC = \frac{1}{2}\left(\frac{\sum_{i=1}^{k} \frac{TP_i}{TP_i + FN_i}}{k} + \frac{\sum_{i=1}^{k} \frac{TN_i}{TN_i + FP_i}}{k} \right); i = 1,2,\ldots,k \quad (2.47)$$

Kategori berdasarkan nilai AUC dapat disajikan ke dalam Tabel 2.1 berikut :

Tabel 2. 2 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 : Bekka, Djemaa, & Alitouche (2013)

2.9 *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* , partisis 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 $D_1, D_2, ..., D_k$ akan menjadi data *training* kemudian D_2 akan menjadi data *testing*, begitu seterusnya (Han, Kamber, & Pei, 2012).

Dalam pembagian data *training* dan *testing* agar representatif dapat digunakan stratifikasi. Cara kerja dalam stratifikasi yaitu 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 set*. Adapun kelebihan dari *stratified k-fold cross validation* adalah menghindari adanya *overfitting* pada data training (Zhang, Wu, & Wang, 2011).

2.10 Desa Tertinggal

Desa tertinggal adalah desa yang belum terpenuhi SPM (Standar Pelayanan Minimal) Desa pada aspek kebutuhan sosial dasar, infrastruktur dasar, sarana dasar, pelayanan umum, dan penyelenggaraan pemerintahan. Beberapa tantangan yang dihadapi dalam pemenuhan SPM Desa antara lain kondisi dan kebutuhan antara satu desa dengan desa lainnya yang berbeda-beda sehingga standar pelayanan minimalnya tidak dapat diseragamkan baik aspek maupun volumenya, selanjutnya adalah tersedianya sumber daya yang masih terbatas baik sumber daya manusia maupun penganggarannya dan belum terbaginya kewenangan/urusan dari pemerintah pusat, pemerintah daerah, dan pemerintah desa dalam pemenuhan standar pelayanan minimal desa. Indeks Pembangunan Desa (IPD) disusun sebagai upaya untuk mengakomodasi beberapa aspek pemenuhan SPM Desa sebagaimana tertuang dalam UU Nomor 6 Tahun 2014 tentang desa tersebut walaupun tidak dapat

mencakup seluruhnya karena adanya keterbatasan data. IPD merupakan indeks komposit tertimbang dari 42 indikator yang secara substansi dan bersama-sama menggambarkan tingkat pembangunan di desa. Dalam rangka menilai tingkat kemajuan atau perkembangan desa, maka desa dibagi menjadi 3 (tiga) klasifikasi, yaitu desa mandiri, desa berkembang, dan desa tertinggal. Secara teknis, desa dikatakan tertinggal ketika memiliki IPD kurang dari sama dengan 50, untuk desa berkembang memiliki IPD lebih dari 50 namun kurang dari atau sama dengan 75, sedangkan desa mandiri memiliki IPD lebih dari 75. Pada penelitian ini dilakukan penggabungan kelompok desa berkembang dan mandiri menjadi kelas desa tidak tertinggal. Perhitungan IPD setiap desa diformulasikan sebagai berikut (BPS, 2015).

$$\text{IPD} = (\sum_{i=1}^{42} b_i * V_i) * 20 \qquad (2.48)$$

dengan *b* merupakan pembobot indikator dan *V* adalah skor indikator yang telah ditentukan berdasarkan metode *Principal Component Analysis* (PCA).

Terdapat 5 aspek atau dimensi yang digunakan dalam pemenuhan SPM Desa untuk pembangunan desa sebagaimana tertuang dalam UU Nomor 6 ahun 2014 tentang desa sebagai berikut.

Tabel 2. 3 Lima Dimensi dalam Pemenuhan SPM Desa

Dimensi	Variabel
Pelayanan Dasar	Ketersediaan dan akses terhadap fasilitas pendidikan : a. Ketersediaan TK b. Ketersediaan SD c. Ketersediaan SMP d. Ketersedian SMA Ketersediaan dan akses terhadap kesehatan : a. Ketersediaan rumah sakit b. Ketersediaan rumah sakit bersalin c. Ketersediaan puskesmas d. Ketersediaan tempat prakter doktor

Tabel 2. 3 Lima Dimensi dalam Pemenuhan SPM Desa (Lanjutan)

Dimensi	Variabel
Pelayanan Dasar	e. Ketersediaan poliklinik/balai pengobatan f. Ketersediaan tempat praktek bidan g. Ketersediaan poskesdes h. Ketersediaan polindes i. Ketersediaan apotek
Kondisi infrastruktur	Ketersediaan infrastruktur ekonomi : a. Ketersediaan minimarket maupun toko kelontong b. Ketersediaan pasar c. Ketersediaan restoran d. Ketersediaan rumah makan e. Ketersediaan akomodasi hotel/ penginapan, f. Ketersediaan bank Ketersediaan infrastruktur energi : a. Ketersediaan listrik b. Ketersediaan penerangan jalan c. Ketersediaan bahan bakar untuk memasak Ketersediaan infastruktur air bersih dan sanitasi : a. Ketersediaan sumber air minum b. Ketersediaan sumber air mandi/cuci c. Ketersediaan fasilitas buang air besar Ketersediaan dan kualitas infrastruktur komunikasi dan informasi: a. Ketersediaan komunikasi menggunakan telepon seluler b. Ketersediaan internet c. Ketersedian pengiriman pos/barang
Aksesibilitas/ transportasi	Ketersediaan dan akses terhadap sarana transportasi : a. Ketersediaan lalu lintas dan kualitas jalan b. Ketersediaan dan operasional angkutan umum c. Waktu tempuh per kilometer transportasi ke kantor camat

Tabel 2. 3 Lima Dimensi dalam Pemenuhan SPM Desa (Lanjutan)

Dimensi	Variabel
Aksesibilitas/ transportasi	d. Biaya per kilometer transportasi ke kantor camat e. Waktu tempuh per kilometer transportasi ke kantor bupati/walikota f. Biaya per kilometer transportasi ke kantor bupati/walikota
Pelayanan Umum	Penanganan kesehatan masyarakat : a. Penanganan kejadan luar biasa (KLB) b. Penaganan gizi buruk c. ketersediaan fasilitas olahraga seperti ketersediaan lapangan olahraga dan kelompok kegiatan olahraga
Penyelenggaraan Pemerintahan	a. Kelengkapan pemerintahan desa b. Otonomi desa c. Aset/kekayaan desa, serta d. kualitas sumberdaya manusia seperti kualitas SDM kepala desa daan sekretaris desa.

(*Halaman ini sengaja dikosongkan*)

BAB III

METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian adalah data sekunder yang diperoleh dari data Potensi Desa (PODES) Provinsi Jawa Timur 2014 yang dikeluarkan oleh Badan Pusat Statistik. Berdasarkan data yang telah diperoleh, digunakan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi di Jawa Timur dengan 115 desa tertinggal dan 1007 desa tidak tertinggal sehingga rasio data sebesar 1:9. Pendataan Podes telah dilaksanakan sejak tahun 1980 bersamaan dengan penyelenggaraan Sensus Penduduk 1980. Pengumpulan data Podes dilakukan sebanyak 3 kali dalam kurun waktu 10 tahun, sebagai bagian dari rangkaian kegiatan Sensus. Podes dilakukan 2 tahun sebelum pelaksanaan sensus untuk mendukung kelancaran pelaksanaan sensus.

3.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini dipilih berdasarkan referensi dari 5 aspek atau dimensi yang digunakan dalam pemenuhan SPM Desa untuk pembangunan desa sebagaimana tertuang dalam UU Nomor 6 tahun 2014 tentang desa. Dalam penelitian ini digunakan 8 variabel prediktor, sedangkan variabel respon (Y) merupakan status ketertinggalan desa yang memiliki 2 kategori yaitu 0 untuk desa tidak tertinggal dan 1 untuk desa tertinggal. Berikut adalah variabel prediktor yang digunakan dalam penelitian.

1. Pelayanan Dasar
   - X_1 : Rasio banyaknya SD/MI terhadap total jumlah siswa
   - X_2 : Rasio banyaknya tempat praktek bidan terhadap total penduduk
   - X_3 : Rasio banyaknya poskesdes terhadap total penduduk
2. Kondisi Infrastruktur
   - X_4 : Rasio banyaknya toko 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. Penyelenggaraan Pemerintahan
   X_8 : Rasio pendapatan asli desa terhadap total penduduk

Konsep dan definisi yang digunakan mengacu pada BPS yaitu:

1. 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 atau lingkungan, faktor kelembagaan, faktor sarana/ prasarana dan akses serta faktor sosial penduduk.
2. Rasio banyaknya SD/MI yaitu jumlah sekolah SD/MI dibagi total jumlah siswa dikali 100.
3. Rasio banyaknya tempat praktik bidan adalah jumlah tempat praktik bidan dibagi total penduduk dikali 100. Bidan adalah petugas paramedik yang berdomisili atau tinggal di desa atau kelurahan atau yang bertugas sebagai bidan di desa dengan SK.
4. Rasio banyaknya poskesdes adalah jumlah poskesdes dibagi total penduduk dikali 100. Poskesdes adalah upaya kesehatan bersumber daya masyarakat (UKBM) yang dibentuk di desa dalam rangka menyediakan pelayanan kesehatan dasar masyarakat desa.
5. Rasio banyaknya toko kelontong yaitu jumlah toko dibagi total penduduk dikali 100. Toko/warung kelontong adalah bangunan (kedai) yang menjual beraneka barang secara eceran.
6. Rasio banyaknya keluarga pengguna listrik adalah jumlah keluarga pengguna listrik PLN dan non PLN dibagi total penduduk dikali 100.

7. Jarak tempuh per kilometer ke kantor camat, merupakan jarak yang harus ditempuh oleh penduduk dari kantor kepala desa/lurah ke kantor camat dalam kilometer.
8. Rasio banyaknya penderita gizi buruk adalah jumlah penderita gizi buruk selama 3 tahun terakhir dibagi total penduduk dikali 100.
9. Rasio pendapatan asli desa adalah 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 dikali 100.

Struktur data pada penelitian ini ditampilkan pada Tabel 3.1 dan selengkapnya dapat dilihat pada Lampiran 1.

Tabel 3. 1 Struktur Data Penelitian

Desa	Respon	Prediktor			
	Status	X_1	X_2	...	X_8
1	0	0.830	0.000	...	1.073
2	0	1.096	0.022	...	1.172
3	0	0.574	0.025	...	1.043
⋮	⋮	⋮	⋮	⋮	⋮
1122	1	1.333	0.000	...	1.094

3.3 Langkah Analisis

Langkah analisis yang digunakan pada penelitian ini adalah sebagai berikut.

1. Mendeskripsikan karakteristik desa berdasarkan variabel yang diduga mempengaruhi status ketertinggalan desa.
   a. Melakukan *preprocessing data* dilakukan pemilihan data variabel dan *filtering* data sesuai jenis data .
   b. Melakukan analisis statistika deskriptif dari data pada variabel yang diduga mempengaruhi status desa tertinggal yang telah dilakukan *prepocessing*.
2. Melakukan klasifikasi data *imbalanced* dengan bantuan *software R* menggunakan metode :
   a. Klasifikasi dengan Regresi Logistik
      i. Pemeriksaan kasus multikolinearitas menggunakan nilai VIF pada masing-masing variabel prediktor (*X)*.

ii. Membagi data menjadi data *training* dan *testing* menggunakan 10-*fold cross validation* dengan metode stratifikasi.
iii. Melakukan klasifikasi data *training* menggunakan Regresi Logistik.
iv. Menghitung ketepatan klasifikasi pada data *testing* dengan semua variabel menggunakan persamaan (2.40) sampai (2.43) dan persamaan (2.47).
v. Melakukan uji signifikansi parameter berdasarkan pemilihan *fold* yang memiliki ketepatan klasifikasi terbaik.
vi. Melakukan pemilihan variabel signifikan berdasarkan metode *backward elimination*.
vii. Mengulangi langkah i, ii, iii, dan iv menggunakan variabel signifikan yang diperoleh dari langkah vi.

b. Klasifikasi dengan Regresi Logistik Ridge
i. Membagi data menjadi data *training* dan *testing* menggunakan 10-*fold cross validation* dengan metode stratifikasi.
ii. Melakukan klasifikasi data *training* menggunakan Regresi Logistik Ridge.
iii. Menghitung ketepatan klasifikasi pada data *testing* dengan semua variabel menggunakan persamaan (2.40) sampai (2.43) dan persamaan (2.47).
iv. Menghitung ketepatan klasifikasi dengan variabel signifikan yang diperoleh dari langkah 2.a.(vi).

c. Klasifikasi dengan Analisis Diskriminan Kernel
i. Membagi data menjadi data *training* dan *testing* menggunakan 10-*fold cross validation* dengan metode stratifikasi.
ii. Melakukan uji asumsi yang meliputi uji homogenitas, uji multivariat normal, dan uji beda rata-rata antar kelompok.
iii. Jika seluruh asumsi terpenuhi, maka data status ketertinggalan desa diklasifikasikan dengan analisis

diskriminan linier. Jika seluruh asumsi tidak terpenuhi, maka digunakan analisis diskriminan kernel dengan pendekatan *Fisher*.

iv. Melakukan klasifikasi data *training* menggunakan Analisis Diskriminan Kernel.

v. Menghitung ketepatan klasifikasi pada data *testing* dengan semua variabel menggunakan persamaan (2.40) sampai (2.43) dan persamaan (2.47).

vi. Menghitung ketepatan klasifikasi menggunakan variabel signifikan yang diperoleh pada langkah 2.a.(vi).

3. Melakukan penanganan kondisi data *imbalanced* dengan metode *Combine sampling* pada Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan menggunakan bantuan *software R*.

   a. Melakukan penanganan kondisi data *imbalanced* dengan metode sampling SMOTE sebagai berikut.

      i. Menentukan jumlah data kelas mayoritas dan kelas minoritas

      ii. Menentukan persentase SMOTE yang digunakan (N%) dengan cara (jumlah data kelas mayoritas/jumlah data kelas minoritas) x 100%

      iii. Menentukan data *k-Nearest Neighbour* (x_{knn}) dengan jarak terdekat dari setiap data minoritas yang akan disintesis.

      iv. Menghitung data sintetis dengan persamaan (2.1) yaitu $x_{syn} = x_i + (x_{knn} - x_i) \times \delta$, dimana δ antara 0 dan 1.

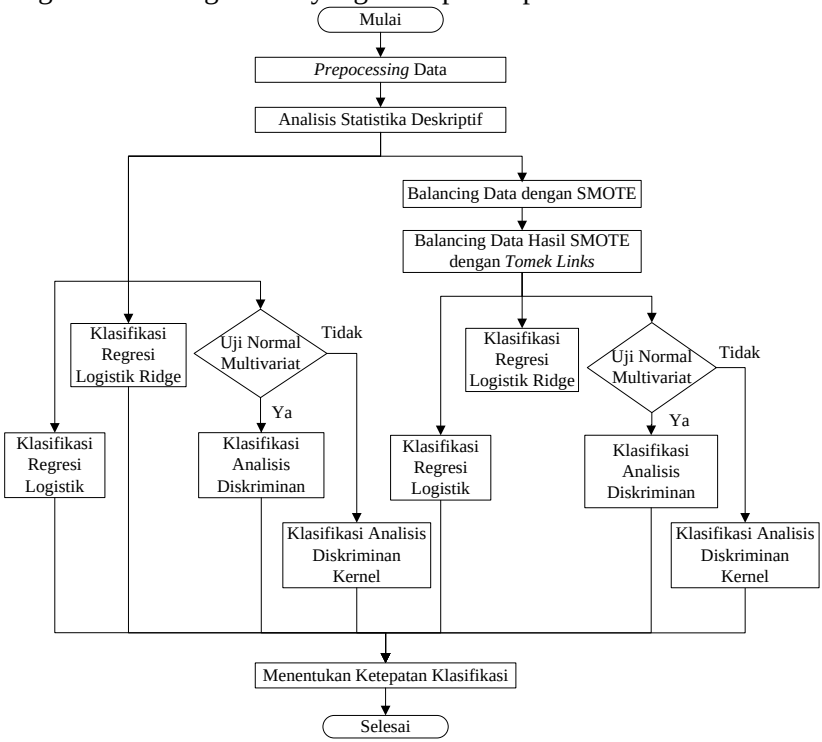
   b. Hasil dari sampling SMOTE, kemudian disampling kembali dengan metode *Tomek Links.* Prosedur untuk menemukan *Tomek Links* adalah bekerja dengan pengecekan setiap data dari kelas yang berbeda. Apabila ditemukan sepasang data yang memiliki kelas label berbeda dan merupakan kasus *Tomek Links*, maka data dari kelas mayoritas akan dihapus dari data *training*

sampai menghasilkan data *training* yang bersih dari *noise* dan *borderline*.

c. Melakukan klasifikasi menggunakan data *balanced* dengan Regresi Logistik seperti langkah 2.a.
d. Melakukan klasifikasi menggunakan data *balanced* dengan Regresi Logistik Ridge seperti langkah 2.b.
e. Melakukan klasifikasi menggunakan data *balanced* dengan Analisis Diskriminan Kernel seperti langkah 2.c.

4. Membandingkan efektivitas *Combine Sampling* pada data semua variabel dan data dengan variabel signifikan dengan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.

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



Gambar 3. 1 Diagram Alir Penelitian

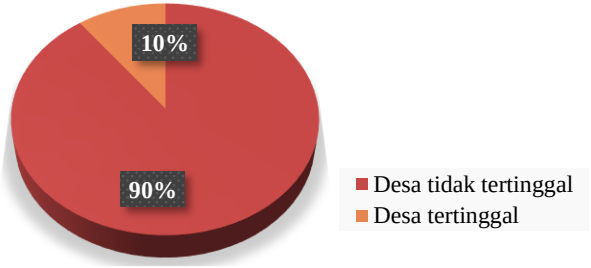
BAB IV

ANALISIS DAN PEMBAHASAN

Pada bab ini akan diuraikan mengenai karakteristik data PODES Jawa Timur tahun 2014 dan hasil klasifikasi desa tertinggal di Jawa Timur dengan data *imbalanced* (data asli) dan data yang telah *balanced* (dilakukan resampling dengan *Combine sampling*) menggunakan metode Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel. Pembagian data testing dan data training digunakan 10-*fold cross validation* dengan metode stratifikasi. Selanjutnya hasil dari masing-masing metode akan dibandingkan untuk memilih metode yang dianggap terbaik dengan menghitung ketepatan klasifikasinya menggunakan nilai akurasi dan *G-mean*.

4.1 Deskripsi Karakteristik Data Status Ketertinggalan Desa

Data PODES tahun 2014 yang digunakan pada penelitian ini adalah desa di Jawa Timur yang dipilih berdasarkan 5 kabupaten yang memiliki persentase jumlah desa tertinggal tertinggi. Apabila dilihat 5 urutan teratas kabupaten yang memiliki desa tertinggal, terdapat Kabupaten Bangkalan yang memiliki persentase desa tertinggal tertinggi, kemudian terdapat Kabupaten Situbondo diurutan kedua, Kabupaten Sumenep urutan ketiga, Kabupaten Bondowoso di urutan keempat, dan yang kelima ada Kabupaten Sampang. Berikut ini ditampilkan persentase jumlah desa tertinggal dan tidak tertinggal pada data 5 kabupaten tersebut.



Gambar 4. 1 Rasio Desa Tertinggal pada Data 5 Kabupaten di Jawa Timur

Berdasarkan Gambar 4.2 dapat diketahui bahwa dari 5 kabupaten tersebut, diperoleh persentase desa tertinggal sebanyak 10% atau 115 desa, sedangkan desa tidak tertinggal sebesar 90% atau 1007 desa. Selanjutnya variabel penelitian yang akan dianalisis statistika deskriptif pada masing-masing indikator status ketertinggalan desa berdasarkan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi yaitu jumlah SD/MI, jumlah tempat praktik bidan, jumlah poskesdes, jumlah toko kelontong, jumlah keluarga pengguna listrik, jarak tempuh per kilometer ke kantor camat, jumlah penderita gizi buruk, dan jumlah pendapatan asli desa dapat dilihat dalam Tabel 4.1 berikut.

Tabel 4. 1 Statistika Deskriptif 5 Kabupaten yang Memiliki Desa Tertinggal Tertinggi

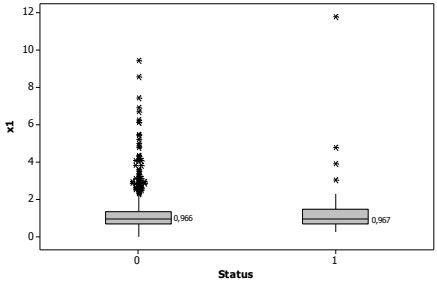
Variabel	Status	Rata-Rata	Varians	Min	Max
Jumlah SD/MI	0	4	8,11	0	27
	1	3	3,28	1	9
Jumlah Tempat Praktik Bidan	0	1	1,23	0	16
	1	1	0,58	0	4
Jumlah Poskesdes	0	1	0,37	0	11
	1	0	0,30	0	2
Jumlah Toko Kelontong	0	35	2092,51	0	480
	1	14	155,10	0	61
Jumlah Keluarga Pengguna Listrik	0	1251	647587,20	52	9154
	1	930	259182,10	115	2513
Jarak ke Kantor Camat	0	5,34	56,85	1	197
	1	10,86	245,03	1	164
Jumlah Penderita Gizi Buruk	0	1	15,20	0	98
	1	1	3,32	0	16
Jumlah Pendapatan Asli Desa (jutaan rupiah)	0	23,80	1300,10	0	386
	1	14,00	324,13	0	89

Keterangan kategori desa :
0 = desa tidak tertinggal
1 = desa tertinggal

Tabel 4.1 menunjukkan bahwa rata-rata jumlah SD/MI untuk desa tidak tertinggal adalah 4 sekolah dengan Desa Tlambah

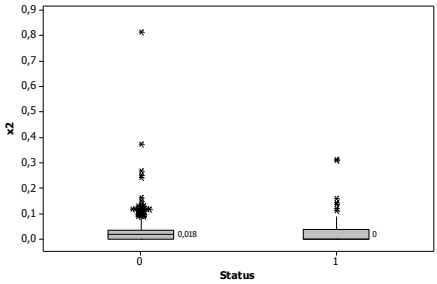
Kabupaten Sampang memiliki jumlah SD/MI terbanyak, sedangkan masih terdapat 11 desa yang tidak memiliki SD/MI. Rata-rata jumlah SD/MI pada desa tertinggal adalah sebanyak 3 sekolah dengan Desa Pajeruan dan Desa Palenggiyan Kabupaten Sampang memiliki jumlah SD/MI terbanyak. Pada desa tidak tertinggal maupun desa tertinggal rata-rata memiliki 1 tempat praktik. Jumlah tempat praktik terbanyak di desa tidak tertinggal berada di Desa Kolor Kabupaten Sumenep, tetapi masih terdapat 162 desa yang tidak memiliki tempat praktik bidan . Sedangkan pada desa tertinggal jumlah tempat praktik bidan terbanyak berada pada Desa Larangan Timur Kabupaten Bangkalan dan sebanyak 37 desa belum memiliki tempat praktik bidan. Rata-rata jumlah poskesdes disetiap desa tidak tertinggal ada 1 unit dengan jumlah terbanyak ada di Desa Payudan Dundang Kabupaten Sumenep, sedangkan terdapat 424 desa belum memiliki poskesdes. Selain itu, disetiap desa tertinggal rata-rata tidak memiliki poskesdes, hal ini disebabkan karena sebanyak 63 desa masih belum memiliki poskesdes dan jumlah poskesdes terbanyak ada di desa Montorna dan Sawah Sumur Kabupaten Sumenep, serta Desa Pendabah Kabupaten Bangkalan. Berdasarkan jumlah toko kelontong, rata-rata jumlahnya ada 35 toko tiap desa dengan Desa Kalianget Barat Kabupaten Sumenep memiliki jumlah toko kelontong terbanyak dan masih terdapat 15 desa yang tidak memiliki toko kelontong. Pada desa tertinggal, rata-rata jumlah toko kelontong ada 14 toko dengan Desa Kalisari Kabupaten Situbondo memiliki jumlah toko kelontong terbanyak, sedangkan masih terdapat 5 desa yang tidak memiliki toko kelontong. Jumlah keluarga pengguna listrik pada desa tidak tertinggal memiliki rata-rata sebanyak 1251 keluarga dengan pengguna listrik terbanyak berada di Desa Sumberejo Kabupaten Situbondo dan terendah berada di Desa Cendagah Kabupaten Bangkalan. Selanjutnya, rata-rata jumlah keluarga pengguna listrik pada desa tertinggal adalah sebanyak 930 keluarga dengan Desa Cangkraman Kabupaten Sumenep memiliki jumlah keluarga pengguna listrik terbanyak dan terendah berada di Desa Tlagah Kabupaten Bangkalan. Rata-rata jarak dari kantor desa/

lurah ke kantor Camat pada desa tidak tertinggal sejauh 5,34 km dengan jarak terjauh ditempuh desa Karamian Kabupaten Sumenep dan jarak terdekat sejauh 1 km. Pada desa tertinggal jarak rata-rata yang harus ditempuh ke kantor Camat adalah 10,86 km dengan jarak terjauh ditempuh desa Masakambing Kabupaten Sumenep dan jarak terdekat minimal harus menempuh 1 km. Selanjutnya, jumlah penderita gizi buruk pada desa tidak tertinggal maupun desa tertinggal selama 3 tahun terakhir rata-rata sebanyak 1 orang penderita. Pada desa tidak tertinggal, jumlah penderita gizi buruk terbanyak berada di Desa Sapeken Kabupaten Sumenep dan sudah banyak desa-desa yang tidak memiliki penderita gizi buruk. Sama halnya dengan desa tidak tertinggal, beberapa desa tertinggal sudah tidak memiliki penderita gizi buruk . Jumlah penderita gizi buruk terbanyak pada desa tertinggal berada di desa Kembangsari Kabupaten Situbondo. Jumlah pendapatan asli desa terbesar pada desa tidak tertinggal berada di Desa Nogosari Kabupaten Bondowoso , sedangkan masih terdapat beberapa desa ada yang tidak memiliki pendapatan asli desa dan rata-rata pendapatan setiap desa sebesar Rp23.800.000,-. Pada desa tertinggal, rata-rata jumlah pendapatan asli desa yang diterima sebesar Rp14.000.000,-. Jumlah PAD terbesar berada di Desa Suwaan Kabupaten Bangkalan dan masih terdapat beberapa desa yang tidak memiliki PAD. Apabila dilihat dari nilai varians, variabel jumlah toko kelontong , jumlah keluarga pengguna listrik, jarak ke kantor Camat, dan jumlah pendapatan asli desa memiliki varians cukup tinggi. Hal ini menunjukkan bahwa tingkat keberagaman tiap desa cukup tinggi, dimana beberapa desa masih memiliki ketimpangan yang besar antar satu desa dengan desa yang lain. Selain itu, karakteristik status ketertinggalan desa dapat dilihat nilai rasio masing-masing variabel menggunakan *boxplot*. Karena setiap pelayanan yang ada di desa mempertimbangkan jumlah penduduk yang menetap di suatu daerah tersebut.Berikut adalah *boxplot* dari rasio banyaknya SD/MI.



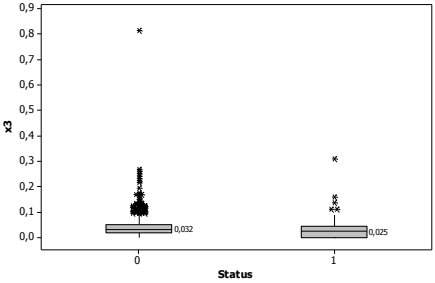
Gambar 4. 2 *Boxplot* Rasio Banyaknya SD/MI

Sekolah Dasar atau MI adalah sarana penunjang yang penting untuk pendidikan anak pada usia dini. Dapat diketahui bahwa dari Gambar 4.3 median kedua kelompok hampir sama yaitu pada desa tertinggal sebesar 0,967 sedangkan pada desa tidk tertinggal sebesar 0,966. Selain itu varians pada desa tertinggal lebih besar dibandingkan desa tidak tertinggal yang dapat dilihat dari lebar *boxplot* pada masing-masing kelompok. Selanjutnya *boxplot* pada rasio banyaknya tempat praktik bidan.



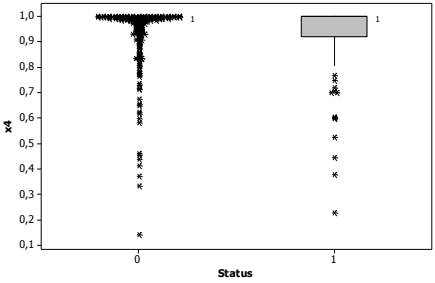
Gambar 4. 3 *Boxplot* Rasio Banyaknya Tempat Praktik Bidan

Median rasio banyaknya tempat praktik bidan ternyata lebih tinggi pada desa tidak tertinggal. Sepert terlihat pada Gambar 4.4 diketahui bahwa median pada desa tertinggal sebesar 0,018 sedangkan desa tertinggal sebesar 0. Kemudian dilihat pula *boxplot* rasio banyaknya poskesdes sebagai berikut.



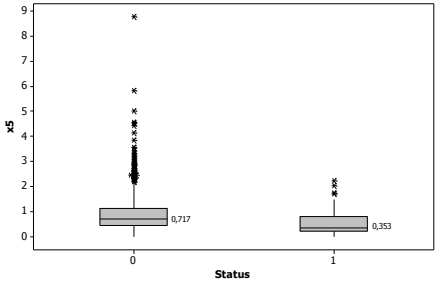
Gambar 4. 4 *Boxplot* Rasio Banyaknya Poskesdes

Pada rasio banyaknya poskesdes diketahui bahwa median di desa tidak tertinggal lebih tinggi dibandingkan di desa tidak tertinggal yaitu 0,032 pada desa tidak tertinggal dan 0,025 pada desa tertinggal. Selain itu dapat pula diketahui keragaman rasio banyaknya poskesdes lebih besar pada desa tidak tertinggal yang dapat dilihat dari lebar *boxplot*. Selanjutnya pada rasio toko kelontong dapat dilihat *boxplot* sebagai berikut.



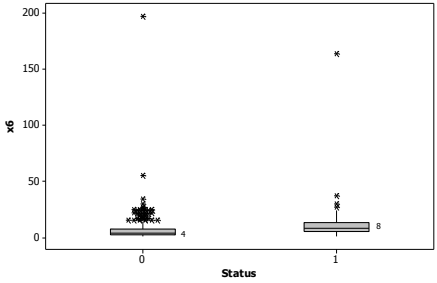
Gambar 4. 5 *Boxplot* Rasio Banyaknya Toko Kelontong

Dapat diketahui pada Gambar 4.7 bahwa rasio banyaknya toko kelontong memiliki median yang sama baik pada desa tertinggal maupun desa tidak tertinggal yaiu sebesar 1. Tetapi varians pada desa tertinggal lebih besar dibandingkan pada desa tidak tertinggal yang menunjukkan bahwa tingkat keragaman rasio jumlah toko kelontong pada masing-masing desa masih cukup tinggi. Selain itu, pada rasio banyaknya keluarga pengguna listrik dapat dilihat *boxplot* sebagai berikut.



Gambar 4. 6 *Boxplot* Rasio Banyaknya Keluarga Pengguna Listrik

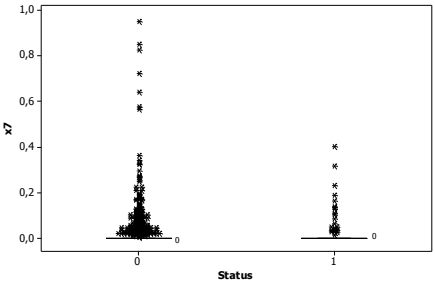
Rasio banyaknya pengguna listrik memiliki median yang lebih tinggi pada desa tidak tertinggal yaitu sebesar 0,717. Sedangkan pada desa tertinggal diperoleh median sebesar 0,353. Jika dilihat dari keragaman varians, pada desa tidak tertinggal memiliki varians yang lebih tinggi dibandingkan pada desa tertinggal. Pada jarak tempuh per kilometer ke kantor camat, dapat dilihat pula *boxplot* sebagai berikut.



Gambar 4. 7 *Boxplot* Jarak Tempuh Ke Kantor Camat

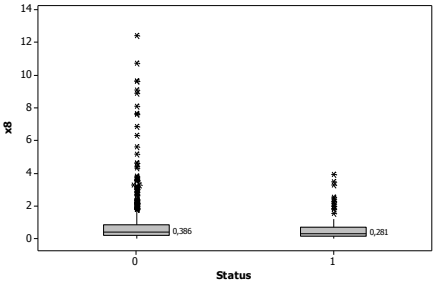
Semakin jauhnya jarak tempuh suatu desa ke kantor kecamatan, dapat menjadi salah satu indikator penentu desa tertinggal. Dapat diketahui pada Gambar 4.9 bahwa pada desa tertinggal memiliki median sebesar 8 km yang lebih besar dibandingkan pada desa tidak tertinggal yaitu memiliki median sebesar 4 km. Tetapi varians pada desa tertinggal lebih besar dibandingkan pada desa tidak tertinggal yang terlihat dari lebar

boxplot masing-masing kelompok. Selain itu, rasio banyaknya penderita gizi buruk dapat ditampilkan pada *boxplot* berikut.



Gambar 4. 8 *Boxplot* Rasio Banyaknya Penderita Gizi Buruk

Gambar 4.10 menunjukkan bahwa pada rasio banyaknya penderita gizi buruk, median desa tertinggal dan desa tertinggal sama yaitu sebesar 0. Kemudian pada rasio pendapatan asli desa dapat dilihat *boxplot* sebagai berikut.



Gambar 4. 9 *Boxplot* Rasio Pendapatan Asli Desa

Pada rasio pendapatan asli desa dapat diketahui bahwa median pada desa tidak tertinggal lebih besar dibandingkan desa tertinggal dan varians pada desa tidak tertinggal juga lebih besar dibandingkan pada desa tidak tertinggal.

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

Data status ketertinggalan desa berdasarkan 5 kabupaten yang memiliki desa tertinggal tertinggi akan dianalisis meng-

gunakan Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel dengan hasil sebagai berikut.

4.2.1 Regresi Logistik pada Data *Imbalanced*

Regresi Logistik merupakan salah metode yang dapat digunakan untuk pengklasifikasian status ketertinggalan desa di Jawa Timur. Pada analisis ini digunakan 8 variabel prediktor, sedangkan variabel respon yang digunakan adalah status ketertiggalan desa yaitu terdapat 115 desa tertinggal dan 1007 desa tidak tertinggal.

A. Regresi Logistik dengan Semua Variabel

Ketepatan klasifikasi Regresi Logistik yang digunakan pada semua variabel nantinya akan dibandingkan efektifitas performansinya dengan menggunakan variabel yang signifikan dengan hasil sebagai berikut.

i. Deteksi Multikolinearitas Semua Variabel

Sebelum dilakukan analisis dengan metode Regresi Logistik, terlebih dahulu dilakukan deteksi multikolinearitas terhadap 8 variabel prediktor yang digunakan dalam analisis. Berikut adalah nilai VIF pada masing-masing variabel yang ditampilkan pada Tabel 4.2.

Tabel 4. 2 Nilai *Variance Inflation Factors* (VIF) Data *Imbalanced*

Variabel	VIF
X_1	1,009
X_2	1,510
X_3	1,538
X_4	1,078
X_5	1,041
X_6	1,061
X_7	1,008
X_8	1,021

Tabel 4.2 menunjukkan bahwa nilai VIF dari masing-masing variabel prediktor memiliki nilai yang kurang dari 5. Hal ini mengindikasikan bahwa tidak terjadi kasus multikolinearitas pada data 5 kabupaten yang memiliki persentase desa tertinggal tertinggi di Jawa Timur. Oleh karena itu, tidak diperlukan penanganan lebih lanjut untuk kasus multikolinearitas.

ii. Ketepatan Klasifikasi Regresi Logistik Semua Variabel

Setelah melakukan pengecekan multikolinearitas pada data, selanjutnya dilakukan pengklasifikasian status ketertinggalan desa di Jawa Timur berdasarkan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi dengan Regresi Logistik menggunakan *stratified* 10-*fold cross validation* yang ditampilkan pada Tabel 4.3.

Tabel 4. 3 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Imbalanced* Seluruh Variabel

Fold	***AUC***		***G-mean***		**Akurasi Total**		**Sensitivitas**		**Spesifisitas**	
	Train	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
1	0.517	0.542	0.197	0.289	0.898	0.903	0.039	0.083	0.996	1.000
2	0.522	0.542	0.220	0.289	0.898	0.903	0.049	0.083	0.994	1.000
3	0.531	0.495	0.259	0.000	0.900	0.893	0.067	0.000	0.996	0.990
4	0.527	**0.581**	0.240	**0.422**	0.900	**0.902**	0.058	**0.182**	0.997	**0.980**
5	0.526	0.500	0.240	0.000	0.898	0.902	0.058	0.000	0.994	1.000
6	0.541	0.495	0.293	0.000	0.901	0.893	0.087	0.000	0.994	0.990
7	0.526	0.500	0.240	0.000	0.898	0.902	0.058	0.000	0.994	1.000
8	0.526	0.500	0.241	0.000	0.899	0.893	0.058	0.000	0.994	1.000
9	0.522	0.500	0.220	0.000	0.899	0.893	0.049	0.000	0.996	1.000
10	0.530	0.537	0.260	0.287	0.897	0.893	0.068	0.083	0.991	0.990
Mean	0.527	0.519	0.241	0.129	0.899	0.897	0.059	0.043	0.995	0.995
Stdev	0.006	0.029	0.026	0.171	0.001	0.005	0.013	0.063	0.001	0.007

Berdasarkan Tabel 4.3 dapat diketahui bahwa model terbaik yang didapat pada *fold* ke-4 karena memiliki nilai AUC, G-*mean*, akurasi total, sensitivitas, dan spesifisistas yang cenderung tinggi pada data testing, dimana nilai ini lebih besar dibanding *fold* lainnya. Selanjutnya, diperoleh rata-rata AUC sebesar 52,7%. Sedangkan rata-rata nilai G-*mean* hanya diperoleh sebesar 12,9% pada data testing. Sedangkan untuk nilai akurasi total pada data testing rata yang didapat cukup besar yaitu sebesar 89,7%. Ketepatan klasifikasi menggunakan regresi logistik diperoleh hasil yang kurang baik. Hal ini diindikasikan karena variabel respon yang digunakan jumlahnya tidak seimbang atau *imbalanced*, sehingga *classifier* cenderung mengklasifikasikan ke dalam kelas mayoritas sehingga kesalahan klasifikasinya cukup tinggi. Hal ini dapat dilihat dari rata-rata nilai sensitivitas pada data testing sebesar 4,3% yang sangat kecil jika dibandingkan nilai spesifitas

pada data testing yaitu sebesar 99,5%. Disamping itu, dapat dilihat bahwa standar deviasi dari masing-masing ketepatan klasifikasi memiliki nilai yang kecil. Hal ini menandakan bahwa keragaman nilai ketepatan klasifikasi pada masing-masing *fold* kecil atau tidak jauh berbeda.

Setelah melakukan ketepatan klasifikasi dengan semua variabel, selanjutnya dilakukan pengujian signifikansi parameter untuk mengetahui variabel yang berpengaruh signifikan. Pada Tabel 4.4, diketahui model terbaik yaitu pada *fold* ke-4. Sehingga dilakukan pengujian parameter pada *fold* ke-4. Apabila dilakukan pengujian secara serentak dengan menggunakan nilai *Likelihood Ratio Test*, diperoleh hasil uji serentak yaitu tolak H_0 karena nilai $G > \chi^2_{(8;0,10)}$ yaitu 592,016 >13,36. Sehingga kesimpulannya secara serentak model pada *fold* ke-4 minimal ada 1 variabel yang berpengaruh signifikan. Selanjutnya dilakukan pengujian secara parsial dengan hasil sebagai berikut.

Tabel 4. 4 Hasil Uji Parsial Pada Regresi Logistik Data *Imbalanced*

Variabel	**Koefisien**	***Std. Error***	$\mathbf{Z_{Hitung}}$	***P-Value***
Konstan	1,582	1,030	1,536	0,125
X_1	0,087	0,103	0,851	0,395
X_2	4,330	3,320	1,304	0,192
X_3	-4,112	3,748	-1,097	0,273
X_4	-3,155	1,078	-2,926	0,003
X_5	-1,375	0,289	-4,755	0,000
X_6	0,035	0,011	3,222	0,001
X_7	0,489	1,353	0,361	0,718
X_8	-0,155	0,144	-1,077	0,282

Dapat diketahui dari Tabel 4.4 bahwa variabel yang signifikan adalah rasio banyaknya toko kelontong (X_4), rasio banyaknya keluarga pengguna listrik (X_5), dan jarak tempuh per kilometer ke kantor camat (X_6) karena memiliki p-*value*<(α=0,10). Pada model tersebut masih terdapat beberapa variabel yang tidak signifikan pada α=0,10. Maka akan dilakukan *backward elimination* untuk memilih variabel signifikan dengan cara mengeluarkan variabel yang paling tidak berpengaruh secara

bertahap. Berikut hasil *backward elimination* yang ditampilkan pada Tabel 4.5.

Tabel 4. 5 Hasil Uji Parsial Pada Regresi Logistik Data *Imbalanced* Variabel Signifikan

Variabel	Koefisien	*Std. Error*	Z_{Hitung}	*P-Value*
Konstan	1,607	1,028	1,563	0,118
X_4	-3,221	1,071	-3,009	0,003
X_5	-1,361	0,281	-4,840	0,003
X_6	0,037	0,011	3,199	0,001

Tabel 4.5 menunjukkan bahwa hasil *backward elimination*, variabel yang signifikan sebanyak empat yaitu rasio banyaknya toko kelontong (X_4), rasio banyaknya keluarga pengguna listrik (X_5), dan jarak tempuh per kilometer ke kantor camat (X_6). Variabel yang signifikan selanjutnya akan digunakan pula pada Analisis Regresi Logistik Ridge dan Analisis Diskriminan Kernel.

B. Regresi Logistik dengan Variabel yang Signifikan

Telah diketahui sebelumnya bahwa terdapat 3 variabel yang signifikan. Seperti dengan analisis Regresi Logistik sebelumnya, terlebih dahulu dilakukan pengecekan multikolinearitas pada variabel yang signifikan.

Tabel 4. 6 Nilai *VIF* Data *Imbalanced* Variabel Signifikan

Variabel	VIF
X_4	1,072
X_5	1,021
X_6	1,051

Nilai VIF yang ditampilkan pada Tabel 4.6 untuk variabel yang signifikan, tidak ada yang lebih dari 5. Hal ini mengindikasikan bahwa tidak terjadi kasus multikolinearitas pada data yang signifikan. Pada data yang memiliki variabel signifikan, selanjutnya dilakukan pengklasifikasian dengan Regresi Logistik menggunakan *stratified* 10-*fold cross validation* yang ditampilkan pada Tabel 4.7.

Tabel 4. 7 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Imbalanced* Variabel Signifikan

	Training	Testing	Stdv Testing
Rata-rata AUC	0,522	0,515	0,029
Rata-rata G-*mean*	0,216	0,100	0,156

Tabel 4. 7 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Imbalanced* Variabel Signifikan (Lanjutan)

	Training	Testing	Stdv Testing
Rata-rata akurasi total	0,898	0,897	0,005
Rata-rata sensitivitas	0,049	0,035	0,059
Rata-rata spesifisitas	0,995	0,995	0,007

Tabel 4.7 menunjukkan bahwa nilai G-*mean* hanya diperoleh sebesar 10% pada data testing. Sedangkan untuk nilai akurasi total pada data testing cukup besar yaitu sebesar 89,7%. Nilai sensitivitas diperoleh sebesar 3,5% pada data testing dan nilai spesifisitas diperoleh sebesar 99,5%%. Nilai sensitivitas menunjukkan ketepatan klasifikasi pada data minoritas, dimana diperoleh hasil klasifikasi yang kurang baik. Sedangkan sensitivitas tinggi menunjukkan ketepatan klasifikasi pada data mayoritas sudah baik. Jika dibandingkan dengan hasil semua variabel, ketepatan klasifikasi yang diperoleh pada data yang signifikan tidak jauh berbeda dan hanya berbeda pada nilai G-*mean* dimana pada semua variabel diperoleh nilai yang lebih tinggi. Disamping itu, dapat diketahui pula bahwa standar deviasi dari masing-masing ketepatan klasifikasi cukup rendah yang mengindikasikan dari 10 *fold* tidak memiliki nilai yang jauh berbeda.

4.2.2 Regresi Logistik Ridge pada Data *Imbalanced*

Setelah dilakukan klasifikasi dengan Regresi Logistik, selanjutnya data status ketertinggalan desa diklasifikasikan menggunakan Regresi Logistik Ridge. Regresi Logistik Ridge merupakan *classifier* pembanding Regresi Logistik dimana digunakan untuk mengakomodasi jika terjadi korelasi antar variabel.

A. Ketepatan Klasifikasi Regresi Logistik Ridge dengan Semua Variabel

Berikut adalah ketepatan klasifikasi dari data training dan data testing menggunakan semua variabel dengan *stratified* 10-*fold cross validation*.

Tabel 4. 8 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge dengan Data *Imbalanced* Seluruh Variabel

	Training	Testing	Stdv Testing
Rata-rata AUC	0,506	0,515	0,028
Rata-rata G-*mean*	0,105	0,100	0,165
Rata-rata akurasi total	0,896	0,898	0,005
Rata-rata sensitivitas	0,014	0,035	0,062
Rata-rata spesifisitas	0,997	0,996	0,007

Berdasarkan Tabel 4.8 dapat diketahui rata-rata AUC pada data testing sebesar 51,5%. Kemudian diperoleh rata-rata G-*mean* pada data testing sebesar 10%. Sedangkan nilai akurasi total pada data testing sebesar 89,8% yang menunjukkan akurasi tinggi. Nilai sensitivitas pada data testing hanya sebesar 3,5% pada data testing yang menunjukkan banyaknya terjadi kesalahan klasifikasi pada data minor. Sedangkan nilai sepesifisitasnya diperoleh sebesar 99,6% yang menunjukkan bahwa klasifikasi pada data mayoritas sudah baik. Hal ini menunjukkan bahwa *classifier* cenderung mengklasifikasikan ke dalam data mayoritas. Standar deviasi masing-masing ketepatan klasifikasi cukup kecil yang menunjukkan bahwa dari masing-masing *fold* memiliki keragaman yang kecil perbedaan yang jauh berbeda.

B. Ketepatan Klasifikasi Regresi Logistik Ridge dengan Variabel Signifikan

Setelah dilakukan ketepatan klasifikasi pada semua variabel, selanjutnya dihitung pula ketepatan klasifikasi pada variabel signifikan menurut hasil Regresi Logistik dengan *stratified* 10-*fold cross validation* berikut.

Tabel 4. 9 Hasil Ketepatan Klasifikasi Regresi Logistik Ridge dengan Data *Imbalanced* Variabel Signifikan

	Training	Testing	Stdv Testing
Rata-rata AUC	0,512	0,515	0,028
Rata-rata G-*mean*	0,141	0,100	0,156
Rata-rata akurasi total	0,897	0,898	0,005
Rata-rata sensitivitas	0,026	0,035	0,059
Rata-rata spesifisitas	0,996	0,996	0,007

Rata-rata AUC yang ditunjukkan pada Tabel 4.9 adalah sebesar 51,5% pada data testing, sedangkan rata-rata G-*mean* pada

data testing diperoleh sebesar 10%. Selanjutnya nilai akurasi total pada data testing sebesar 89,8% yang menunjukkan akurasi tinggi. Tetapi jika dilihat dari sensitivitas ternyata hanya diperoleh sebesar 3,5%. Hal ini menunjukkan bahwa banyak terjadi kesalahan klasifikasi pada data minoritas. Kemudian nilai senstifitasnya diperoleh sebesar 99,6% pada data testing yang menunjukkan bahwa klasifikasi pada data mayoritas sudah baik. Hal ini menunjukkan bahwa *classifier* cenderung mengklasifikasikan ke dalam data mayoritas. Standar deviasi masing-masing ketepatan klasifikasi cukup kecil yang menunjukkan bahwa dari masing-masing *fold* tidak memiliki perbedaan yang jauh berbeda. Jika dibandingkan dengan semua variabel, diperoleh ketepatan klasifikasi yang sama dengan variabel yang signifikan pada data testing, perbedaannya hanya pada data training dimana pada data variabel signifikan mayoritas memiliki ketepatan klasifikasi yang lebih tinggi.

4.2.3 Analisis Diskriminan Kernel pada Data *Imbalanced*

Analisis diskriminan digunakan untuk mengetahui performa klasifikasi ketertinggalan desa dari data 5 kabupaten yang memiliki persentase desa tertinggal tertinggi. Pada penelitian ini akan dilakukan pengujian pada semua variabel dan variabel yang signifikan dengan Analisis Diskriminan Kernel.

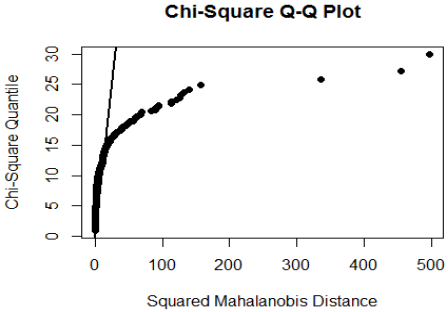
A. Analisis Diskriminan Kernel dengan Semua Variabel

Dalam melakukan analisis diskriminan, terlebih dahulu dilakukan pengujian asumsi apakah telah memenuhi atau belum. Uji asumsi yang dilakukan adalah uji normal multivariat dan uji homogenitas. Berikut adalah hasil uji asumsi dari data asli yang memiliki variabel respon *imbalanced* atau tidak seimbang.

i. Uji Normal Multivariat

Pada pengujian normal multivariat dilakukan dengan *mardia's test*. Hasil dari uji normal multivariat menghasilkan *p-value* sebesar 0. Oleh karena itu, diperoleh keputusan tolak H_0 karena *p-value* kurang dari $\alpha = 0,05$. Hal ini menunjukkan bahwa data status ketertinggalan desa di Jawa Timur tidak berdistribusi normal multivariat, sehingga asumsi distribusi normal multivariat

tidak terpenuhi. Berikut ini adalah QQ-plot dari data penelitian yang disajikan dalam Gambar 4.12.



Gambar 4. 10 Chi-Squared QQ-Plot Data PODES 2014 Semua Variabel

Berdasarkan Gambar 4.3 dapat diketahui bahwa banyak sebaran titik hitam tersebar jauh dari garis, hal ini mengindikasikan bahwa data tidak berdistribusi normal multivariat.

ii. Uji Homogenitas

Pengujian homogenitas dilakukan dengan menghitung statistik uji berdasarkan persamaan (2.5) yang telah dipaparkan pada bab II Tinjauan Pustaka sebelumnya. Hasil uji homogenitas diperoleh *p-value* sebesar 0 dan nilai *Chi-Square* sebesar 469,55. Karena *p-value* kurang dari $\alpha = 0,05$, maka diperoleh keputusan tolak H_0. Hal ini menunjukkan bahwa matriks varians kovarians pada data status ketertinggalan desa di Jawa Timur tidak homogen, sehingga asumsi homogenitas data tidak terpenuhi.

iii. Hasil Ketepatan Klasifikasi dengan Semua Variabel

Setelah diketahui bahwa asumsi uji multivariat normal dan homogenitas data tidak terpenuhi, maka analisis diskriminan yang digunakan adalah analisis diskrminan kernel. Berikut adalah hasil ketepatan klasifikasi status ketertinggalan desa di Jawa Timur pada Tabel 4.10.

Tabel 4. 10 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel dengan Data *Imbalanced* Semua Variabel

	Training	**Testing**	**Stdv Testing**
Rata-rata AUC	0,604	0,584	0,044
Rata-rata G-*mean*	0,479	0,428	0,104
Rata-rata akurasi total	0,895	0,890	0,016

Tabel 4. 10 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel dengan Data *Imbalanced* Semua Variabel (Lanjutan)

	Training	**Testing**	**Stdv Testing**
Rata-rata sensitivitas	0,237	0,199	0,090
Rata-rata spesifisitas	0,970	0,969	0,019

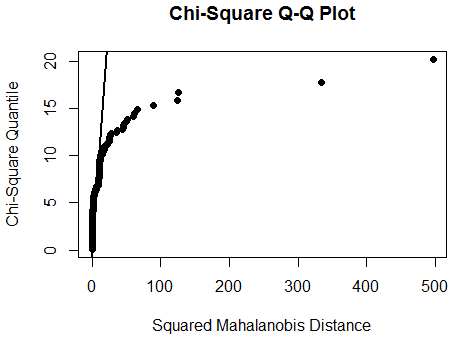
Rata-rata G-*mean* yang diperoleh dari Tabel 4.10 adalah 42,8% pada data testing. Sedangkan nilai akurasi total yang diperoleh pada data testing yaitu sebesar 89% yang dapat dikatakan tinggi. Selanjutnya, rata-rata nilai sensitivas kecil yaitu hanya sebesar 19,9% pada data testing dan rata-rata nilai spesifisitas yang cukup tinggi yaitu sebesar 96,9% pada data testing. Nilai sensitivitas diperoleh hasil yang kecil karena banyak terjadi kesalahan klasifikasi pada kelas minoritas, sedangkan nilai spesifisitas yang tinggi menandakan bahwa klasifikasi pada data kelas mayoritas sudah baik. Hal ini mengindikasikan bahwa *classifier* cenderung mengklasifikasikan ke dalam data mayoritas. Dapat diketahui pula bahwa standar deviasi dari masing-masing *fold* kecil, sehingga dapat diasumsikan bahwa ketepatan klasifikasi yang diperoleh hampir sama.

B. Analisis Diskriminan Kernel dengan Variabel Signifikan

Setelah melakukan analisis dengan semua variabel, selanjutnya kan dibandingkan ketepatan klasifikasi dari variabel yang signifikan berdasarkan Regresi Logistik dengan hasil sebagai berikut.

i. Uji Normal Multivariat

Pada pengujian menggunakan variabel yang signifikan diperoleh hasil dari uji normal multivariat menghasilkan *p-value* sebesar 0. Oleh karena itu, diperoleh keputusan tolak H_0 karena *p-value* kurang dari $\alpha = 0,05$. Hal ini menunjukkan bahwa data dengan variabel signifikan tidak berdistribusi normal multivariat, sehingga asumsi distribusi normal multivariat tidak terpenuhi. Jika dilihat dari QQ-plot diketahui bahwa diketahui bahwa banyak sebaran titik hitam tersebar jauh dari garis, hal ini mengindikasikan bahwa data tidak berdistribusi normal multivariat. Berikut ini adalah QQ-plot dari data penelitian yang disajikan dalam Gambar 4.4.



Gambar 4. 11 Chi-Squared QQ-Plot Data PODES 2014 Variabel Signifikan

ii. Uji Homogenitas

Hasil uji homogenitas pada data variabel yang signifikan diperoleh *p-value* sebesar 0 dan nilai Chi-Square sebesar 366,56. Karena *p-value* kurang dari $\alpha = 0,05$, maka diperoleh keputusan tolak H_0. Hal ini menunjukkan bahwa matriks varians kovarians pada data pada variabel yang signifikan tidak homogen, sehingga asumsi homogenitas data tidak terpenuhi.

iii. Hasil Ketepatan Klasifikasi dengan Variabel Signifikan

Setelah diketahui bahwa asumsi uji multivariat normal dan homogenitas data tidak terpenuhi, maka analisis diskriminan yang digunakan adalah analisis diskrminan kernel. Selanjutnya dengan data pada variabel yang signifikan diperoleh ketepatan klasifikasi dengan Analisis Diskriminan Kernel menggunakan *stratified* 10-*fold cross validation* dengan hasil sebagai berikut.

Tabel 4. 11 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel dengan Data *Imbalanced* Variabel Signifikan

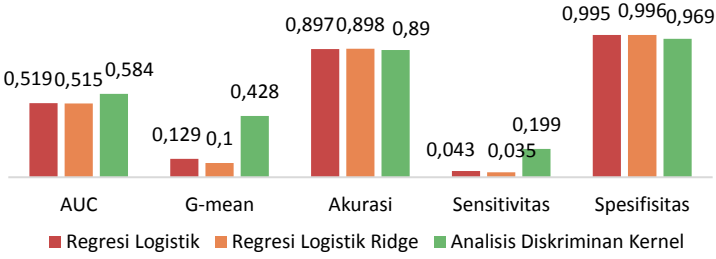
	Training	**Testing**	**Stdv Testing**
Rata-rata AUC	0,603	0,584	0,053
Rata-rata G-*mean*	0,477	0,409	0,165
Rata-rata akurasi total	0,895	0,892	0,017
Rata-rata sensitivitas	0,236	0,197	0,102
Rata-rata spesifisitas	0,971	0,971	0,018

Tabel 4.11 menunjukkan bahwa diperoleh rata-rata AUC sebesar 58,4%, sedangkan diperoleh rata-rata G-*mean* cukup kecil yaitu sebesar 40,9% pada data testing. Kemudian rata-rata akurasi

total pada data testing diperoleh nilai yang cukup tinggi yaitu 89,2%. Selain itu, rata-rata sensitivitas yang diperoleh cukup kecil yaitu 19,7 % pada data testing yang mengindikasikan bahwa banyak terjadi kesalahan klasifikasi pada data minoritas. Sedangkan rata-rata spesifisitas diperoleh sebesar 97,1% pada data testing yang menandakan bahwa data mayoritas memiliki ketepatan klasifikasi yang baik. Hal ini mengindikasikan bahwa *classifier* cenderung mengklasifikasikan ke dalam data mayoritas. Selanjutnya dapat dilihat standar deviasi dari masing-masing ketepatan klasifikasi diperoleh nilai yang kecil, sehingga dapat diindikasikan bahwa masing-masing *fold* memiliki keragaman ketepatan klasifikasi yang kecil. Apabila dibandingkan ketepatan klasifikasi dengan Analisis Diskriminan Kernel pada data semua variabel dan variabel yang signifikan, diperoleh hasil yang tidak jauh berbeda. Tetapi pada nilai G-*mean* pada semua variabel memiliki rata-rata yang lebih tinggi dibandingkan dengan variabel yang signifikan.

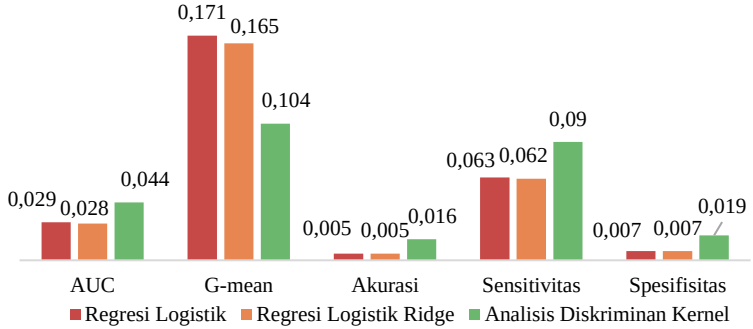
4.2.4 Analisis Gabungan Pada Data *Imbalanced* Semua Variabel dan Variabel Signifikan

Sebelumnya telah dilakukan klasifikasi status ketertinggalan desa berdasarkan 5 kabupaten yang memiliki persentase desa tertinggal tertinggi dengan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan semua variabel maupun menggunakan variabel yang signifikan. Berikut adalah perbandingan ketiga metode dengan semua variabel.



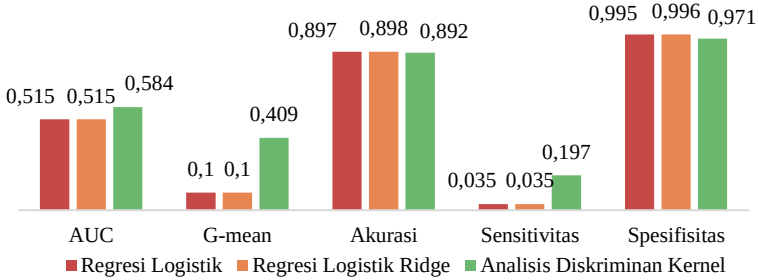
Gambar 4. 12 Perbandingan Ketepatan Klasifikasi Data Im*balanced* Semua Variabel

Gambar 4.12 menunjukkan bahwa Regresi Logistik dan Regresi Logistik Ridge memiliki ketepatan klasifikasi yang tidak jauh berbeda. Begitu pula dengan Analisis Diskriminan Kernel, memiliki nilai yang lebih tinggi dibandingkan kedua metode tersebut. Selain itu, pada ketiga metode diperoleh rata-rata AUC, G-*mean* dan sensitivitas yang kecil. Sedangkan akurasi total dan spesifisitas diperoleh rata-rata yang tinggi. Hal ini mengindikasikan bahwa terjadi suatu masalah dimana *classifier* cenderung mengklasifikasikan ke data mayoritas. Berdasarkan ketiga metode tersebut dengan menggunakan semua variabel, metode yang memiliki rata-rata AUC, G-*mean*, akurasi total, sensitivitas, dan spesifisitas tertinggi adalah Analisis Diskriminan Kernel. Berikut ini adalah perbandingan ketiga metode dengan variabel signifikan.



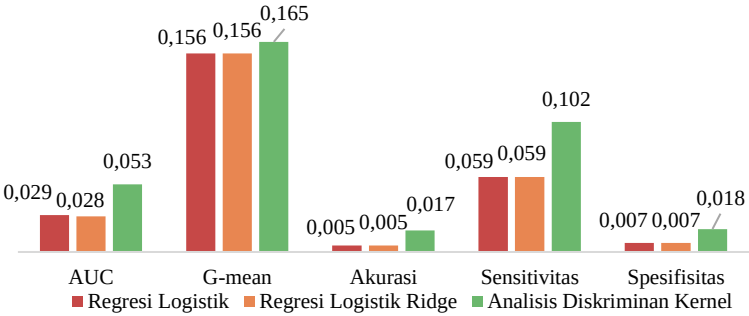
Gambar 4. 13 Perbandingan Standar Deviasi Data Im*balanced* SemuaVariabel

Perbandingan standar deviasi pada Gambar 4.13 menunjukkan bahwa Regresi Logistik dan Regresi Logistik Ridge memiliki standar deviasi yang hampir sam, sedangakan Analisis Diskriminan Kernel memiliki standar deviasi yang cukup berbeda dengan kedua metode tersebut. Analisis Diskriminan memiliki standar deviasi yang tinggi pada AUC, akurasi total, sensitivitas, dan spesifisitas.



Gambar 4. 14 Perbandingan Ketepatan Klasifikasi Data Im*balanced* Variabel Signifikan

Hampir sama dengan hasil semua variabel, berdasarkan Gambar 4.14 diketahui bahwa Regrsei Logistik dan Regresi Logistik Rdge memiliki ketepatan klasifikasi yang tidak jauh berbeda, bahkan memiliki rata-rata AUC dan G-mean yang sama yaitu sebesar 51,5% dan 10%. Begitu pula dengan Analisis Diskriminan memiliki hasil yang tidak jauh berbeda, tetapi masih lebih tinggi dibandingkan kedua metode lainnya. Jika dibandingkan dengan data semua variabel, pada variabel signifikan ketepatan klasifikasi yang diperoleh cenderung turun.



Gambar 4. 15 Perbandingan Standar Deviasi Data Im*balanced* Variabel Signifikan

Gambar 4.15 menunjukkan bahwa seperti pada data semua variabel, Regresi Logistik dan Regresi Logistik Ridge memiliki standar deviasi yang tidak jauh berbeda. Tetapi dari 3 metode tersebut, Analisis Diskriminan Kernel memiliki standar deviasi

yang lebih tinggi dibandingkan kedua metode tersebut pada semua ukuran ketepatan klasifikasi.

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

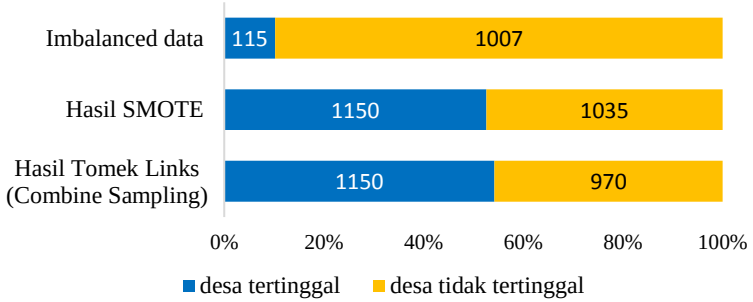
Data status ketertinggalan desa dari 5 kabupaten yang memiliki persentase desa tertinggal tertinggi, mempunyai komposisi variabel respon yang tidak seimbang, yaitu jumlah desa tertinggal sebanyak 115 unit desa dan desa tidak tertinggal sebanyak 1007 desa. Oleh karena itu, untuk meningkatkan akurasi dilakukan resampling data menggunakan *combine sampling* dengan hasil sebagai berikut.

4.3.1 Metode *Combine Sampling*

Telah diketahui bahwa data PODES Jawa Timur tahun 2014 yang digunakan memiliki komposisi variabel respon yaitu desa tertinggal dan desa tidak tertinggal yang tidak seimbang atau *imbalanced*. Oleh karena itu, dilakukan resampling data dengan metode *combine sampling* untuk meningkatkan ketepatan klasifikasi.

Metode *combine sampling* merupakan perpaduan metode *oversampling* dan *undersampling* yaitu SMOTE dan *Tomek Links*. Penggunaan kedua metode dilakukan secara berurutan. Langkah awal, data *imbalanced* dilakukan resampling menggunakan SMOTE sehingga data kelas minoritas akan seimbang dengan kelas mayoritas. Setelah itu, data hasil resampling dengan SMOTE, dilanjutkan resampling kembali menggunakan *Tomek Links* untuk menghapus data *noise* ataupun *borderline*. Hasil penanganan dengan metode *combine sampling* menunjukkan perubahan persentase dari setiap kelas minoritas maupun kelas mayoritas. Data hasil *combine sampling* menghasilkan komposisi data yang cukup seimbang, yaitu persentase desa tertinggal dan desa tidak tertinggal adalah 54,87% : 45,13%. Selanjutnya data *combine sampling* akan dilakukan analisis Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel. Berikut ditampilkan perubahan yang terjadi pada data *imbalanced* menjadi

balanced dengan SMOTE, kemudian dilakukan resampling kembali dengan Tomek Links.



Gambar 4. 16 Komposisi Data *Imbalanced* dengan Data *Balanced*

4.3.2 Analisis Regresi Logistik pada Data *Balanced*

Setelah data *balanced* dengan metode *combine sampling*, selanjutnya data akan dianalisis menggunakan metode Regresi Logistik dengan hasil sebagai berikut.

A. Regresi Logistik pada Data *Balanced* dengan Semua Variabel

Seperti sebelumnya, sebelum dianalisis dengan Regresi Logistik terlebih dahulu dilakukan deteksi multikolinearitas pada data dengan menggunakan semua variabel penelitian.

i. Deteksi Multikolinearitas pada data *Balanced*

Seperti dengan analisis Regresi Logistik sebelumnya, akan dilakukan pengecekan multikolinearitas pada data yang telah *balanced* yang ditampilkan pada Tabel 4.12.

Tabel 4. 12 Nilai *VIF* dari Data *Balanced*

Variabel	VIF
X_1	1,014
X_2	1,297
X_3	1,346
X_4	1,153
X_5	1,089
X_6	1,072
X_7	1,019
X_8	1,058

Tabel 4.12 menunjukkan bahwa nilai VIF dari masing-masing variabel prediktor dari data hasil *combine sampling* memiliki nilai yang kurang dari 5. Hal ini mengindikasikan bahwa tidak terjadi kasus multikolinearitas pada data penelitian. Oleh karena itu, tidak diperlukan penanganan lebih lanjut untuk kasus multikolinearitas.

ii. Ketepatan Klasifikasi Regresi Logistik dengan Data *Balanced*

Setelah melakukan pengecekan multikolinearitas pada data, selanjutnya dilakukan pengklasifikasian status ketertinggalan desa di Jawa Timur dari data hasil *combine sampling* menggunakan semua variabel dengan *stratified* 10-*fold cross validation* yang ditampilkan pada Tabel 4.13.

Tabel 4. 13 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Balanced*

Fold	*AUC*		*G-mean*		Akurasi Total		Sensitivitas		Spesifisitas	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
1	0.784	0.754	0.784	0.754	0.786	0.755	0.803	0.765	0.765	0.742
2	0.786	0.751	0.785	0.749	0.787	0.755	0.799	0.800	0.772	0.701
3	0.784	0.793	0.784	0.793	0.785	0.792	0.798	0.791	0.770	0.794
4	0.776	**0.828**	0.775	**0.828**	0.777	**0.830**	0.794	**0.852**	0.757	**0.804**
5	0.779	0.822	0.778	0.822	0.780	0.821	0.799	0.809	0.758	0.835
6	0.783	0.765	0.783	0.764	0.784	0.769	0.796	0.809	0.770	0.722
7	0.783	0.793	0.783	0.793	0.785	0.792	0.798	0.791	0.769	0.794
8	0.787	0.735	0.787	0.735	0.788	0.736	0.800	0.748	0.774	0.722
9	0.785	0.763	0.785	0.763	0.787	0.764	0.800	0.774	0.771	0.753
10	0.778	0.787	0.778	0.786	0.780	0.788	0.796	0.800	0.761	0.773
Mean	0.783	0.779	0.782	0.779	0.784	0.780	0.798	0.794	0.767	0.764
Stdev	0.004	0.031	0.004	0.031	0.004	0.030	0.002	0.028	0.006	0.043

Tabel 4.13 menunjukkan bahwa model terbaik yang didapat pada *fold* ke-4 karena memiliki nilai AUC, G-*mean*, akurasi total, sensitivitas, dan spesifisistas yang cenderung tinggi pada data testing, dimana nilai ini lebih besar dibanding *fold* lainnya. Diketahui bahwa rata-rata AUC sebesar 77,9% pada data testing, sedangkan rata-rata G-*mean* pada data testing adalah sebesar 77,9%. Sedangkan untuk nilai akurasi total pada data testing rata-rata yang didapat tidak jauh berbeda yaitu sebesar 78%. Selanjutnya diperoleh rata-rata sensitivitas sebesar 79,4% pada data testing dan rata-rata spesifisitas sebesar 76,4% pada data

testing. Ketepatan klasifikasi pada data *balanced* nilainya lebih stabil jika dibandingkan dengan data *imbalanced*. Hal ini ditunjukkan dari nilai senstivitas dan spesifisitas tidak jauh berbeda dan cukup tinggi yang menandakan bahwa setelah data *balanced,* ketepatan klasifikasi dapat bertambah. Disamping itu, dapat dilihat bahwa standar deviasi dari masing-masing ketepatan klasifikasi memiliki nilai kecil yang berarti bahwa nilai ketepatan klasifikasi pada masing-masing *fold* tidak jauh berbeda.

Setelah melakukan ketepatan klasifikasi dengan semua variabel pada data *balanced*, diketahui bahwa pada *fold* ke-4 diperoleh nilai G-*mean*, akurasi total, sensitivitas, dan spesifisias yang tinggi jika dibandingkan *fold* lainnya. Oleh karena itu, selanjutnya dilakukan pengujian signifikansi parameter menngunakan *fold* ke-4. Hasil uji serentak menggunakan *Likelihood Ratio Test* diperoleh nilai *G* sebesar 1.899,403. Keputusan yang diperoleh tolak H_0 karena nilai $G > \chi^2_{(8;0,10)}\alpha$ yaitu 1.899,403 >13,36 yang berarti bahwa secara serentak model pada *fold* ke-4 minimal ada 1 variabel yang berpengaruh signifikan. Selanjutnya dilakukan pengujian parsial untuk mengetahui variabel apa saja yang signifikan dengan hasil sebagai berikut.

Tabel 4. 14 Hasil Uji Parsial Pada Regresi Logistik Data *Balanced*

Variabel	**Koefisien**	***Std. Error***	$\mathbf{Z_{Hitung}}$	***P-Value***
Konstan	0,647	0,833	0,777	0,437
X_1	-0,130	0,072	-1,815	0,069
X_2	9,453	2,001	4,725	0,000
X_3	-10,573	2,022	-5,228	0,000
X_4	-1,196	0,843	-1,418	0,156
X_5	-1,079	0,122	-8,840	0,000
X_6	0,285	0,018	15,697	0,000
X_7	-1,117	0,990	-1,128	0,260
X_8	-0,173	0,062	-2,774	0,006

Tabel 4.14 menunjukkan bahwa terdapat 6 variabel yang signifikan karena memiliki p-*value*<(α=0,10) dan variabel yang tidak signifikan adalah rasio banyaknya toko kelontong (X_4) dan rasio banyaknya penderita gizi buruk (X_7). Pada model tersebut masih terdapat beberapa variabel yang tidak signifikan pada

α=0,10. Maka akan dilakukan *backward elimination* untuk memilih variabel signifikan dengan cara mengeluarkan variabel yang paling tidak berpengaruh secara bertahap dengan hasil sebagai berikut.

Tabel 4. 15 Hasil Uji Parsial Pada Regresi Logistik Data *Imbalanced*

Variabel	Koefisien	*Std. Error*	Z_{Hitung}	*P-Value*
Konstan	-0,534	0,171	-3,130	0,002
X_1	-0,138	0,072	-1,929	0,054
X_2	9,392	2,004	4,687	0,000
X_3	-10,622	2,026	-5,244	0,000
X_5	-1,112	0,120	-9,229	0,000
X_6	0,292	0,018	16,292	0,000
X_8	-0,183	0,062	-2,937	0,003

Hasil *backward elimination* pada Tabel 4.15 menunjukkan bahwa variabel yang signifikan sebanyak enam yaitu rasio banyaknya SD/MI (X_1), rasio banyaknya tempat praktik bidan (X_2), rasio banyaknya poskesdes (X_3), rasio banyaknya keluarga pengguna listrik (X_5), jarak tempuh per kilometer ke kantor camat (X_6), dan rasio pendapatan asli desa (X_8).

B. Regresi Logistik dengan Variabel yang Signifikan

Setelah diperoleh variabel yang signifikan, selanjutnya terlebih dahulu dilakukan pengecekan multikolinearitas pada variabel yang signifikan. Berikut adalah hasil uji multikolinearitas yang ditampilkan pada tabel 4.16.

Tabel 4. 16 Nilai *VIF* Data *Imbalanced* Variabel Signifikan

Variabel	VIF
X_1	1,007
X_2	1,291
X_3	1,344
X_5	1,031
X_6	1,023
X_8	1,042

Nilai VIF yang ditampilkan pada Tabel 4.16 untuk variabel yang signifikan, tidak ada yang lebih dari 5. Hal ini mengindikasikan bahwa tidak terjadi kasus multikolinearitas pada data *balanced* yang signifikan.

i. Ketepatan Klasifikasi Regresi Logistik Variabel Signifikan

Data *balanced* yang signifikan akan diklasifikasikan dengan Regresi Logistik menggunakan *stratified* 10-*fold cross validation* yang ditampilkan pada Tabel 4.17.

Tabel 4. 17 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Balanced* Variabel Signifikan

	Training	**Testing**	**Stdv Testing**
Rata-rata AUC	0,777	0,776	0,033
Rata-rata G-*mean*	0,777	0,776	0,033
Rata-rata akurasi total	0,779	0,778	0,032
Rata-rata sensitivitas	0,796	0,797	0,029
Rata-rata spesifisitas	0,758	0,756	0,045

Rata-rata AUC diperoleh sebesar 77,6% pada data testing, sedangkan rata-rata G-*mean* pada Tabel 4.17 diperoleh sebesar 77,6% pada data testing. Sedangkan nilai akurasi total yang diperoleh tidak jauh berbeda yaitu 77,8%. Nilai sensitivitas diperoleh sebesar 79,7% pada data testing dan nilai spesifisitas diperoleh sebesar 75,6% pada data testing. Jika dibandingkan dengan hasil semua variabel, ketepatan klasifikasi yang diperoleh pada data yang signifikan hasilnya lebih rendah daripada data dengan semua variabel. Selanjutnya dapat diketahui pula bahwa standar deviasi dari masing-masing ketepatan klasifikasi cukup rendah yang mengindikasikan dari 10 *fold* tidak memiliki nilai yang jauh berbeda.

4.3.3 Regresi Logistik Ridge pada Data *Balanced*

Data hasil *combine sampling* setelah dianalisis dengan Regresi Logistik, selanjutnya dianalisis dengan Regresi Logistik Ridge sebagai *classifier* pembanding baik pada data dengan menggunakan semua variabel maupun menggunakan variabel signifikan menurut hasil Regresi Logistik.

A. Ketepatan Klasifikasi Regresi Logistik Ridge dengan Semua Variabel

Ketepatan klasifikasikasi pada data *balanced* dengan semua variabel diperoleh hasil sebagai berikut.

Tabel 4. 18 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Balanced* Semua Variabel

	Training	Testing	Stdv Testing
Rata-rata AUC	0,784	0,780	0,033
Rata-rata G-*mean*	0,784	0,779	0,033
Rata-rata akurasi total	0,785	0,781	0,032
Rata-rata sensitivitas	0,799	0,796	0,028
Rata-rata spesifisitas	0,768	0,764	0,045

Berdasarkan Tabel 4.18 diperoleh rata-rata AUC sebesar 78% pada data testing, sedangkan rata-rata G-*mean* sebesar 77,9% pada data testing. Sedangkan nilai akurasi total yang diperoleh tidak jauh berbeda yaitu 78,1%. Nilai sensitivitas diperoleh sebesar 79,6% pada data testing dan nilai spesifisitas diperoleh sebesar 76,4% pada data testing. Jika dilihat dari keseluruhan nilai ketepatan klasifikasi, diperoleh hasil yang tidak jauh berbeda. Disamping itu, dapat diketahui pula bahwa standar deviasi dari masing-masing ketepatan klasifikasi cukup rendah yang mengindikasikan dari 10 *fold* memiliki keragaman yang kecil.

B. Ketepatan klasifikasi Regresi Logistik Ridge Variabel Signifikan

Selanjutnya dilakukan klasifikasi dengan Regresi Logistik Ridge pada data *balanced* yang signifikan menurut metode Regresi Logistik sebagai berikut.

Tabel 4. 19 Hasil Ketepatan Klasifikasi Regresi Logistik dengan Data *Balanced* Variabel Signifikan

	Training	Testing	Stdv Testing
Rata-rata AUC	0,777	0,777	0,032
Rata-rata G-*mean*	0,777	0,777	0,032
Rata-rata akurasi total	0,779	0,779	0,031
Rata-rata sensitivitas	0,796	0,795	0,031
Rata-rata spesifisitas	0,759	0,760	0,043

Tabel 4.19 menunjukkan bahwa diperoleh rata-rata AUC dan G-*mean* yang sama yaitu sebesar 77,7% pada data testing.

Sedangkan nilai akurasi total yang diperoleh tidak jauh berbeda yaitu 77,9%. Nilai sensitivitas diperoleh sebesar 79,5% pada data testing dan nilai spesifisitas diperoleh sebesar 76% pada data testing. Disamping itu, diketahui bahwa standar deviasi dari masing-masing ketepatan klasifikasi cukup rendah yang mengindikasikan dari 10 *fold* memiliki nilai yang tidak jauh berbeda. Apabila dibandingkan metode Regresi Logistik Ridge dengan semua variabel dan variabel yang signifikan, diperoleh ketepatan klasifikasi yang lebih tinggi pada data dengan semua variabel.

4.3.4 Analisis Diskriminan Kernel pada Data *Balanced*

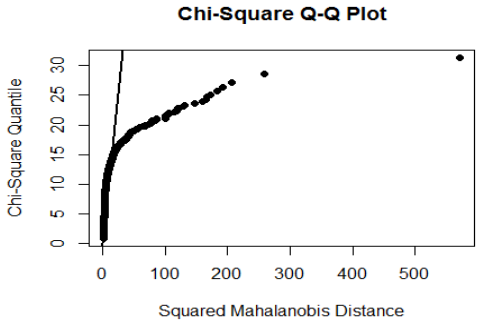
Data hasil dari *combine sampling* selanjutnya dianalisis menggunakan analisis diskriminan baik pada semua variabel maupun variabel yang signifikan.

A. Analisis Diskriminan Kernel dengan Semua Variabel

Sama dengan analisis sebelumnya, terlebih dahulu dilakukan pengujian asumsi pada analisis diskriminan apakah telah memenuhi atau belum. Uji asumsi yang dilakukan adalah uji normal multivariat dan uji homogenitas. Berikut adalah hasil uji asumsi dari data yang memiliki variabel respon *balanced*.

i. Uji Normal Multivariat

Hasil dari uji normal multivariat pada data *balanced* menggunakan semua variabel menghasilkan *p-value* sebesar 0. Oleh karena itu, diperoleh keputusan tolak H_0 karena *p-value* kurang dari $\alpha = 0,05$. Hal ini menunjukkan bahwa data hasil *combine sampling* tidak berdistribusi normal multivariat, sehingga asumsi distribusi normal multivariat tidak terpenuhi. Selain itu dapat dilihat dari hasil QQ-plot dari data penelitian dapat diketahui bahwa banyak sebaran titik hitam tersebar jauh dari garis, hal ini mengindikasikan bahwa data yang telah *balanced* tidak berdistribusi normal multivariat, dimana disajikan dalam Gambar 4.7 berikut.



Gambar 4. 17 *Chi-Squared* QQ-Plot Data Hasil Data *Combine Sampling* Semua Variabel

ii. Uji Homogenitas

Pada data hasil *combine sampling* diperoleh hasil uji homogenitas dengan *p-value* sebesar 0 dan nilai Chi-Square sebesar 2319,8. Karena *p-value* kurang dari $\alpha = 0,05$, maka diperoleh keputusan tolak H_0. Hal ini menunjukkan bahwa matriks varians kovarians pada data *balanced* tidak homogen, sehingga asumsi homogenitas data tidak terpenuhi.

iii. Hasil Ketepatan Klasifikasi dengan Data *Balanced*

Setelah diketahui bahwa asumsi uji multivariat normal dan homogenitas data tidak terpenuhi, maka analisis diskriminan yang digunakan adalah analisis diskriminan kernel. Selanjutnya dapat dilihat nilai ketepatan klasifikasi status ketertinggalan desa di Jawa Timur pada Tabel 4.20 pada data hasil *combine sampling* yang telah *balanced*.

Tabel 4. 20 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel dengan Data *Imbalanced* Semua Variabel

	Training	Testing	Stdv Testing
Rata-rata AUC	0,772	0,770	0,038
Rata-rata G-*mean*	0,770	0,767	0,040
Rata-rata akurasi total	0,777	0,775	0,036
Rata-rata sensitivitas	0,834	0,830	0,026
Rata-rata spesifisitas	0,710	0,710	0,061

Tabel 4.20 menunjukkan bahwa diperoleh rata-rata AUC sebesar 77% pada data testing, sedangkan rata-rata G-*mean* yaitu sebesar 76,7% pada data testing. Kemudian rata-rata akurasi total

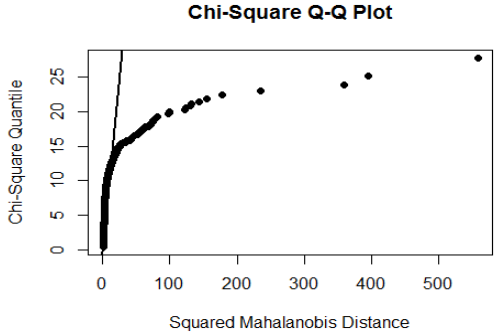
pada data testing diperoleh sebesar 77,5%. Selain itu, rata-rata sensitivitas yang diperoleh cukup besar sebesar 83%. Sedangkan rata-rata spesifisitas diperoleh sebesar 71%. Rata-rata ketepatan klasifikasi yang diperoleh tidak jauh berbeda dan cukup tinggi. Selanjutnya dapat dilihat standar deviasi dari masing-masing ketepatan klasifikasi diperoleh nilai yang kecil, sehingga dapat diindikasikan bahwa masing-masing *fold* memiliki keragaman ketepatan klasifikasi yang kecil.

B. Analisis Diskriminan Kernel dengan Variabel Signifikan

Setelah melakukan analisis pada semua variabel dengan Analisis Diskriminan Kernel, selanjutnya dilakukan analisis pada variabel yang signifikan berdasarkan hasil Regresi Logistik. Terlebih dahulu dilakukan uji asumsi dengan hasil sebagai berikut.

i. Uji Normal Multivariat

Hasil dari uji normal multivariat pada data *balanced* menggunakan variabel signifikan menghasilkan *p-value* sebesar 0. Oleh karena itu, diperoleh keputusan tolak H_0 karena *p-value* kurang dari $\alpha = 0,05$. Hal ini menunjukkan bahwa data *balanced* dengan variabel signifikan tidak berdistribusi normal multivariat,. Berikut ini adalah QQ-plot dari data penelitian yang disajikan dalam Gambar 4.18.



Gambar 4. 18 *Chi-Squared* QQ-Plot Data Hasil Data *Combine Sampling* Variabel Signifikan

Berdasarkan Gambar 4.18 dapat diketahui bahwa banyak sebaran titik hitam tersebar jauh dari garis, hal ini mengindikasikan

bahwa data yang telah *balanced* dengan variabel signifikan tidak berdistribusi normal multivariat.

ii. Uji Homogenitas

Setelah dilakukan uji normal multivariat, asumsi kedua yang akan diuji adalah homogenitas data. Pada data *balanced* menggunakan variabel signifikan diperoleh hasil uji homogenitas dengan *p-value* sebesar 0 dan nilai *Chi-Square* sebesar 1934,2. Karena *p-value* kurang dari $\alpha = 0,05$, maka diperoleh keputusan tolak H_0. Hal ini menunjukkan bahwa matriks varians kovarians pada data *balanced* dengan variabel signifikan tidak homogen, sehingga asumsi homogenitas data tidak terpenuhi.

iii. Hasil Ketepatan Klasifikasi dengan Data *Balanced*

Pada data *balanced* dengan variabel signifikan diketahui bahwa asumsi multivariat normal dan homogenitas tidak terpenuhi, sehingga digunakan Analisis Diskriminan Kernel. Berikut adalah ketepatan klasifikasi yang diperoleh.

Tabel 4. 21 Hasil Ketepatan Klasifikasi Analisis Diskriminan Kernel dengan Data *Imbalanced* Variabel Signifikan

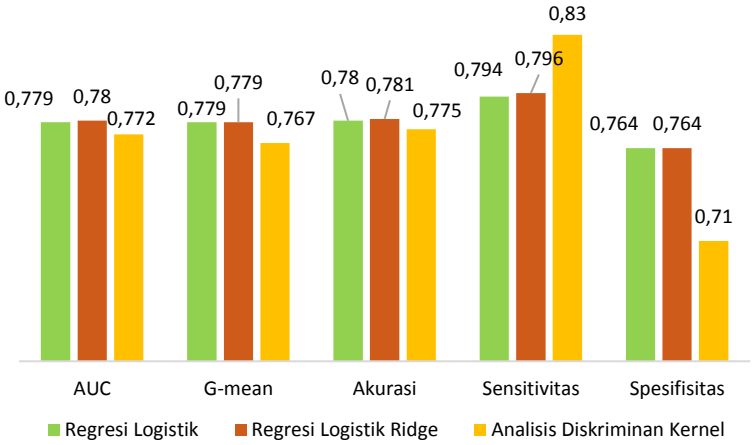
	Training	**Testing**	**Stdv Testing**
Rata-rata AUC	0,772	0,769	0,038
Rata-rata G-*mean*	0,769	0,766	0,040
Rata-rata akurasi total	0,777	0,774	0,037
Rata-rata sensitivitas	0,833	0,829	0,026
Rata-rata spesifisitas	0,710	0,709	0,062

Dapat diketahui pada Tabel 4.21, diperoleh rata-rata AUC pada data testing sebesar 76,9%, sedangkan diperoleh rata-rata G-*mean* yaitu sebesar 76,6% pada data testing. Kemudian rata-rata akurasi total pada data testing diperoleh sebesar 77,4%. Selain itu, rata-rata sensitivitas yang diperoleh cukup besar sebesar 82,9%. Sedangkan rata-rata spesifisitas diperoleh sebesar 70,9%. Rata-rata ketepatan klasifikasi yang diperoleh tidak jauh berbeda dan cukup tinggi. Tetapi jika dibandingkan dengan Analisis Diskriminan Kernel menggunakan semua variabel dan variabel signifikan, diketahui bahwa hasil semua variabel lebih tinggi dibandingkan dengan variabel yang signifikan walaupun selisishnya tidak

banyak. Selanjutnya dapat dilihat standar deviasi dari masing-masing ketepatan klasifikasi diperoleh nilai yang kecil, sehingga dapat diindikasikan bahwa masing-masing *fold* memiliki nilai yang tidak jauh berbeda.

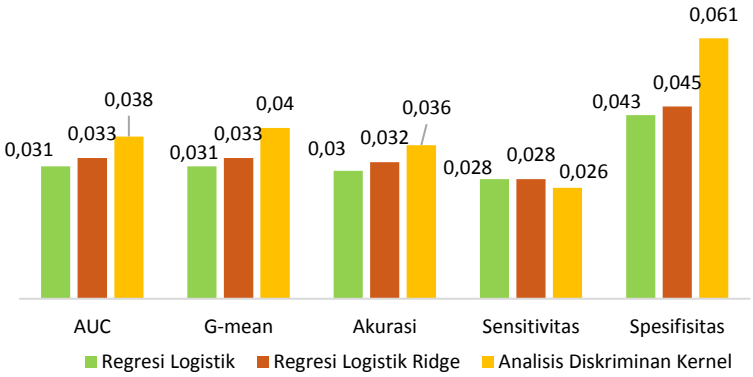
4.3.5 Analisis Gabungan Data *Balanced* Semua Variabel dan Variabel Signifikan

Pada data yang telah *balanced* dengan metode *combine sampling*, 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 denagn semua variabel.



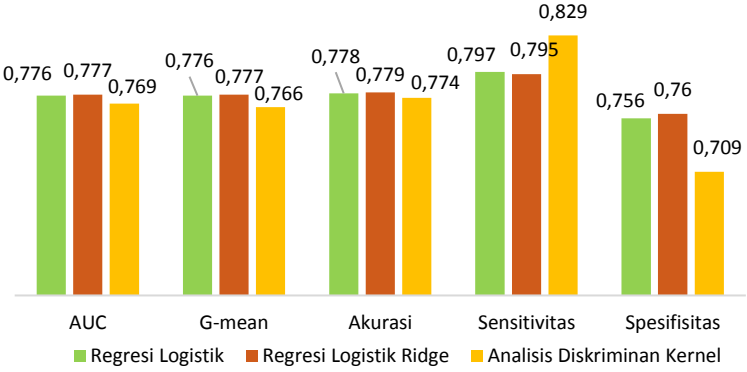
Gambar 4. 19 Perbandingan Ketepatan Klasifikasi Data *Balanced* Semua Variabel

Berdasarkan Tabel 4.19 dapat diketahui bahwa masing-masing ketepatan klasifikasi memiliki nilai yang tidak jauh berbeda. Pada Regresi Logistik Ridge memiliki nilai G-mean yang sama serta nilai AUC dan akurasi total memiliki selisih yang kecil. Tetapi jika dibandingkan dari ketiga metode, Regresi Logistik Ridge lebih unggul pada nilai AUC, akurasi total, dan spesifisitas. Sehingga dengan semua variabel pada data *balanced* baik diklasifikasikan dengan Regresi Logistik Ridge.



Gambar 4. 20 Perbandingan Standar Deviasi Data *Balanced* Semua Variabel

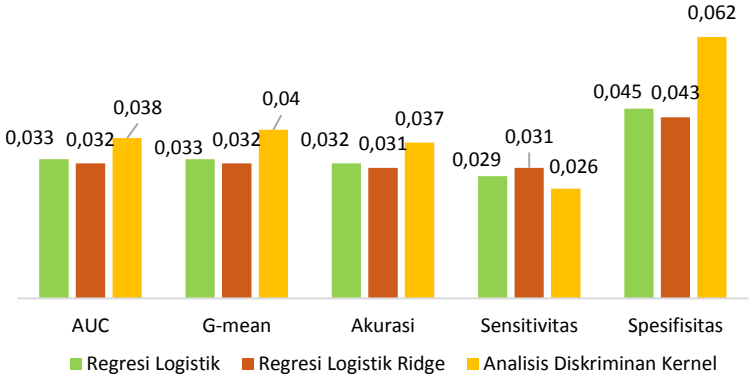
Pada Gambar 4.20 dapat diketahui perbandingan standar deviasi menggunakan semua variabel, dimana Analisis Diskriminan Kernel memiliki standar deviasi yang lebih tinggi daripada metode lainnya, walaupun selisih standar deviasi dari masing-masing metode tidak terlalu besar. Sedangkan dengan menggunakan variabel yang signifikan diperoleh hasil sebagai berikut.



Gambar 4. 21 Perbandingan Ketepatan Klasifikasi Data *Balanced* Variabel Signifikan

Tidak jauh berbeda dengan hasil semua variabel, Gambar 4.21 menunjukkan bahwa dari ketiga metode tersebut dengan

menggunakan variabel yang signifikan, Regresi Logistik Ridge memiliki hasil yang tinggi pada nilai AUC, G-*mean*, akurasi total, dan spesifisitas.



Gambar 4. 22 Perbandingan Standar Deviasi Data *Balanced* Variabel Signifikan

Hasil perbandingan standar deviasi pada variabel signifikan memiliki hasil yang sama dengan semua variabel. Dimana Analisis Diskriminan Kernel memiliki standar deviasi paling tinggi dibandingkan Regresi Logistik maupun Regresi Logistik Ridge.

4.4 Efektivitas Metode *Combine Sampling*

Setelah sebelumnya dilakukan analisis pada data status ketertinggalan desa di Jawa Timur berdasarkan 5 kabupaten yng memiliki persentase desa tertinggal tertinggi dengan metode Regresi Logistik, Regresi Logistik Ridge dan Analisis diskriminan Kernel. Kemudian dilakukan pula resampling data dengan metode *combine sampling* yang kemudian dianalisis dengan ketiga metode tersebut. Selanjutnya akan dibandingkan ketepatan klasifikasi dari masing-masing metode baik pada semua variabel maupun variabel yang signifikan.

4.4.1 Efektivitas *Combine Sampling* pada Semua Variabel

Berikut adalah perbandingan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan

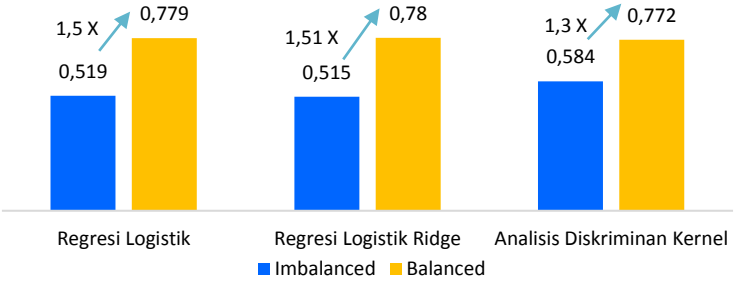
semua variabel pada data testing baik dengan data *imbalanced* maupun *balanced.*

Tabel 4. 22 Evaluasi Ketepatan Klasifikasi dengan Semua Variabel

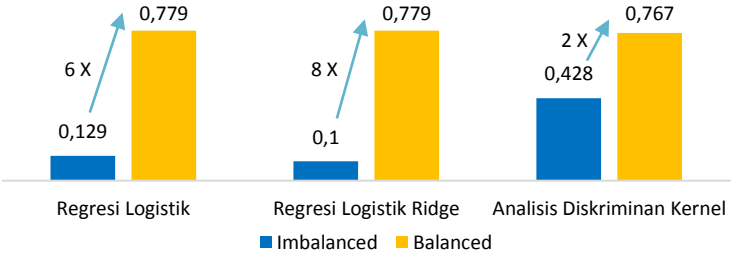
		AUC	G-*mean*	Akurasi	Sens	Stdv AUC	Stdv G-*mean*	Stdv Sens
Regresi Logistik	1	0,519	0,129	0,897	0,043	0.029	0.171	0.063
	2	0,779	0,779	0,780	0,794	0.031	0.031	0.028
Regresi Logistik Ridge	1	0,515	0,100	0,898	0,035	0.028	0.165	0.062
	2	0,780	0,779	0,781	0,796	0.033	0.033	0.028
Analisis Diskriminan Kernel	1	0,584	0,428	0,890	0,199	0.044	0.104	0.09
	2	0,772	0,767	0,775	0,830	0.038	0.04	0.026

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

Tabel 4.22 menunjukkan bahwa akurasi total data yang telah *balanced* setelah dilakukan resampling dengan *combine sampling* memiliki tingkat akurasi yang lebih tinggi dari data *imbalanced* baik dalam metode Regresi Logistik, Regresi Logistik Ridge maupun Analisis Diskriminan Kernel. Meskipun terdapat nilai akurasi total yang lebih tinggi pada data *imbalanced*, hal ini menunjukkan bahwa akurasi total kurang baik mengklasifikasikan kelas data yang tidak seimbang yang dapat dilihat dari nilai G-*mean* yang sangat kecil serta perbedaan nilai sensitivitas serta spesifisitas yang tinggi pada data *imbalanced*. Sehingga dapat diasumsikan bahwa dengan menyeimbangkan kelas data, maka tingkat ketepatan klasifikasi sebuah model dapat meningkat. Selain itu dari masing-masing metode nilai AUC, akurasi total dan G-*mean* memiliki nilai yang tidak jauh berbeda pada data *balanced*. Pada data *imbalanced*, dengan menggunakan ketiga metode tersebut Analisis Diskriminan kernel memiliki performa yang lebih baik dibanding lainnya. Berdasarkan nilai AUC, G-*mean* dan akurasi total pada data yang menggunakan semua variabel, metode Regresi Logistik Ridge baik untuk mengklasifikasikan status ketertinggalan desa di Jawa Timur pada data *balanced*. Berikut adalah visualisasi perbandingan dari masing-masing metode.



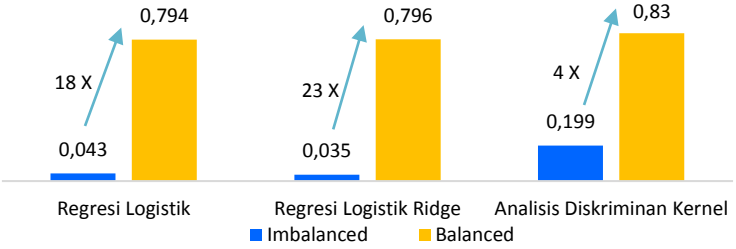
(a)



(b)

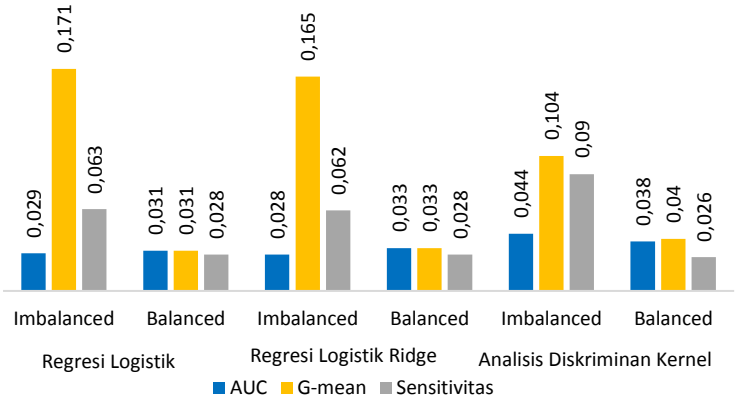
Gambar 4. 23 Perbandingan (a) AUC dan (b) G-*mean* dengan Semua Variabel

Gambar 4.22 menunjukkan bahwa terjadi perubahan yang cukup tinggi pada nilai AUC dan G-*mean* dari data *imbalanced* ke data *balanced*. Dimana kenaikan tertinggi baik AUC maupun G-mean terdapat pada Regresi Logistik Ridge yaitu dari 0,515 menjadi 0,78 dan 0,1 menjadi 0,779. Pada urutan kedua adalah metode Regresi Logistik dan urutan terakhir adalah Analisis Diskriminan Kernel. Kemudian dapat dilihat pula perubahan nilai sensitivitas dari data *imbalanced* ke data *balanced* sebagai berikut.



Gambar 4. 24 Perbandingan Sensitivitas dengan Semua Variabel

Dapat dilihat dari Gambar 4.23 bahwa rata-rata sensitivitas dari data *imbalanced* ke data *balanced* mengalami kenaikan yang signifikan. Dimana kenaikan tertinggi pada metode Regresi Logistik Ridge yang meningkat hingga 23 kali yaitu dari 0,035 menjadi 0,796 setelah data *balanced*. Kemudian diurutan kedua adalah Regresi Logistik dengan kenaikan 18 kali dan terakhir adalah Analisis Diskriminan Kernel yang meningkat 4 kali.



Gambar 4. 25 Perbandingan Standar Deviasi dengan Semua Variabel

Perbandingan standar deviasi pada Gambar 4.24, diketahui bahwa setelah data *balanced* diperoleh standar deviasi yang cenderung lebih kecil daripada data *imbalanced* untuk G-*mean* dan sensitivitas. Sedangkan standar deviasi pada AUC cenderung turun. Apabila dilihat pada data *balanced* diperoleh standar deviasi yang lebih stabil dibandingkan dengan data *imbalanced*. Tetapi secara keseluruhan standar deviasi yang diperoleh nilainya kecil yang menunjukkan keberagaman dari masing-masing *fold* kecil atau memiliki nilai yang tidak jauh berbeda.

4.4.2 Efektivitas *Combine Sampling* pada Variabel Signifikan

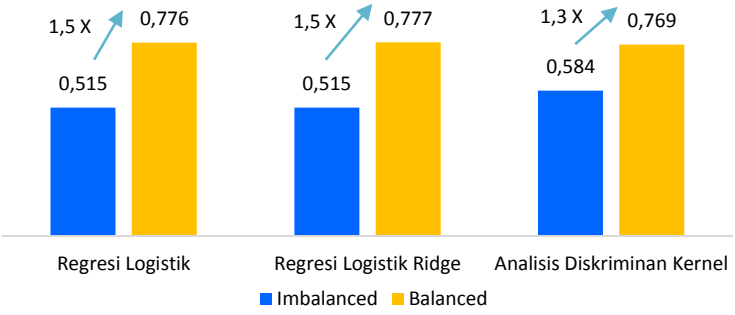
Setelah dilakukan perbandingan metode dengan semua variabel, selanjutnya dilakukan perbandingan ketiga metode dengan variabel signifikan baik pada data *balanced* maupun *imbalanced*.

Tabel 4. 23 Evaluasi Ketepatan Klasifikasi dengan Variabel Signifikan

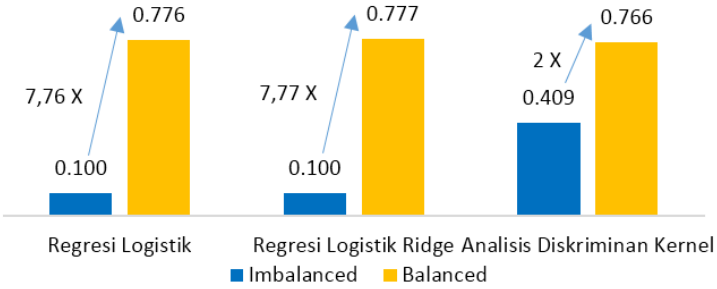
		AUC	G-*mean*	Akurasi	Sens	Stdv AUC	Stdv G-*mean*	Stdv Sens
Regresi Logistik	1	0,515	0,100	0,897	0,035	0,029	0,156	0,059
	2	0,776	0,776	0,778	0,797	0,033	0,033	0,029
Regresi Logistik Ridge	1	0,515	0,100	0,898	0,035	0,028	0,156	0,059
	2	0,777	0,777	0,779	0,795	0,032	0,032	0,031
Analisis Diskriminan Kernel	1	0,584	0,409	0,892	0,197	0,053	0,165	0,102
	2	0,769	0,766	0,774	0,829	0,038	0,04	0,026

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

Tidak jauh berbeda dengan hasil ketepatan klasifikasi menggunakan semua variabel, dari Tabel 2.23 dapat diketahui hasil data *balanced* menggunakan variabel signifikan memiliki ketepatan klasifikasi yang lebih tinggi dan cenderung stabil dari data *imbalanced*. Selain itu pada data *imbalanced*, metode Analisis Diskriminan Kernel memiliki performansi ketepatan klasifikasi yang lebih baik dibandingkan Regresi Logistik dan Regresi logistik Ridge. Oleh karena itu, dengan menggunakan variabel signifikan data status ketertinggalan desa baik diklasifikasikan dengan Regresi Logistik Ridge karena memiliki nilai ketepatan klasifikasi paling baik diantara metode lainnya pada data *balanced*. Kemudian dapat dilihat pula visualisasi dari ketepatan klasifikasi ketiga metode dengan hasil sebagai berikut.



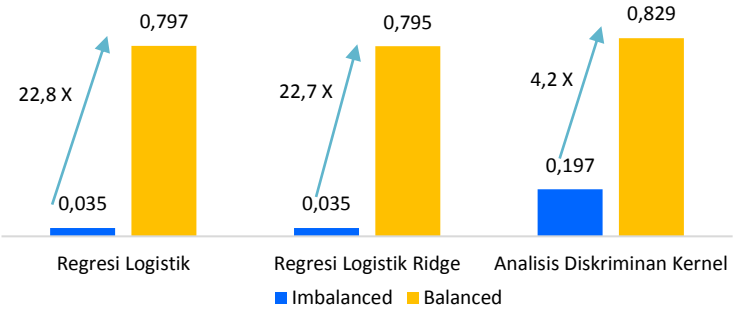
(a)



(b)

Gambar 4. 26 Perbandingan (a) AUC dan (b) G-mean dengan Variabel Signifikan

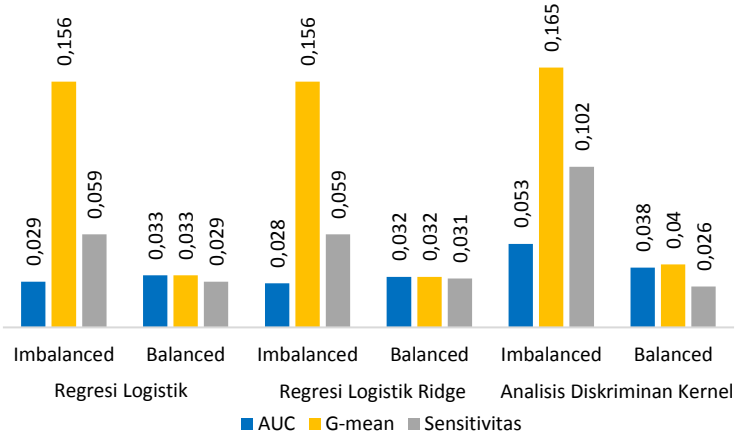
Pada perbandingan nilai AUC dan G-*mean* dapat diketahui bahwa terjadi kenaikan yang cukup signifikan dari data *imbalanced* ke data *balanced*. Berdasarkan Gambar 4.26 dapat diketahui bahwa Regresi Logistik Ridge memiliki kenaikan tertinggi baik pada AUC maupun G-*mean* yaitu dari 0,515 menjadi 0,777 dan 0,100 menjadi 0,777. Kemudian diurutan kedua adalah metode Regresi Logistik, sedangkan metode Analisis Diskriminan Kernel mengalami peningkatan yang terendah. Kemudian dilihat pula kenaikan sensitivitas pada data *imbalanced* ke data *balanced*.



Gambar 4. 27 Perbandingan Sensitivitas dengan Variabel Signifikan

Perbandingan sensitivitas pada Gambar 4.27 menunjukkan bahwa terjadi kenaikan yang signifikan dari data im*balanced* ke data *balanced*. Dapat diketahui pula bahwa dengan menggunakan variabel yang signifikan, kenaikan sensivitas tertinggi berada pada

metode Regresi Logistik yaitu dari 0,035 menjadi 0,797. Pada urutan kedua terdapat Regresi Logistik Ridge dan urtan yang memiliki kenaikan sensitivitas terendah adalah Analisis Diskriminan Kernel.



Gambar 4. 28 Perbandingan Standar Deviasi dengan Variabel Signifikan

Gambar 4.20 menunjukkan bahwa standar deviasi dari data *imbalanced* ke data *balanced* pada AUC cenderung mengalami kenaikan, sedangkan G-*mean* dan sensitivitas mengalami penurunan pada ketiga metode. Jika dibandingkan, nilai standar deviasi dari data *balanced* lebih stabil dibandingkan data *imbalanced* karena memiliki nilai yang tidak jauh berbeda.

(*Halaman ini sengaja dikosongkan*)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan, diperoleh kesimpulan sebagai berikut.

1. Berdasarkan statistika deskriptif dari masing-masing variabel prediktor pada data 5 kabupaten yang memiliki persentase desa tertinggal tertinggi, kebanyakan desa yang memiliki nilai tertinggi pada masing-masing variabel didominasi oleh desa dari Kabupaten Sumenep. Selain itu, masih banyak variabel yang memiliki nilai terendah 0 dimana hal ini mengindikasikan bahwa masih banyak desa yang belum memiliki tempat praktik bidan, poskesdes, toko kelontong, penderita gizi buruk, dan pendapatan asli desa. Apabila dilihat dari nilai varians, variabel jumlah toko kelontong , jumlah keluarga pengguna listrik, jarak ke kantor Camat, dan jumlah pendapatan asli desa memiliki varians cukup tinggi yang menunjukkan bahwa tingkat keberagaman tiap desa cukup tinggi, dimana beberapa desa masih memiliki ketimpangan yang besar antar satu desa dengan desa yang lain. Selain itu, dari data rasio dapat diketahui bahwa dari masing-masing variabel memiliki perbedaan median antara desa tertinggal dan desa tidak tertinggal, tetapi median tertinggi didominasi pada desa tidak tertinggal.
2. Klasifikasi pada data *imbalanced* diperoleh hasil bahwa ketepatan klasifikasi dengan menggunakan semua variabel lebih tinggi dibandingkan dengan variabel signifikan walaupun selisihnya tidak jauh berbeda. Ketiga metode klasifikasi tersebut juga memiliki ketepatan klasifikasi yang hampir sama. Tetapi nilai ketepatan klasifikasi pada Analisis Diskriminan Kernel dengan menggunakan semua variabel memiliki nilai yang lebih tinggi. Dimana diperoleh AUC, G-mean, akurasi total, sensitivitas dan spesifisitas secara berurutan adalah 58,4%, 42,8%, 89%, 19,9% dan 96,9%.

3. Pada data hasil *combine sampling*, dengan menggunakan semua variabel memiliki hasil yang lebih tinggi dibandingkan dengan variabel signifikan. Ketiga metode tersebut memiliki ketepatan klasifikasi yang tidak jauh berbeda, tetapi metode Regresi Logistik Ridge memiliki nilai ketepatan klasifikasi yang lebih tinggi dibandingkan metode lainnya dengan menggunakan data semua variabel. Dimana diperoleh AUC, G-mean, akurasi total, sensitivitas dan spesifisitas secara berurutan adalah 78%, 77,9%, 78,1%, 79,6%, dan 46,4% sehingga Regresi Logistik Ridge baik untuk menjadi *classifier* data status ketertinggalan desa di Jawa Timur.
4. Penerapan metode *Combine Sampling* mampu meningkatkan ketepatan klasifikasi (AUC, G-mean, dan sensitivitas) dari data *imbalanced* ke data *balanced* baik pada data semua variabel maupun data signifikan. Peningkatan ketepatan klasifikasi yang tertinggi terdapat pada Regresi Logistik dan Regresi Logistik Ridge dengan menggunakan semua variabel. Kemudian peningkatan kriteria ketepatan klasifikasi terbesar terdapat pada nilai sensitivitas dengan peningkatan mencapai sebesar 23 kali.

5.2 Saran

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

1. Menerapkan metode *Combine Sampling* pada kasus *multiclass* dan menggunakan data campuran (data numerik dan data kategorik).
2. Menggunakan metode *balancing* data dan metode klasifikasi yang lainnya sebagai pembanding untuk mendapatkan ketepatan klasifikasi yang lebih baik.

DAFTAR PUSTAKA

Agresti, A. (2002). *Categorical Data Analysis Second Edition.* Florida: John Wiley & Sons, Inc.

Agusta, I. (2007). Desa Tertinggal di Indonesia. *Jurnal Transdisiplin Sosiologi, Komunikasi, dan Ekologi Manusia*, 233-252.

Azkiya, M., Mukid, M., & Ispriyanti, D. (2015). Klasifikasi Nasabah Kredit Bank "X" di Provinsi Lampung Menggunakan Analisis Diskriminan Kernel. *Jurnal Gaussian, 4*, 937-946.

Bappenas, & Badan Pusat Statistik . (2015). *Indeks Pembangunan Desa 2014: Tantangan Pemenuhan Standar Pelayanan Minimum Desa.* Jakarta: Badan Pusat Statistik.

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).

Barandela, R., Sanchez, J. S., Garcia, V., & Rangel, E. (2003). Strategies for Learning in Class Imbalance Problem. *Pattern Reognition*, 849-851.

Batista, G., Bazzan, A., & Monard, M. (2003). *Balancing Training data for Automated Annotation of Keywords : a Case Study.*

Batista, G., Prati, R. C., & Monard, M. C. (2004). A Study of The Behavior of Several Methods For Balancing Machine Learning Training Data. *SIGKDD Explorations Newsletter 6*, 20-29.

BPS. (2015). *Indeks Pembangunan Desa 2014.* Jakarta: Badan Pusat Statistik.

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

Choi, J. M. (2010). A Selective Sampling Method for Imabalnced Data Learning on Support Vector Machine. *Graduate Theses and Dissertations*.

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 Edition.* Canada: John Wiley & Sons.

Gaudio, R. D., Batista, G., & Branco, A. (2013). Coping With Highly Imbalanced Dataset: A Case Study With Definition Extraction In A Multilingual Setting. *Natural Languange Engineering*, 1-33.

Hair, J., Anderson, R., Babin, B. J., & Black, W. C. (2006). *Multivariate Data Analysis.* USA: Prentice Hall.

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.

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression : Biased Estimation For Nonorthogonal Problems. *Technometrics, 12*(1).

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

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

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

Kemendesa. (2016). Dipetik Januari 4, 2018, dari Direktorat Jenderal Pembangunan Daerah Tertinggal: http://ditjenpdt.kemendesa.go.id

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

Khaulasari, H., Purnami, S. W., & Prastyo, D. D. (2016). *Combine Sampling - Least Square Support Vector Machine Untuk Klasifikasi Multi Class Imbalanced Data.* Surabaya: Tesis, Statistika FMIPA-ITS.

Li, Y., Gong, S., & Liddell, H. (2001). Kernel Discriminant Analysis.

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

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

Masruroh, I. (2011). *Pemilihan Model Regresi Linier Berganda Pada Kasus Multikolinieritas Dengan Metode Regresi Komponen Utama (Principal Component Regression) Dan Regresi Gulud (Ridge Regression).* FMIPA-Universitas Brawijaya.

Maumere, D., & Ratnasari, V. (2015). *Permodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur Dengan Menggunakan Metode Regresi Logistik Ridge.* Surabaya: Statistika FMIPA-ITS.

Midi, H., Sarkar, S., & Rana, S. (2010). Collinearity Diagnostics of Binary Logistic Regression Model. *Journal of Interdisciplinary Mathematics*, pp. 253-267.

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: IlmuKomputer.com.

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

Rossi, R. J. (2009). *Applied Biostatistics for the Health Sciences.* United State of America: John Wiley & Sons, Inc.

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

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

Sambodo, H. P., Purnami, S., & Rahayu, S. (2013). *Ketepatan Klasifikasi Status Ketertinggalan Desa Dengan Pendekatan Reduce Support Vector Machine (RSVM) di Provinsi Jawa Timur*. Surabaya: Tesis, Statistika FMIPA-ITS.

Sharma, S. (1996). *Applied Multivariate Technique.* United States: 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.

Sulasih, D., & Purnami, S. (2016). *Rare Event Weighted Logistic Regression Untuk Klasifikasi Imbalanced Data.* Surabaya: Thesis, Statistika FMIPA-ITS.

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

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

Wahyuningtias, Y., & Otok, B. (2012). *Evaluasi Ketepatan Klasifikasi Kelulusan Tes Keterampilan Seleksi Nasional Masuk Perguruan Tinggi Bidang Olahraga dengan Analisis Diskriminan Kernel.* Surabaya: Statistika FMIPA-ITS.

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

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.

(*Halaman ini sengaja dikosongkan*)

LAMPIRAN

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

Desa	X_1	X_2	X_3	⋯	X_8	Status
1	0.83	0.000	0.02	⋯	1.073	0
2	1.096	0.022	0.022	⋯	1.172	0
3	0.574	0.025	0.025	⋯	1.043	0
4	0.164	0.035	0.035	⋯	1.701	0
5	1.176	0.034	0.034	⋯	1.56	0
6	0.836	0.03	0.03	⋯	0.506	0
7	0.833	0.000	0.032	⋯	0.600	0
8	1.087	0.016	0.016	⋯	1.749	0
9	0.606	0.00	0.029	⋯	0.732	0
10	0.515	0.034	0.034	⋯	2.172	0
11	0.432	0.027	0.054	⋯	0.642	0
12	0.885	0.029	0.029	⋯	0.087	0
⋮	⋮	⋮	⋮	⋱	⋮	⋮
1111	0.759	0.022	0.022	⋯	0.065	0
1112	1.116	0.027	0.027	⋯	0.081	0
1113	0.862	0.137	0.137	⋯	0.412	1
1114	0.935	0.144	0.000	⋯	0.432	1
1115	1.015	0.000	0.000	⋯	0.323	1
1116	0.734	0.031	0.031	⋯	0.092	0
1117	0.907	0.034	0.000	⋯	0.101	0
1118	0.803	0.081	0.000	⋯	0.244	1
1119	0.953	0.019	0.029	⋯	0.242	0
1120	1.115	0.000	0.012	⋯	0.294	0
1121	1.367	0.025	0.000	⋯	0.627	0
1122	1.333	0.000	0.000	⋯	1.094	1

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

Desa	X_1	X_2	X_3	⋯	X_8	**Status**
1	1.008	0.000	0.001	⋯	0.313	1
2	0.592	0.011	0.021	⋯	0.581	0
3	1.440	0.017	0.033	⋯	0.250	0
4	0.661	0.112	0.112	⋯	3.470	1
5	1.604	0.000	0.049	⋯	0.730	0
6	0.791	0.007	0.043	⋯	0.033	1
7	1.311	0.057	0.000	⋯	0.323	1
8	0.334	0.035	0.070	⋯	0.697	0
9	0.581	0.000	0.194	⋯	0.583	0
10	0.639	0.010	0.015	⋯	0.553	1
11	1.527	0.025	0.050	⋯	0.523	0
12	7.010	0.019	0.033	⋯	0.358	1
⋮	⋮	⋮	⋮	⋱	⋮	⋮
2109	0.589	0.000	0.014	⋯	0.043	0
2110	0.870	0.048	0.096	⋯	1.780	0
2111	0.651	0.000	0.000	⋯	0.133	1
2112	0.773	0.007	0.059	⋯	0.118	0
2113	1.501	0.000	0.061	⋯	0.433	1
2114	0.873	0.000	0.034	⋯	0.401	0
2115	0.974	0.009	0.027	⋯	0.027	0
2116	1.320	0.022	0.044	⋯	0.111	0
2117	1.333	0.000	0.095	⋯	0.473	0
2118	0.817	0.000	0.025	⋯	0.256	1
2119	1.511	0.060	0.003	⋯	2.261	1
2120	1.740	0.000	0.011	⋯	0.407	1

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

a. Data Imbalanced dengan Semua Variabel

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data1[, -9])
Chi-Sq (approx.) = 469.55, df = 36, p-value < 2.2e-16
```

b. Data Imbalanced dengan Variabel Signifikan

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data1[, -4])
Chi-Sq (approx.) = 366.56, df = 6, p-value < 2.2e-16
```

c. Data Balanced dengan Semua Variabel

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data[, -9])
Chi-Sq (approx.) = 2319.8, df = 36, p-value < 2.2e-16
```

d. Data Balanced dengan Variabel Signifikan

```
Box's M-test for Homogeneity of Covariance Matrices
data:  data.matrix(data1[, -7])
Chi-Sq (approx.) = 1934.2, df = 21, p-value < 2.2e-16
```

Lampiran 4. Hasil Uji Distribusi Normal Multivariat

a. Data Imbalanced dengan Semua Variabel

```
$multivariateNormality
             Test         Statistic p value Result
1 Mardia Skewness  91880.677817661       0     NO
2 Mardia Kurtosis 840.022534376288       0     NO
3             MVN              <NA>    <NA>     NO
```

b. Data Imbalanced dengan Variabel Signifikan

```
             Test         Statistic p value Result
1 Mardia Skewness 52967.4556953527       0     NO
2 Mardia Kurtosis 913.934769139669       0     NO
3             MVN              <NA>    <NA>     NO
```

c. Data Balanced dengan Semua Variabel

```
$multivariateNormality
             Test         Statistic p value Result
1 Mardia Skewness 87983.8260100704       0     NO
2 Mardia Kurtosis 731.426412745582       0     NO
3             MVN              <NA>    <NA>     NO
```

d. Data Balanced dengan Variabel Signifikan

```
$multivariateNormality
             Test         Statistic p value Result
1 Mardia Skewness 92296.9499249139       0     NO
2 Mardia Kurtosis 894.076375477098       0     NO
3             MVN              <NA>    <NA>     NO
```

Lampiran 5. Hasil Uji Signifikansi Parameter Regresi Logistik Ridge

a. Hasil Data *Imbalanced*

```
logisticRidge(formula = Status ~ ., data = train, family = binom
ial(link = "logit"))
Coefficients:
```

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	1.281	NA	NA	NA	NA	
x1	0.049	1.451	2.177	0.666	0.505	
x2	-2.602	-3.576	2.321	-1.541	0.123	
x3	1.501	1.959	2.316	0.846	0.398	
X4	-0.566	-12.999	2.362	-5.504	0.000	***
X5	-3.234	-7.524	1.901	-3.957	0.000	***
X6	0.024	7.059	2.040	3.460	0.001	***
X7	0.212	0.521	2.230	0.234	0.815	
X8	-0.089	-3.222	2.370	-1.360	0.174	

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Ridge paramter: 0.02205763, chosen automatically, computed using
4 PCs
```

b. Hasil Data *Balanced*

```
logisticRidge(formula = Status ~ ., data = train, family = binom
ial(link = "logit"))
Coefficients:
```

	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)	
(Intercept)	-0.476	NA	NA	NA	NA	
x1	-0.13	-5.29	2.80	-1.89	0.06	.
x2	9.01	16.03	3.47	4.62	0.00	***
x3	-10.36	-17.30	3.30	-5.25	0.00	***
x5	-1.10	-31.03	3.33	-9.33	0.00	***
x6	0.28	100.75	6.05	16.65	0.00	***
x8	-0.18	-7.58	2.56	-2.96	0.00	**

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Ridge paramter: 0.0005251979, chosen automatically, computed usi
ng 5 PCs
```

Lampiran 6. *Syntax* Uji Asumsi Analisis Diskriminan

a. Syntax Uji Homogenitas

```
library(biotools)
data=read.csv("E:/KULIAH/TA/Data/Hasil/data5.csv",header=T,se
p=";")
head(data)
Uji_Homogen<-boxM(data.matrix(data)[,-9], data[,9])
Uji_Homogen
```

b. *Syntax* Uji Normal Multivariat

```
library(MVN)
data=read.csv("E:/KULIAH/TA/Data/Hasil/data5.csv",header=T,se
p=";")
uji_multinorm<-mvn(data1, 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_multinorm
```

Lampiran 7. *Syntax Combine Sampling*

```
#SMOTE
data<-read.csv("E:/KULIAH/TA/Data/Data 5 kab/data5.csv",
header=TRUE, sep=";")
library(unbalanced)
set.seed(12345)
Y<-as.factor(data$Status)
X<-data[,-9]
databaru<-ubSMOTE(X,Y, perc.over = 900, k=100, perc.under =
100, verbose=TRUE)
newdata<-cbind(databaru$X, databaru$Y)
write.csv(newdata,"E:/KULIAH/TA/Data/Data 5
kab/SMOTEkab.csv")

#Tomek Links
data1=read.csv("E:/KULIAH/TA/Data/Data 5
kab/SMOTE5kab.csv", header=TRUE, sep=";")
Y<-as.factor(data1$Status)
X<-data1[,-9]
databaru<-ubTomek(X,Y,  verbose=TRUE)
newdata<-cbind(databaru$X, databaru$Y)
write.csv(newdata,"E:/KULIAH/TA/Data/Data 5 kab/ST.csv")
```

Lampiran 8. *Syntax* Regresi Logistik

```
data=read.csv("E:/KULIAH/TA/Data/Hasil/data5.csv",header=T,se
p=";")
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)
mean(GmeanTrain)
mean(GmeanTest)
#export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                      TotalAccuracyTest,
```

Lampiran 8. *Syntax* Regresi Logistik (Lanjutan)

```
                SensTrain,
                SensTest,
                SpesTrain,
                SpesTest,
                G-mean-Train,
                G-mean-Test)
hasil-mean-=data.frame(-mean-(TotalAccuracyTrain),
                -mean-(TotalAccuracyTest),
                -mean-(SensTrain),
                -mean-(SensTest),
                -mean-(SpesTrain),
                -mean-(SpesTest),
                -mean-(G-mean-Train),
                -mean-(G-mean-Test))
hasil-mean-
write.csv(hasiltotal, file = "E:/KULIAH/TA/Data/Hasil/
Reglog_hasil_total.csv")
write.csv(hasil-mean-,file="E:/KULIAH/TA/Data/Hasil/
Reglog_hasil_-mean-.csv")
```

Lampiran 9. *Syntax* Regresi Logistik Ridge

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

#COBA
data<-read.csv("E:/KULIAH/TA/Data/Data 5
kab/data5sig.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)
GmeanTrain=rep(0,r)
GmeanTest=rep(0,r)

#MODEL ANDISKER
for(i in 1:r)
{
  train=data[-fold[[1]],]
  test=data[fold[[1]],]
  model = logisticRidge(Status~., data=as.data.frame(train),
lambda ="automatic" )
```

Lampiran 9. *Syntax* Regresi Logistik Ridge (Lanjutan)

```
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))
  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,
```

Lampiran 9. *Syntax* Regresi Logistik Ridge (Lanjutan)

```
                SpesTrain,
                SpesTest,
                GmeanTrain,
                GmeanTest)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                mean(TotalAccuracyTest),
                mean(SensTrain),
                mean(SensTest),
                mean(SpesTrain),
                mean(SpesTest),
                 mean(GmeanTrain),
                mean(GmeanTest))
hasilmean
write.csv(hasiltotal, file = "E:/KULIAH/TA/Data/Data 5
kab/data5sig_Reglogridge_total.csv")
write.csv(hasilmean,file="E:/KULIAH/TA/Data/Data 5
kab/data5sig_Reglogridge_total_mean.csv")
summary(model)
```

Lampiran 10. *Syntax* Analisis Diskriminan Kernel

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

data<-read.csv("E:/KULIAH/TA/Data/Hasil/Asli_andisker.csv",
header=TRUE, sep=";")
Y<-as.factor(data$Status)
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[[i]],]
  test=data[fold[[i]],]
  model = kfda(trainData=train,kernel.name="rbfdot")
  predtrain1=kfda.predict(model, train)
  predtrain2=predtrain1$class
  predtrain3=as.vector(predtrain2)
  predtrain=as.numeric(predtrain3)

  predtest1=kfda.predict(model, test)
  predtest2=predtest1$class
  predtest3=as.vector(predtest2)
```

Lampiran 10. *Syntax* Analisis Diskriminan Kernel (Lanjutan)

```
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,
                      SensTrain,
                      SensTest,
                      SpesTrain,
```

Lampiran 10. *Syntax* Analisis Diskriminan Kernel (Lanjutan)

```
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:/KULIAH/TA/Data/Hasil/
Andisker_hasil_total.csv ")
write.csv(hasilmean,file=" E:/KULIAH/TA/Data/Hasil/
Andisker _hasil_total_mean.csv")
```

Lampiran 11. 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	:	Dewi Lutfia Pratiwi
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 :

"Klasifikasi Imbalanced Data Status Ketertinggalan Desa di Jawa timur Berdasarkan Penerapan Metode Combine Sampling pada Regresi Logistik Ridge dan Analisis Diskriminan Kernel "

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

BIODATA PENULIS

Penulis dengan nama lengkap Dewi Lutfia Pratiwi dilahirkan di Kabupaten Madiun pada 21 Desember 1995. Penulis menempuh pendidikan formal di SDN Wungu 01, SMPN 4 Madiun, dan SMAN 2 Madiun. Kemudian penulis diterima sebagai Mahasiswa Departemen Statistika ITS melalui jalur SNMPTN pada tahun 2014. Selama masa perkuliahan, penulis aktif di berbagai kepanitiaan dan organisasi. Penulis aktif dalam organisasi Divisi *Statistics Computer Course* (SCC) HIMASTA-ITS selama dua periode. Di bidang akademik, penulis diberi kesempatan untuk menjadi Juara II *Indonesian Research Competition* dalam 3rd ISCO, finalis Konferensi Matematika di Universitas Indonesia, dan Juara 1 LKTIN di Universitas Mulawarman. Selain itu, penulis juga berkesempatan menjadi peserta *Regional Conference on Student Activism* (RECONSA) 2018 di Universiti Teknologi PETRONAS dan mendapat penghargaan *best gold paper* pada acara tersebut. Penulis juga pernah diberi kesempatan menjadi asisten dosen mata kuliah Analisis Multivariat serta telah mengikuti beberapa kegiatan *survey* sebagai pengaplikasian ilmu statistika. Apabila pembaca ingin memberi kritik dan saran serta diskusi lebih lanjut mengenai Tugas Akhir ini, dapat menghubungi penulis melalui email dewilp21@gmail.com atau nomor telepon 085736761966.