TUGAS AKHIR- SS141501

PENDEKATAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* DALAM MENANGANI KLASIFIKASI *IMBALANCED* DATA BINER
(STUDI KASUS: STATUS KETERTINGGALAN DESA DI JAWA TIMUR)

CANGGIH SHOFFI IMANWARDHANI
NRP 062114 4000 0051

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 – SS 141501

PENDEKATAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* DALAM MENANGANI KLASIFIKASI *IMBALANCED* DATA BINER

(STUDI KASUS: STATUS KETERTINGGALAN DESA DI PROVINSI JAWA TIMUR)

CANGGIH SHOFFI IMANWARDHANI
NRP 062114 4000 0051

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

APPROACHING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE IN HANDLING CLASSIFICATION IMBALANCED DATA BINARY
(CASE STUDY: UNDERDEVELOPED VILLAGES STATUS IN EAST JAVA)

CANGGIH SHOFFI IMANWARDHANI
SN 062114 4000 0051

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

PENDEKATAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* DALAM MENANGANI KLASIFIKASI *IMBALANCED* DATA BINER (STUDI KASUS: STATUS KETERTINGGALAN DESA DI PROVINSI JAWA TIMUR)

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 :
Canggih Shoffi Imanwardhani
NRP. 062114 4000 0051

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

PENDEKATAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* DALAM MENANGANI KLASIFIKASI *IMBALANCED* DATA BINER
(STUDI KASUS: STATUS KETERTINGGALAN DESA DI PROVINSI JAWA TIMUR)

Nama Mahasiswa : Canggih Shoffi Imanwardhani
NRP : 062114 4000 0051
Departemen : Statistika
Dosen Pembimbing : Dr. Santi Puteri Rahayu, M.Si

Abstrak

Klasifikasi pada data imbalanced menghasilkan ketepatan akurasi yang jelek dan cenderung memprediksi ke kelas mayoritas. Untuk menyeimbangkan proporsi kelas dilakukan resampling data minoritas dengan Synthetic Minority Oversampling Technique (SMOTE). Metode klasifikasi yang akan digunakan adalah Regresi Logistik (LR), Regresi Logistik Ridge (LR Ridge) dan Analisis Diskriminan Kernel (ADK). Tujuan dari penelitian ini adalah menganalisis efektifitas SMOTE dalam meningkatkan ketepatan akurasi klasifikasi. Data yang digunakan adalah desa 5 Kabupaten di Jawa Timur yang berjumlah 1.122 desa dengan kelompok berstatus desa tertinggal sebanyak 115 desa. Dengan menggunakan partisi data stratified 10-fold cross validation didapatkan nilai rata-rata AUC, G-mean dan sensitivitas yang kecil pada data imbalanced. Untuk data balanced, setelah dilakukan resampling kelas minoritas dengan SMOTE didapatkan peningkatan nilai rata-rata AUC, G-mean dan sensitivitas yaitu menjadi sekitar 76% serta standar deviasi yang dihasilkan juga lebih kecil dibandingkan klasifikasi data imbalanced. Pada data balanced, LR dengan semua variabel memberikan nilai AUC (76,4%) dan G-mean (76,35%) yang sedikit lebih tinggi dibandingkan metode lain. Peningkatan indikator klasifikasi tertinggi terjadi di nilai sensitivitas yang mencapai 22 kali lipat. Peningkatan nilai G-mean dan sensitivitas tertinggi pada kombinasi SMOTE dan LR Ridge dengan semua variabel.

Kata kunci: Analisis Diskriminan Kernel, Imbalanced Data, Regresi Logistik, Regresi Logistik Ridge, SMOTE.

(*Halaman ini sengaja dikosongkan*)

APPROACHING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) IN HANDLING CLASSIFICATION IMBALANCED DATA BINARY (CASE STUDY: UNDERDEVELOP VILLAGES IN EAST JAVA)

Name	**: Canggih Shoffi Imanwardhani**
Student Number	**: 062114 4000 0051**
Department	**: Statistics**
Supervisor	**: Dr. Santi Puteri Rahayu, M.Si**

Abstrack

The classification of imbalanced data turnout poor accuracy and tend to predicts the majority class. To balance the proportion of classes was used resampling minority data with Synthetic Minority Oversampling Technique (SMOTE). The method of classification to be used is Logistic Regression (LR), Ridge Logistic Regression (LR Ridge) and Kernel Discriminant Analysis (ADK). The purpose of this study is to analyze the effectiveness of SMOTE in improving accuracy. The data used are 5 regency villages in East Java, amounting to 1122 villages which group on underdeveloped villages are 115 villages. The classification used partion data with stratified 10-fold cross validation. Performance classification in imbalanced data gain high total accuracy but low in AUC, G-mean and sensitivity. After balancing with SMOTE, the average of AUC, G-mean and sensitivity were increase around 76% and value of standar deviation were also smaller than imbalanced data classification. In the balanced data, LR with all variables gives high AUC (76,4%) and G-mean (76,35%) values that wew slightly higher than other methods. The highest increase of indicator classification occurred in the sensitivity value which reached 22 times. Highest increased G-mean and sensitivity in combination of SMOTE with LR Ridge.

Keywords : Imbalanced Data, Kernel Discriminant Analysis, Logisctic Regression, Logisctic Ridge Regression, SMOTE

(*Halaman ini sengaja dikosongkan*)

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT, karena atas rahmat-Nya lah peneliti dapat menyelesaikan Tugas Akhir yang berjudul **"Pendekatan *Synthetic Minority Oversampling Technique* dalam menangani *Imbalanced Data Biner* (Studi Kasus: Status Ketertinggalan Desa di Provinsi Jawa Timur)"**. Penulis menyadari selama proses pengerjaan Tugas Akhir tidak terlepas dari bantuan berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terimakasih kepada:

1. Dr. Santi Puteri Rahayu M.Si. selaku dosen pembimbing yang telah bersedia meluangkan waktunya untuk memberikan bimbingan, saran serta dukungan dengan sabar selama proses penyelesaian Tugas Akhir.
2. M. Sjahid Akbar, M.Si dan Imam Safawi, M.Si selaku dosen penguji yang telah banyak memberi masukan kepada penulis.
3. Dr. Suhartono dan Dr. Sutikno, M.Si. selaku Kepala Departemen Statistika dan Ketua Prodi Sarjana Departemen Statistika FMKSD-ITS yang telah memfasilitasi penulis selama proses perkuliahan.
4. Dr. Purhadi, M.Sc. selaku dosen wali yang telah banyak memberikan arahan selama proses perkuliahan penulis.
5. Kedua orang tua serta keluarga besar penulis yang senantiasa memberikan doa dan dukungan kepada penulis.
6. Seluruh pihak yang membantu penulis selama proses perkuliahan dan proses pengerjaan Tugas Akhir.

Penulis sangat mengharapkan kritik dan saran untuk membuat Tugas Akhir ini lebih baik. Besar harapan penulis agar Tugas Akhir ini bermanfaat bagi seluruh pihak.

Surabaya, Juli 2018

Penulis

(*Halaman ini sengaja dikosongkan*)

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

(*Halaman ini sengaja dikosongkan*)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Klasifikasi merupakan proses menemukan sekumpulan model atau fungsi yang menggambarkan dan membedakan konsep atau kelas-kelas data, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui (Han, Kamber, & Pei, 2012). Beberapa metode/algoritma pemodelan yang dikembangkan untuk membantu dalam masalah klasifikasi antara lain Regresi Logistik, Analisis Diskriminan, *Classification Adaptive Regression Tree* (CART), *Multivariate Adaptive Regression Spline* (MARS), *k-Nearest Neighbour* (k-NN), *Support Vektor Mechine* (SVM), dan lain-lain. Teknik pemodelan ini telah banyak diterapkan pada data *balanced*, apabila klasfikasi diterapkan pada data *imbalanced* akan terjadi kesalahan prediksi. Klasifikasi pada *imbalanced data* akan cenderung mengklasifikasikan kelas mayoritas dan mengabaikan kelas minoritas dan menghasilkan ketepatan akurasi yang kurang baik. Metode klasifikasi Regresi Logistik menunjukkan performa yang buruk ketika bekerja pada data dengan proporsi kelas yang tidak seimbang (*imbalanced data*) (King & Zeng, 2001).

Untuk mengatasi masalah *imbalanced data* dapat menggunakan pendekatan *Synthetic Minority Oversampling Technique* (SMOTE) yang pertama kali diperkenalkan oleh Chawla (2002). Pendekatan ini bekerja dengan membuat "*synthetic*" data, yaitu mereplikasi data dari data minor. SMOTE dapat membangun data sampel kelas minoritas secara sintesis dengan cara proses interpolasi antar *instance* (observasi) pada kelas minoritas yang terletak berdekatan. SMOTE merupakan teknik *oversampling* yang baik dan efektif untuk menangani *overfitting* pada proses *oversampling* untuk menangani ketidakseimbangan di kelas minoritas (Chawla et al., 2002).

Beberapa penelitian tentang klasifikasi *imbalanced* data yang telah dilakukan dengan menggunakan metode SMOTE adalah Novritasari dan Purnami (2015), melakukan klasifikasi

kerentanan seseorang terserang penyakit stroke di Jawa Timur dengan menggunakan SMOTE dan SVM (*Suppor Vektor Machine*). Penelitian tersebut menggunakan data sebanyak 65918 observasi, rasio penderita stroke dan bukan penderita stroke adalah 1:129. Setelah dilakukan SMOTE, proporsi data menjadi seimbang dan didapatkan hasil ketepatan klasifikasi akurasi, sensitivitas, dan spesifitas yang lebih tinggi daripada data stroke awal yang *imbalanced*. Penelitian lain dilakukan oleh Hairani, Setiawan, dan Adji (2016) berjudul Metode Klasifikasi Data Mining dan Teknik Sampling SMOTE Menangani *Class Imbalance* Untuk Segmentasi Customer Pada Industri Perbankan, menunjukkan metode J48+SMOTE mampu menangani *class imbalance* pada *dataset* Bank Direct Marketing pada industri perbankan dengan ketepatan klasifikasi sebesar 93,21%. Siringoringo (2018) menerapkan SMOTE dan *k-Nearest Neighbor* untuk klasifikasi data tidak seimbang pada data *Credit Card Fraud* yang terdiri dari 29.976 data dengan *imbalanced ratio* sebesar 3,521. Menggunakan skema evaluasi 10-*cross fold validation* diperoleh peningkatan performansi klasifikasi mencapai 80% dengan menerapkan penggabungan SMOTE dan *kNN*.

Metode klasifikasi yang umum digunakan adalah Regresi Logistik. Analisis regresi logistik merupakan analisis statistik yang digunakan untuk memodelkan hubungan antara variabel dependen dengan variabel independen, dimana variabel dependennya memiliki dua kemungkinan nilai kategorik (dikotomus). Menurut Antipov & Pokryshevskaya (2010) regresi logistik sangat menarik karena beberapa hal, yaitu (1) secara konsep sederhana, (2) mudah diinterpretasikan, dan (3) terbukti dapat menyediakan hasil yang akurat dan baik. Regresi logistik merupakan metode klasifikasi linier, dimana klasifikasi tersebut memberikan keuntungan dalam prosedur *training* dan *testing* yang efisien, terutama ketika diimplementasikan pada data besar dan berdimensi tinggi (Yuan, Ho, & Lin, 2012). Namun Regresi Logistik belum mampu mengatasi masalah apabila didalam model terdapat hubungan (korelasi) yang tinggi antar variabel prediktor yang disebut multikolinieritas. Efek dari multikolinieritas ini dapat mengakibatkan estimasi parameter

dari model yang didapatkan menjadi bias dan varians yang besar (Sungkono & Nugrahaningsih, 2017). Ozkale et al (2017) merekomendasikan estimasi parameter *ridge logistik* sebagai alternatif dalam mengatasi multikolinieritas dalam data.

Analisis Regresi Logistik Ridge dilakukan dengan penambahan suatu bilangan positif kecil θ pada estimasi parameter yang disebut *ridge parameter*. Regresi Logistik Ridge memiliki estimasi yang bias, estimasi ini memiliki variansi estimator koefisien maksimum (Ryan, 2009). Penelitian mengenai Regresi Logistik Ridge pernah dilakukan oleh Cessie dan Houwelingen (1992), Sunyoto (2009) dan Putra (2015). Cessie dan Houwelingen menunjukkan estimator *ridge* dikombinasikan dengan Regresi Logistik dapat memperbaiki model dan menambah akurasi prediksi observasi untuk data kanker ovarium. Sunyoto melakukan prediksi menentukan keberhasilan Siswa SMA Negeri 1 Kediri yang diterima di Perguruan Tinggi Negeri. Putra meneliti tentang Indeks Pembangunan Manusia di Jawa Timur menggunakan Regresi Logistik Ridge dengan jumlah kabupaten/kota di Jawa Timur sebanyak 38 data dan *imbalanced ratio* sebesar 5,333. Menggunakan metode *backward elimination* dari Regresi Logistik Ridge didapatkan ketepatan klasifikasi sebesar 97,37%.

Oleh karena itu, pada penelitian ini digunakan metode Regresi Logistik Ridge menggunakan pendekaan SMOTE untuk menangani *imbalanced data*. Salah satu studi kasus yang memiliki *imbalanced data* adalah data klasifikasi desa tertinggal di Jawa Timur tahun 2014 yang dikeluarkan oleh BPS. BPS mengelompokkan desa yang berstatus tertinggal di Provinsi Jawa Timur sebanyak 208 desa dan desa yang tidak tertinggal sebanyak 7.513 desa. Terdapat data yang tidak seimbang antara jumlah desa tertinggal dan desa yang tidak tertinggal dengan rasio *imbalance* sebesar 1:36. Pengklasifikasian desa tertinggal ditujukan untuk memetakan kondisi desa di Indonesia berdasarkan tingkat perkembangannya, menetapkan sasaran/target pembangunan dalam lima tahun kedepan, serta memotret kinerja pembangunan yang sudah dilaksanakan di desa. Direktorat Jendral Pembangunan Daerah Tertinggal (Ditjen PDT) mengelompokkan daerah tertinggal di Indonesia,

untuk provinsi Jawa Timur daerah yang termasuk tertinggal adalah Kabupaten Sampang, Kabupaten Bangkalan, Kabupaten Situbondo, dan Kabupaten Bondowoso. Selain itu, persentase desa tertinggal tertinggi berdasarkan pengelompokan oleh BPS tahun 2014 adalah Kabupaten Bangkalan, Kabupaten Situbondo, Kabupaten Sumenep, Kabupaten Bondowoso, dan Kabupaten Sampang.

Klasifikasi status desa di Indonesia dilakukan oleh BPS pada tahun 2014, dengan menggunakan metode PCA. Data yang digunakan untuk penentuan desa tertinggal pada tahun 2014 adalah data pendataan Podes dengan menggunakan 42 indikator yang terdapat pada 12 variabel. Penelitian mengenai identifikasi desa tertinggal telah dilakukan oleh beberapa peneliti. Sambodo, Purnami, dan Rahayu (2014) melakukan klasifikasi status ketertinggalan desa di Jawa Timur sebanyak 8502 data menggunakan pendekatan *Reduce Support Vektor Machine* (RSVM) didapatkan akurasi tertinggi sebesar 71,65% dengan *10-fold CV* dan *subset matrix* sebesar 10%. Penelitian lain yang menggunakan studi kasus studi kasus klasifikasi desa tertinggal adalah Klasifikasikan *imbalanced data* pada dengan metode *Rare Event Weighted Logistik Regression* (RE-WLR) (Sulasih, Purnami, & Rahayu, 2016). Unit yang digunakan adalah desa di Jawa Timur sebanyak 7.721 desa dimana 207 diantaranya termasuk dalam desa tertinggal dengan rasio desa tertinggal 1:36. Menggunakan RE-WLR didapatkan akurasi total sebesar 98,04%. Trisaputra & Nida (2016) melakukan eksplorasi dan klasifikasi desa tertinggal di Indonesia menggunakan pendekatan *data mining* dengan metode *decision tree* dan Regresi Logistik. Data yang digunakan terdiri dari 77.961 desa di Indonesia dan variabel yang diteliti sebanyak 205 variabel. Dengan menggunakan kombinasi metode tersebut menghasilkan akurasi klasifikasi desa tertinggal sebesar 77%.

Selain menggunakan Regresi Logistik, Regresi Logistik Ridge, sebagai pembanding digunakan metode klasifikasi Analisis Diskriminan Kernel. Analisis Diskriminan Kernel adalah pendekatan diskriminan nonlinier pada bentuk maupun teksturnya yang tidak memerlukan asumsi apapun (Li, 2002). Li menyebutkan bahwa metode Analisis Diskriminan Kernel

memberikan kinerja terbaik daripada metode *Principal Component Analysis* (PCA), Kernel PCA, dan Analisis Diskriminan Linier. Fungsi kernel yang digunakan pada penelitian ini adalah kernel Gaussian *Radial Basis Function* (RBF). Fungsi kernel Gaussian (RBF) dipilih karena paling sering digunakan dan dapat mengidentikkan dua kelas distribusi (You, Hamsici, & Matinez, 2010). Selain itu kelebihan dari fungsi kernel menurut Hsu, Chang, & Lin (2003) dapat digunakan pada data yang besar. Penelitian mengenai Analisis Diskriminan Kernel dilakukan oleh Wahyuningtyas dan Otok (2012) untuk mengklasifikasikan kelulusan tes keterampilan seleksi nasional masuk perguruan tinggi bidang olahraga. Peneliti tersebut menunjukkan nilai pengklasifikasian metode Analisis Diskrimininan Kernel lebih akurat dibandingkan metode diskriminan linier. Nilai akurasi yang didapatkan sebesar 96,82%. Septianingrum, Syafitri, & Soleh (2010) menerapkan Analisis Diskriminan Kernel pada Komponen Kimia Aktif Tanaman Obat Herbal, dari 12 prediksi hanya 2 prediksi yang melakukan kesalahan penempatan klasifikasi. Djuraidah dan Aunuddin (2004) juga membuktikan keakurasian analisis diskriminan kernel dibandingkan analisis diskriminan linier Fisher. Dalam pengelompokan analisis diskriminan kernel memberikan kesalahan klasifikasi yang lebih kecil dibanding dengan analisis diskriminan Fisher.

1.2 Rumusan Masalah

Ketepatan klasifikasi pada data dengan kelas yang tidak seimbang (*imbalanced*) akan memberikan hasil yang buruk daripada klasifikasi pada data yang seimbang (*balanced*). Untuk meningkatkan ketepatan klasifikasi data *imbalanced* dilakukan penerapan SMOTE. Berdasarkan uraian tersebut, permasalahan utama dalam penelitian ini adalah bagaimana efektifitas kinerja SMOTE dalam meningkatkan ketepatan klasifikasi data *imbalanced* menggunakan metode klasifikasi Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel. Untuk itu perlu dianalisis terkait karakteristik desa tertinggal di Jawa Timur. Kemudian, dilakukan perbandingan ketepatan

klasifikasi masing-masing metode baik pada data *imbalanced* maupun data *balanced*.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan sebelumnya, tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

1. Mendeskripsikan karakteristik desa 5 Kabupaten di Jawa Timur berdasarkan variabel yang diduga mempengaruhi status ketertinggalan desa
2. Menganalisis ketepatan klasifikasi pada data *imbalanced* menggunakan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.
3. Menganalisis ketepatan klasifikasi pada data *balanced* menggunakan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.
4. Menganalisis efektivitas metode SMOTE dalam meningkatkan ketepatan klasifikasi pada masing-masing metode.

1.4 Manfaat Penelitian

Manfaat yang ingin dicapai dari penelitian tugas akhir ini adalah sebagai berikut:

1. Menambah wawasan keilmuan mengenai permasalahan dan penanganan *imbalanced data* dengan penerapan metode SMOTE, di klasifikasi dengan metode Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.
2. Dapat dijadikan prediksi status desa tertinggal secara berulang-ulang sehingga dapat membantu perencanaan dan pengambilan keputusan secara lebih efisien dan tepat sasaran dengan bantuan *machine learning* khususnya saat dilakukan pemekaran desa.

1.5 Batasan Masalah

Batasan masalah dalam penelitian ini adalah data yang digunakan merupakan data Potensi Desa (PODES) di Jawa Timur pada tahun 2014, dikarenakan survey terbaru akan dilakukan pada pertengahan tahun 2018. Sampel yang diklasifikasikan adalah unit desa di 5 Kabupaten dengan

persentase desa tertinggal tertinggi di Jawa Timur dengan jumlah desa sebanyak 1.122. Kabupaten tersebut adalah Bangkalan, Bondowoso, Sampang, Situbondo, dan Sumenep. Variabel yang digunakan dalam penelitian ini adalah berskala numerik. Partisi data *training* dan *testing* menggunakan 10-*fold cross validation stratified*. Analisis Diskriminan menggunakan pendekatan fungsi kernel Gaussian *Radial Basis Function* (RBF). Variabel yang diduga mempengaruhi status desa disesuaikan dengan kondisi yang ada di desa dan yang tidak memiliki banyak nilai 0 (nol).

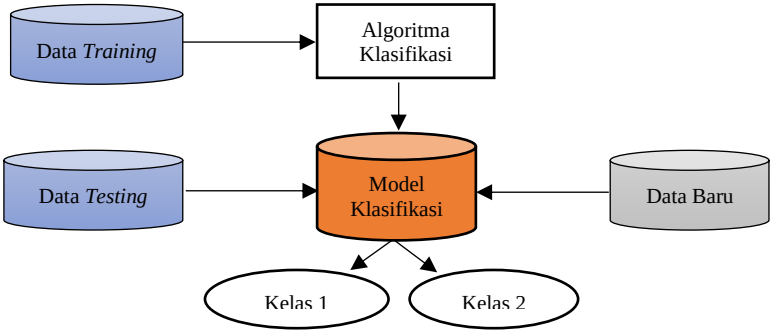
(*Halaman ini sengaja dikosongkan*)

BAB II
TINJAUAN PUSTAKA

Pada bab dua ini dibahas mengenai landasan teori yang berkaitan dengan klasifikasi, metode *Synthetic Minority Oversampling Technique* (SMOTE), Regresi Logistik, Regresi Logistik Ridge, Analisis Diskriminan Kernel dan ketepatan klasifikasi. Dibagian akhir akan dibahas mengenai Indeks Pembangunan Desa (IPD) dan gambaran umum Provinsi Jawa Timur.

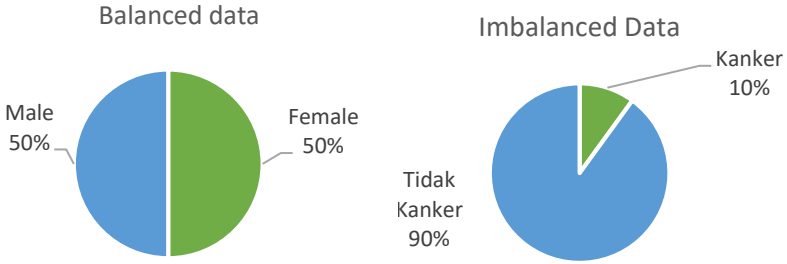
2.1 Klasifikasi

Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan dan membedakan konsep atau kelas-kelas data, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui. Data yang akan dilakukan pengklasifikasian dibagi menjadi 2 kelompok, yaitu data *training* dan data *testing*. Data *training* menggunakan data yang telah diketahui label-labelnya dan kelas setiap observasi, selanjutnya data tersebut digunakan untuk melatih algoritma untuk membangun model atau fungsi yang sesuai. Data *testing* adalah data yang belum diketahui kelas observasinya, data ini akan dipakai untuk mengetes dan mengetahui performa model yang didapatkan dari data *training*. Dari model yang didapat dari data *training*, digunakan untuk melakukan prediksi klasifikasi pada data *testing* atau data baru. Berikut ini adalah ilustrasi dari klasifikasi.



Gambar 2.1 Ilustrasi Proses Klasifikasi

Dalam klasifikasi kelas biner biasanya terdapat dua kondisi himpunan data, yaitu *balanced data* dan *imbalanced* data. *Imbalanced data* terjadi saat kondisi data yang tidak seimbang dengan jumlah data suatu kelas melebihi jumlah data kelas lain, kelas data yang banyak disebut kelas mayoritas sedangkan kelas data yang sedikit disebut kelas minoritas. Klasifikasi pada *imbalanced data* akan cenderung mengklasifikasikan kelas mayoritas dan mengabaikan kelas minoritas. Sebagai contoh dataset yang tidak seimbang adalah dalam suatu kasus terdapat 10 pasien, 9 pasien tidak terdeteksi adanya kanker sedangkan 1 pasien diidentifikasi terdapat sel kanker.



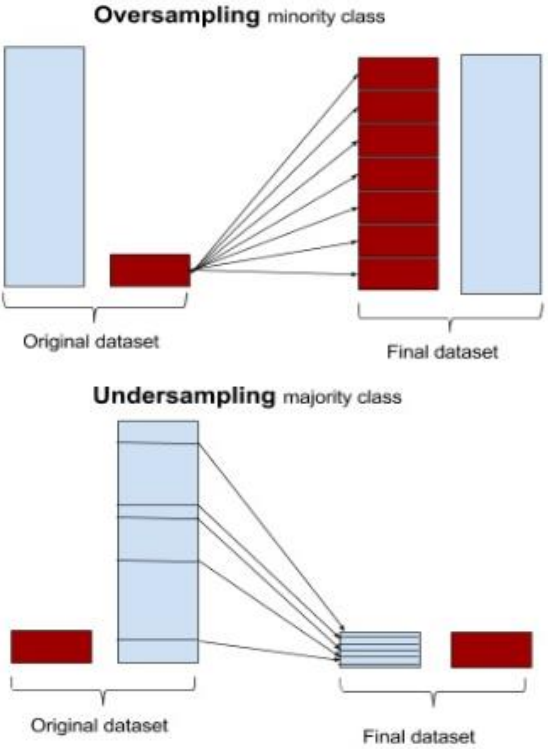
Negatif ≈ Positif Negatif > Positif

Gambar 2.2 Ilustrasi *Balanced Data* dan *Imbalanced Data*

Dari ilustrasi tersebut dapat diketahui ada ketidakseimbangan antara pasien kelas tidak teridentifikasi dengan kelas yang teridentifikasi kanker serta rasio tidak seimbangnya sebesar 1:9, dimana 1 merepresentasikan kelas minoritas (*positive*) sedangkan 100 merepresentasikan kelas mayoritas (*negative*).

Terdapat beberapa pendekatan untuk mengatasi masalah *imbalanced data*, yaitu pendekatan pada level data dengan teknik pengambilan sampel, pendekatan level algoritma, serta metode *ensemble* (Choi, 2010). Teknik pengambilan sampel yang biasanya digunakan untuk mengatasi masalah *imbalanced* data yaitu *over-sampling*, *under-sampling*, dan kombinasi keduanya. Metode *oversampling* dilakukan untuk menyeimbangkan jumlah distribusi data dengan cara meningkatkan jumlah data kelas minoritas. Sedangkan metode

undersampling dilakukan dengan mengurangi jumlah data kelas mayoritas agar data menjadi seimbang. Berikut ini adalah ilustrasi penanganan *imbalanced data* menggunakan metode *undersampling* dan *oversampling*.



Sumber: https://www.kdnuggets.com/2016/08/learning-from-imbalanced-classes.html

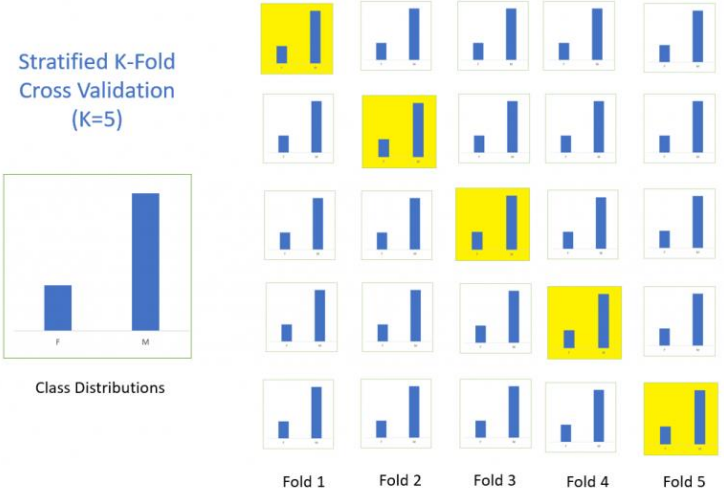
Gambar 2.3 Ilustrasi Penanganan *Imbalanced Data*

2.2 K-Fold Cross Validation

K-fold cross validation (CV) dilakukan untuk membagi data *training* dan data *testing*. *K-fold cross validation* mengulang sebanyak *k*-kali untuk membagi sebuah himpunan contoh secara acak menjadi *k* subset yang saling bebas yaitu D_1, D_2, ..., D_k dengan masing-masing ukuran yang hampir sama. Setiap ulangan disisakan satu subset untuk *testing* dan subset lainnya untuk *training*. Pada iterasi ke-*i*, partisi D_i akan diatur sebagai data *testing* dan partisi yang tersisa lainnya akan

digunakan sebagai data *training* untuk memperoleh model. Artinya, pada iterasi yang pertama, partisi D_2, D_3, ..., D_k akan menjadi data *training* untuk medapatkan model yang pertama yang akan diuji dengan data pada partisi D_1. Pada iterasi kedua partisi D_1, D_3, ..., D_k akan menjadi data *training* kemudian D_2 akan menjadi data *testing*, begitu seterusnya (Han, Kamber, & Pei, 2012). Prosedur ini diulang sebanyak *k*-kali sedemikian sehingga setiap subset digunakan untuk pengujian tepat satu kali.

Pada penelitian ini akan dilakukan *stratified k-fold CV* dengan nilai *k* yang digunakan adalah 10, yang berarti dilakukan 10 kali pengujian akurasi model. *Stratified* adalah proses pengambilan sampel secara random dengan proporsi kelas yang sama setiap *fold*. Adapun kelebihan dari *stratified k-fold cross validation* adalah menghindari adanya *overfitting* pada data training (Zhang, Wu, & Wang, 2011). Menurut Kohavi (1995), *stratified* digabung dengan *10-fold CV* menghasilkan bias dan varians yang rendah. Berikut ini adalah ilustrasi dari *stratified k-fold cross validation* menggunakan $k = 5$.



Sumber: https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2015/11/skfold-768x530.png

Gambar 2.4 Ilustrasi *Stratified K-Fold Cross Validation* (*k*=5)

2.3 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) adalah salah satu turunan dari metode *oversampling*. SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla. Pendekatan ini bekerja dengan membuat replikasi dari data minoritas. Replikasi tersebut dikenal dengan data sintetis (*syntetic data*). Metode SMOTE bekerja dengan mencari *k nearest neighbors* (yaitu ketetanggaan terdekat data sebanyak *k*) untuk setiap data di kelas minoritas, setelah itu dibuat data sintetis sebanyak persentase duplikasi data minor (*percentage oversampling*, N%) yang diinginkan dan *k-nearest neighbors* yang dipilih secara acak (Chawla et al., 2002). (jumlah data kelas mayoritas/jumlah data kelas minoritas) x 100%

$$N\% = \frac{\text{Jumlah data kelas mayoritas}}{\text{Jumlah data kelas minoritas}} \times 100\% \qquad (2.1)$$

Nearest neighbor dipilih berdasarkan jarak Euclidian antara kedua data. Misalkan diberikan dua data dengan *p* dimensi yaitu $X^T = [x_1, x_2, \dots, x_n]$ dan $Y^T = [y_1, y_2, \dots, y_n]$, maka jarak Euclidian $d(x, y)$ adalah

$$x_{knn} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \qquad (2.2)$$

Secara umum, rumus menentukan data sintetis sebagai berikut:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \qquad (2.3)$$

dengan

x_{syn} adalah data sintetis hasil dari replikasi

x_i adalah data yang akan direplikasi

x_{knn} adalah data yang memiliki jarak terdekat dari data yang akan direplikasi

δ adalah bilangan random antara 0 dan 1.

Jika bilangan random mendekati 0, maka data sintetis akan sama dengan data minoritas asal. Jika mendekati 1 maka data sintetis akan sama dengan tetangga terdekat (Santoso et al, 2017). Menurut Santoso et al (2017), jika menggunakan bilangan random sekitar 0,5, kemungkinan data sintesis akan sama dengan data mayoritas.

Contoh penggunaan metode SMOTE pada data adalah sebagai berikut:

No	X_1	X_2	X_3	Y
1	2	29	34	1
2	3	25	30	1
3	3	28	36	1
4	4	29	38	0
5	4	33	39	0

No	X_1	X_2	X_3	Y
6	5	35	40	0
7	5	32	40	0
8	4	34	43	0
9	5	35	42	0
10	5	37	42	0

Data yang akan direplikasi yaitu data dari kelas minoritas (Y=1). Jumlah data minoritas sebanyak 3, sedangkan jumlah data mayoritas (Y=0) sebanyak 7, sehingga nilai persentase SMOTE yang akan digunakan adalah (7/3) x 100% = 233,33%. Hal ini akan dilakukan replikasi 1 kali pada setiap data minoritas dan tetangga data dari data yang akan direplikasi akan dipilih hanya salah satu yang merupakan tetangga data terdekat (x_{knn}). Sehingga dengan persamaan (2.2), maka data setiap kelas minoritas akan menghasilkan satu data replikasi dengan $\delta = 0{,}3$ dan penjelasan pencarian jarak *Euclid* dapat dilakukan 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

$$d\left(\begin{bmatrix}2\\29\\34\end{bmatrix},\begin{bmatrix}3\\25\\30\end{bmatrix}\right)=\sqrt{(2-3)^2+(29-25)^2+(34-30)^2}=5{,}745$$

Observasi 1 dengan observasi 3

$$d\left(\begin{bmatrix}2\\29\\34\end{bmatrix},\begin{bmatrix}3\\28\\36\end{bmatrix}\right)=\sqrt{(2-3)^2+(29-28)^2+(34-36)^2}=2{,}449$$

Berdasarkan perhitungan, didapatkan nilai jarak terdekat ada di perhitungan observasi 1 dengan observasi 3. Sehingga yang digunakan x_{knn} adalah observasi 3. Perhitungan data *sintesis* dengan persamaan (2.3) adalah sebagai berikut:

$$x_{syn} = \begin{bmatrix} 2 \\ 29 \\ 34 \end{bmatrix} + \left(\begin{bmatrix} 3 \\ 28 \\ 36 \end{bmatrix} - \begin{bmatrix} 2 \\ 29 \\ 34 \end{bmatrix} \right) \times 0,3 = \begin{bmatrix} 2 \\ 29 \\ 34 \end{bmatrix} + \begin{bmatrix} 0,3 \\ -0,3 \\ 0,6 \end{bmatrix} = \begin{bmatrix} 2,3 \\ 28,7 \\ 34,6 \end{bmatrix}$$

Dapat diketahui data sintesis yang dihasilkan adalah data dengan $x_{syn(1)} = 2,3$; $x_{syn(2)} = 28,7$; $x_{syn(3)} = 34,6$; dan $y_{\text{syn}} = 1$ dan seterusnya untuk observasi data minoritas yang lain hingga banyak data minoritas tereplikasi sebanyak persentase *oversampling* yang diinginkan.

2.4 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)

$$f(y_i) = (\pi(\mathbf{x}_i))^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}, y = \{0,1\} \qquad (2.4)$$

π_i adalah probabilitas (peluang) dari kejadian ke-i.

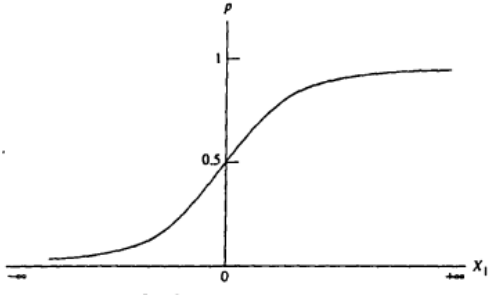
Jika $y_i = 0$, maka $f(y_i) = (\pi_i)^0(1-\pi_i)^{1-0} = (1-\pi_i)$
Jika $y_i = 1$, maka $f(y_i) = (\pi_i)^1(1-\pi_i)^{1-1} = \pi_i$
Menurut Hosmer dan Lemeshow (2000), fungsi logistik yang digunakan untuk memodelkan $\mathbf{x}_i$ dengan nilai ekspektasi y_i nya yaitu

$$\pi(\mathbf{x}_i) = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \qquad (2.5)$$

atau

$$\pi(\mathbf{x}_i) = \frac{1}{1 + e^{-(\mathbf{x}_i^T \boldsymbol{\beta})}} \qquad (2.6)$$

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



Sumber: Sharma,1996

Gambar 2.5 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.

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

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

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

$$\text{Logit}\left[\pi\left(\mathbf{x}_i\right)\right] = \left(\mathbf{x}_i^T\boldsymbol{\beta}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad (2.7)$$

Pengklasifikasian observasi dilakukan dengan terlebih dahulu mengestimasi nilai probabilitas pada persamaan (2.6). 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 (2.4) didapatkan fungsi *likelihood*: (Hosmer, Lemeshow, & Sturdivant, 2013)

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

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

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

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

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

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

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

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

$$\ln \left(L(\mathbf{X}, \boldsymbol{\beta}) \right) = \sum_{i=1}^{n} \left(y_i \left(\mathbf{x}_i{}^T \boldsymbol{\beta} \right) - \ln \left[1+e^{\mathbf{x}_i{}^T \boldsymbol{\beta}} \right] \right) \qquad (2.8)$$

Melalui Persamaan 2.8 dilakukan penurunan terhadap β menjadi Persamaan 2.9

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

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

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

$$\mathbf{X}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0 \qquad (2.10)$$

Persamaan (2.10) didapatkan menggunakan metode Newton-Raphson. Turunan kedua adalah sebagai berikut:

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

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

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

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

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

$$\frac{\partial^2 \ln L\left(\mathbf{X},\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \mathbf{H}\left(\boldsymbol{\beta}\right) = -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T Var\left(\pi_i\right) \qquad (2.11)$$

dimana $\boldsymbol{H}(\boldsymbol{\beta})$ adalah matriks Hessian. Karena turunan kedua selalu bernilai negative, maka didapat bahwa nilai $\boldsymbol{\beta}$ membuat fungsi *likelihood* bernilai ekstrim maksimum. Namun karena hasil turunan pertama pada persamaan (2.9) tidak mendapatkan hasil yang eksplisit atau rumus untuk mencari nilai $\boldsymbol{\beta}$ tidak didapat, maka akan digunakan Deret Taylor. Apabila dilakukan ekspansi berdasarkan Deret Taylor disekitar nilai β, maka didapatkan persamaan (2.11).

$$\frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}} = \frac{\partial L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}}\bigg|_{\hat{\beta}=\hat{\beta}_0} + \frac{\partial^2 L\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}^T}\bigg|_{\hat{\beta}=\hat{\beta}_0} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0\right) \qquad (2.12)$$

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

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

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

Hasil substitusi persamaan (2.7) dan (2.10) ke dalam persamaan (2.12) menghasilkan estimasi parameter $\widehat{\boldsymbol{\beta}}$ ditunjukkan pada persamaan (2.13). (Hosmer, Lemeshow, & Sturdivant, 2013)

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

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

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

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

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{X}^T\mathbf{W}\mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)$$
$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\left[\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)\right]$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left[\mathbf{X}\hat{\boldsymbol{\beta}}_0 + \mathbf{W}^{-1}\left(\mathbf{y} - \hat{\boldsymbol{\pi}}\right)\right]$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \qquad (2.13)$$

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

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

$$\mathbf{z}_i = \text{Logit}\left[\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\right] + \frac{y_i - \hat{\boldsymbol{\pi}}(\mathbf{x}_i)}{\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\left[1-\hat{\boldsymbol{\pi}}(\mathbf{x}_i)\right]} \qquad (2.14)$$

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

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

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 :

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

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 $|Z| > Z_{\frac{\alpha}{2}}$ atau *p-value* < α. Artinya, variabel ke-j berpengaruh signifikan terhadap pembentukan model.

2.5 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).

2.5.1 Multikolinieritas

Multikolinieritas adalah keadaan dimana terdapat korelasi atau hubungan linier antara variabel-variabel prediktor sehingga variabel-variabel tersebut tidak bersifat orthogonal. Variabel prediktor yang bersifat orthogonal adalah variabel yang memiliki nilai korelasinya sama dengan nol. Menurut Montgomery, salah satu ukuran yang dapat digunakan untuk menguji adanya multikolinearitas pada regresi adalah nilai *Variance Inflation Factors (VIF)* yang dihasilkan. Besarnya nilai VIF ini bergantung pada nilai koefisien determinasi (R^2) yang dihasilkan. Apabila nilai *VIF* lebih dari 10, maka dapat diindikasikan terdapat kasus multikolinieritas (Vago & Kemeny, 2006). Nilai *VIF* dinyatakan sebagai berikut. (Hocking, 2003)

$$VIF = \frac{1}{1 - R_{j}^{2}},\ j = 1, 2, \ldots, p \tag{2.22}$$

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

$$R_j^2 = 1 - \frac{SSE}{SST} \qquad (2.23)$$

dimana : $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ dan $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
Pedoman suatu model regresi yang bebas multikolinearitas adalah mempunyai nilai VIF disekitar angka 1.

2.6 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 (2.8) dituliskan:

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

Kemudian, Vago & Kemeny (2006) dengan menerapkan teknik pada Regresi Ridge pada Regresi Logistik, didapatkan fungsi obyektif untuk Regresi Logistik Ridge pada persamaan (2.24).

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

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

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

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

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 $\mathbf{x}_i$ 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.

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

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

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

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

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

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

$$\frac{\partial^2 \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus} \partial \hat{\boldsymbol{\beta}}^{\oplus T}} = -\mathbf{X}^T \mathbf{W} \mathbf{X} - 2\theta \mathbf{I} \qquad (2.25)$$

dengan $\mathbf{W} = diag[\pi_i(\mathbf{x}_i)(1-\pi_i(\mathbf{x}_i))]$

Estimasi parameter Regresi Logistik Ridge menggunakan metode MLE dengan iterasi *Newton-Raphson* yang akan digunakan untuk memaksimumkan fungsi obyektif pada persamaan (2.24). Kemudian diekspansikan di sekitar $\beta^{\oplus}$menurut Deret *Taylor* dan didapatkan persamaan (2.26).

$$\left.\frac{\partial \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0^{\oplus}} = -\left.\frac{\partial^2 \phi(\hat{\boldsymbol{\beta}}^{\oplus})}{\partial \hat{\boldsymbol{\beta}}^{\oplus}\partial \hat{\boldsymbol{\beta}}^{\oplus T}}\right|_{\hat{\beta}^{\oplus}=\hat{\beta}_0^{\oplus}} (\hat{\boldsymbol{\beta}}^{\oplus} - \hat{\boldsymbol{\beta}}_0{}^{\oplus}) \quad (2.26)$$

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

$$\hat{\boldsymbol{\beta}}^{\oplus} = (\mathbf{X}^T\mathbf{W}\mathbf{X} + 2\theta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \quad (2.27)$$

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

2.7 Analisis Diskriminan Linier

Menurut Hair et al. (2006), 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). Ini berarti ada variabel yang hasilnya tergantung dari data variabel independen. Analisis diskriminan digunakan untuk menentukan fungsi yang mebedakan antar kelompok dan mengkelaskan obyek baru ke dalam kelompoknya (Johnson & Wichern, 2007). 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 et al., 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.

a. Uji Distribusi 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 *kurtosis* untuk menguji apakah suatu data berdistribusi normal multivariat. Nilai dari *skewness* dan *kurtosis* data multivariat dapat dihitung dengan persamaan sebagai berikut:

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}} S^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right]^3 \qquad (2.28)$$

dan

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}} S^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right]^2 \qquad (2.29)$$

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

$$z_1 = \frac{(\mathrm{p}+1)(\mathrm{n}+1)(\mathrm{n}+3)}{6[(\mathrm{n}+1)(\mathrm{p}+1)-6]} b_{1,\mathrm{p}} \qquad (2.30)$$

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

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \qquad (2.31)$$

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.

b. Uji Homogenitas

Asumsi lain yang harus terpenuhi adalah matriks varians kovarians 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 (2.32):

$$C = (1-u)\,\mathrm{M} \tag{2.32}$$

dengan

$$u = \left[\sum_{i=1}^{2} \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{2} (n_i - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(\mathrm{p}+1)} \right],$$

$$M = \left[\sum_{i=1}^{2} (n_i \text{-} 1) \right] \ln \left| \mathbf{S}_{pooled} \right| - \sum_{i=1}^{2} \left[(n_i \text{-} 1) \ln \mathbf{S}_i \right]$$

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

n_i merupakan banyaknya data pada kelompok ke-*i*, untuk nilai i=1,2. $\mathbf{S}_{pooled}$ merupakan matriks varians kovarians kelas gabungan, *p* adalah jumlah variabel independen, dan $\boldsymbol{S}_i$ adalah matriks varians kovarians kelompok ke-i. Persamaan untuk mendapatkan matriks $\boldsymbol{S}_i$ adalah sebagai berikut:

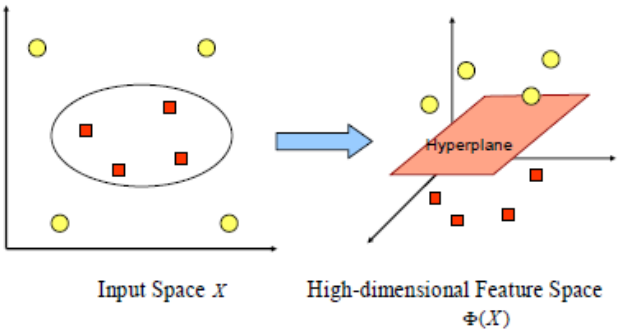
$$\mathbf{S}_i = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

dengan persamaan $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right)$ untuk nilai i = 1,2,…,p dan k = 1,2,…,p. Pada kasus $i = k$ maka nilai $s_{ik} = s_{kk}$ menjadi $s_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2$.

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

2.8 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 (2.6).



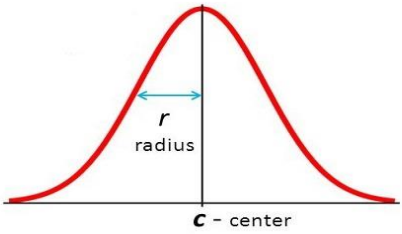
Sumber: (Nugroho, Witarto, & Handoko, 2003)

Gambar 2.6 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 (2.33) dan divisualisasikan pada Gambar (2.7).

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{(\mathrm{x}_i \; x_j)^2}{2\sigma^2}\right) \tag{2.33}$$



Gambar 2. 7 Kernel Gaussian RBF

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

$$J(\boldsymbol{\omega}) = \frac{\omega^T S_B^{\Phi} \omega}{\omega^T S_W^{\Phi} \omega} \tag{2.34}$$

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 (2.34) dan fungsi perluasan dari $\boldsymbol{\omega}$ pada persamaan (2.35).

$$\boldsymbol{\omega} = \sum_{i=1}^{l} \alpha_i \Phi(x_i) \tag{2.35}$$

Persamaan 2.35 dan persamaan m_i^{Φ} menghasilkan persamaan (2.36)

$$\boldsymbol{\omega}^T m_i^{\Phi} = \frac{1}{l_i}\sum_{j=1}^{l}\sum_{k=1}^{l_i}\alpha_j \mathrm{k}(x_j, x_k^i) = \alpha^{\mathrm{T}} M_i \qquad (2.36)$$

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 2.36 diperoleh persamaan baru dari $\boldsymbol{\omega}^T S_B^{\Phi}\boldsymbol{\omega}$ sebagai berikut:

$$\boldsymbol{\omega}^T S_B^{\Phi}\boldsymbol{\omega} = \alpha^T M \alpha \qquad (2.37)$$

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:

$$\boldsymbol{\omega}^T S_W^{\Phi}\boldsymbol{\omega} = \alpha^T N \alpha \qquad (2.38)$$

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 (2.39)

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha} \qquad (2.39)$$

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

$$(\boldsymbol{\omega}, \Phi(x)) = \sum_{i=1}^{l} \alpha k(\boldsymbol{x}_i, \boldsymbol{x}) \qquad (2.40)$$

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 :

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

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

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

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

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

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

2.9 Evaluasi 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 2.1.

Tabel 2. 1 *Confusion Matriks*

		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 kriteria performa klasifikasi yang umum seperti akurasi total, sensitivitas, dan spesifisitas. Akurasi total menunjukkan tingkat keakurasian sistem dalam mengklasifikasikan data dengan benar. Sensitivitas adalah proporsi dari kelas positif yang terprediksi dengan benar atau akurasi kelas yang positif. Sedangkan spesifisitas adalah proporsi dari kelas negatif yang terprediksi dengan benar atau akurasi kelas yang negatif. Berikut ini rumus perhitungan akurasi total klasifikasi, sensitivitas, dan spesifisitas.

$$akurasi\ total = \frac{TP + TN}{TN + TP + FN + FP} \tag{2.41}$$

$$sensitivitas = \frac{TP}{TP + FN} \tag{2.42}$$

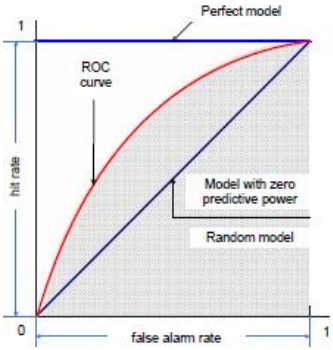
$$spesifisitas = \frac{TN}{TN + FP} \tag{2.43}$$

Perhitungan ketepatan klasifikasi pada kelas tidak seimbang bisa menggunakan *Geometric mean* (*G-mean*). *G-mean* digunakan untuk mengukur performa keseluruhan (*overall classification performance*). Nilai ini akan memaksimalkan keakuratan masing-masing kelas dengan keseimbangan yang baik. (Barandela, Sanchez, Garcia, & Rangel, 2003)

$$G - mean = \sqrt{Sensitivitas \times Spesifisitas} \tag{2.44}$$

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.



Sumber: Haerdle, *et.al.*, 2014

Gambar 2. 8 *ROC Curve*

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 2.8 serta nilai AUC dapat dihitung berdasarkan persamaan (2.45)

$$AUC = \frac{1}{2}\left(sensitivitas + spesifisitas\right) \qquad (2.45)$$

Pengklasifikasian kebaikan model berdasarkan nilai AUC menurut Gorunescu (2011) dalam tabel berikut.

Tabel 2. 2 Kategori Kebaikan Model Berdasarkan AUC

Nilai *AUC*	Kategori
0,90-1,00	*Excellent Classification*
0,80-0,90	*Good Classification*
0,70-0,80	*Fair Classification*
0,60-0,70	*Poor Classification*
0,50-0,60	*Failure*

2.10 Indeks Pembangunan Desa (IPD)

Desa menurut Undang – Undang Nomor 6 Tahun 2014 Tentang Desa adalah kesatuan masyarakat hukum yang memiliki batas wilayah yang berwenang untuk mengatur dan mengurus urusan pemerintah, kepentingan masyarakat setempat. Dalam rangka menilai tingkat kemajuan atau perkembangan desa, maka Desa dibagi menjadi 3 (tiga) klasifikasi, yaitu desa mandiri, desa berkembang, dan desa tertinggal. Desa mandiri adalah desa yang mempunyai

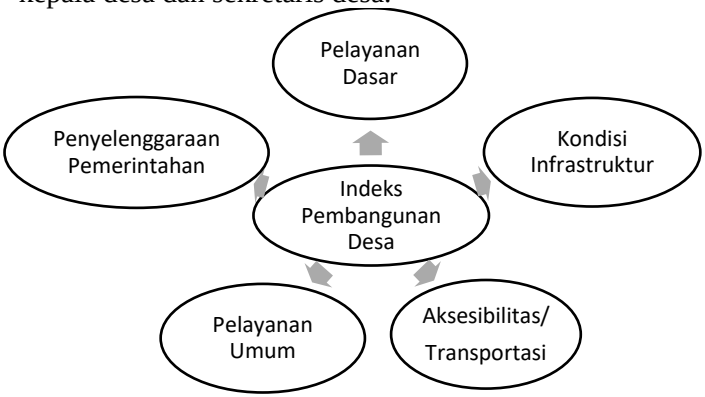
ketersediaan dan akses terhadap pelayanan dasar yang mencukupi, infrastruktur yang memadai, aksesibilitas/ transportasi yang tidak sulit, pelayanan umum yang bagus, serta penyelenggaraan pemerintahan yang sudah sangat baik dan memiliki nilai IPD lebih dari 75. Desa berkembang adalah desa yang mempunyai ketersediaan dan akses terhadap pelayanan dasar, infrastruktur, aksesibilitas/transportasi, pelayanan umum, serta penyelenggaraan pemerintahan yang cukup memadai dan memiliki nilai IPD antara 50 hingga 75. Desa tertinggal adalah desa yang relatif kurang berkembang dan belum terpenuhinya Standar Pelayanan Minimal Desa (SPM Desa) pada aspek kebutuhan sosial dasar, infrastruktur dasar, sarana dasar, pelayanan umum, dan penyelenggaraan pemerintahan dan memiliki nilai IPD kurang dari atau sama dengan 50. Dalam penelitian ini dilakukan penggabungan kelompok desa berkembang dan mandiri menjadi kelas desa tidak tertinggal.

IPD disusun untuk melihat tingkat perkembangan desa di Indonesia. IPD memiliki tujuan sebagai instrument perencanaan pembangunan desa di Indonesia, monitoring dan evaluasi kinerja pembangunan desa, dan pengukuran pencapaian sasaran pembangunan nasional tahun 2015-2019 di Indonesia. Tujuan klasifikasi desa adalah untuk memetakan kondisi desa di Indonesia berdasarkan tingkat perkembangannya, menetapkan target/sasaran pembangunan dalam lima tahun kedepan, memotret kinerja pembangunan yang sudah dilaksanakan. Oleh BPS, IPD dibagi menjadi 5 dimensi dengan menyesuaikan dengan ketersediaan data/variabel dalam data Potensi Desa 2014, yaitu sebagai berikut:

1. Pelayanan dasar, mewakili aspek pelayanan dasar untuk mewujudkan bagian dari kebutuhan dasar, khusus untuk pendidikan dan kesehatan. Variabel yang termasuk sebagai komponen penyusunnya meliputi ketersediaan dan akses terhadap fasilitas pendidikan seperti TK, SD, SMP, dan SMA serta ketersediaan dan akses terhadap fasilitas kesehatan seperti rumah sakit, rumah sakit bersalin, puskesmas, tempat praktik dokter, poliklinik/balai pengobatan, tempat praktik bidan, poskesdes, polindes, dan apotek.

2. Kondisi infrastruktur, mewakili kebutuhan dasar, sarana prasarana, pengembangan ekonomi lokal, dan pemanfaatan sumber daya alam secara berkelanjutan sengan memisahkan aspek aksesibilitas/transportasi. Variabel-variabel penyusunnya mencakup ketersediaan infrastruktur ekonomi seperti kelompok pertokoan, minimarket, maupun toko kelontong, pasar, restoran, rumah makan, akomodasi hotel/penginapan, serta bank. Selanjutnya ketersediaan infrastruktur energi meliputi listrik, penerangan jalan, dan bahan bakar untuk memasak. Ketersediaan infastruktur air bersih dan sanitasi diantaranya seperti sumber air minum, sumber air mandi/cuci, dan fasilitas buang air besar serta ketersediaan dan kualitas infrastruktur komunikasi dan informasi seperti komunikasi menggunakan telepon seluler, internet, dan pengiriman pos/barang.
3. Aksesibiltas/transportasi, dipisahkan sebagai dimensi tersendiri dalam indikator pembangunan desa dengan pertimbangan sarana dan prasarana transportasi memiliki kekhususan dan prioritas pembangunan desa sebagai penghubung kegiatan sosial ekonomi dalam desa. Variabel-variabel penyusunnya meliputi ketersediaan dan akses terhadap sarana transportasi seperti lalu lintas dan kualitas jalan, aksesibilitas jalan, ketersediaan dan operasional angkutan umum, dan aksesibilitas transportasi seperti waktu tempuh per kilometer transportasi ke kantor camat, biaya per kilometer transportasi ke kantor camat, waktu tempuh per kilometer transportasi ke kantor bupati/walikota, dan biaya per kilometer transportasi ke kantor bupati/walikota.
4. Pelayanan umum, merupakan upaya pemenuhanan kebutuhan pelayanan atas barang, jasa, atau pelayanan administratif dengan tujuan memperkuat demokrasi, kohesi sosial, perlindungan lingkungan, dan sebagainya. Variabel-variabel penyusun dimensi ini mencakup penanganan kesehatan masyarakat seperti penanganan kejadian luar biasa (KLB), dan penanganan gizi buruk, serta ketersediaan fasilitas olahraga seperti ketersediaan lapangan olahraga dan kelompok kegiatan olahraga.

5. Penyelenggaraan pemerintahan, merupakan bentuk pelayanan administratif yang diselenggarakan penyelenggara pelayanan bagi warga yang dalam hal ini adalah pemerintah. Oleh karena itu variabel ini perlu diukur dan berdiri sendiri sebagai sebuah indikator pembangunan desa, karena sifatnya sebagai perangkat terlaksananya tujuan pembangunan desa tersebut. Variabel-variabel penyusunnya meliputi kemandirian seperti kelengkapan pemerintahan desa, otonomi desa, dan aset/kekayaan desa, serta kualitas sumber daya manusia seperti kualitas SDM kepala desa dan sekretaris desa.



Gambar 2. 9 Dimensi Indeks Pembangunan Desa menurut BPS

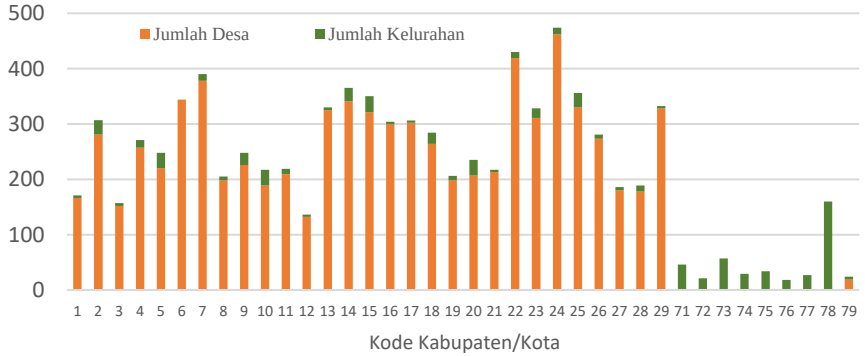
2.11 Gambaran Umum Jawa Timur

Provinsi Jawa Timur adalah salah satu Provinsi di Pulau Jawa, Indonesia. Ibu kota provinsi Jawa Timur terletak di Kota Surabaya. Provinsi Jawa Timur secara astronomis terletak antara 111,0^{o} -114,4^{o} Bujur Timur dan 7,12^{o} -8,48^{o} Lintang Selatan. Jawa Timur dibagi menjadi dua bagian besar yaitu Jawa Timur daratan dan Kepulauan Madura. Provinsi Jawa Timur mempunyai luas wilayah mencapai 47.995 Km^{2}, merupakan provinsi yang memiliki wilayah terluas di Pulau Jawa. Batas wilayah Provinsi Jawa Timur meliputi sebelah utara berbatasan dengan Pulau Kalimantan atau tepatnya dengan Provinsi Kalimantan Selatan; Sebelah Timur Berbatasan dengan Pulau Bali; Sebelah Selatan Berbatasan dengan perairan terbuka, yaitu

Samudra Hindia; dan Sebelah Barat berbatasan dengan Provinsi Jawa Tengah (BAPPEDA PROVINSI JATIM, 2017).

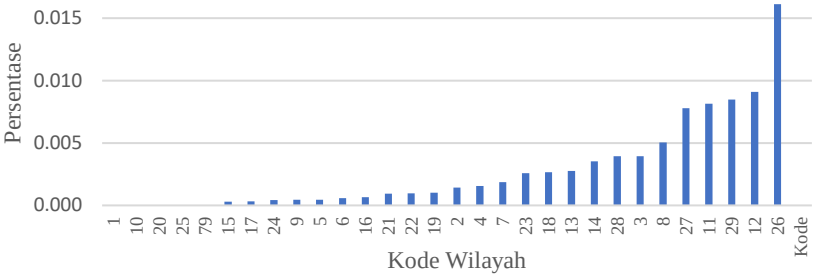
Secara administrasif berdasarkan Peraturan Menteri Dalam Negeri Nomor 56 Tahun 2015 Tentang Kode dan Data Wilayah Administrasi Pemerintahan, kode provinsi Jawa Timur adalah 35, terdiri atas 38 Kabupaten/Kota (29 Kabupaten dan 9 Kota) yang mempunyai 664 Kecamatan dengan 781 Kelurahan dan 7.721 Desa. Setiap kabupaten/kota memiliki satu atau lebih kelurahan, sedangkan kota di Jawa Timur yang memiliki unit desa hanya terdapat di Kota Batu.

Dari keseluruhan desa, 207 desa masuk kategori desa tertinggal, 6.822 desa dalam kategori desa berkembang, dan 693 desa lainnya dalam kategori desa mandiri. Desa berkembang dan desa mandiri diklasifikasikan menjadi desa tidak tertinggal.



Gambar 2. 10 Jumlah Desa dan Kelurahan di Provinsi Jawa Timur

Desa tertinggal di Jawa Timur tersebar di 25 Kabupaten dari 29 Kabupaten yang ada di Jawa Timur. Kabupaten yang tidak memiliki desa yang tertinggal antara lain Kabupaten Pacitan (Kode:01), Kabupaten Banyuwangi (10), Kabupaten Magetan (20), Kabupaten Gresik (25), dan Kota Batu (79). Sedangkan kabupaten dengan persentase status desa tertinggal tertinggi adalah kabupaten Bangkalan (26), kabupaten Situbondo (12), kabupaten Sumenep (29), kabupaten Bondowoso (11), dan kabupaten Sampang (27). Gambar 2.11 adalah visualisasi persentase desa tertinggal setiap Kabupaten/Kota di Provinsi Jawa Timur, terlampir di Lampiran 1.



Gambar 2.11 Persentase Desa Tertinggal di Jawa Timur

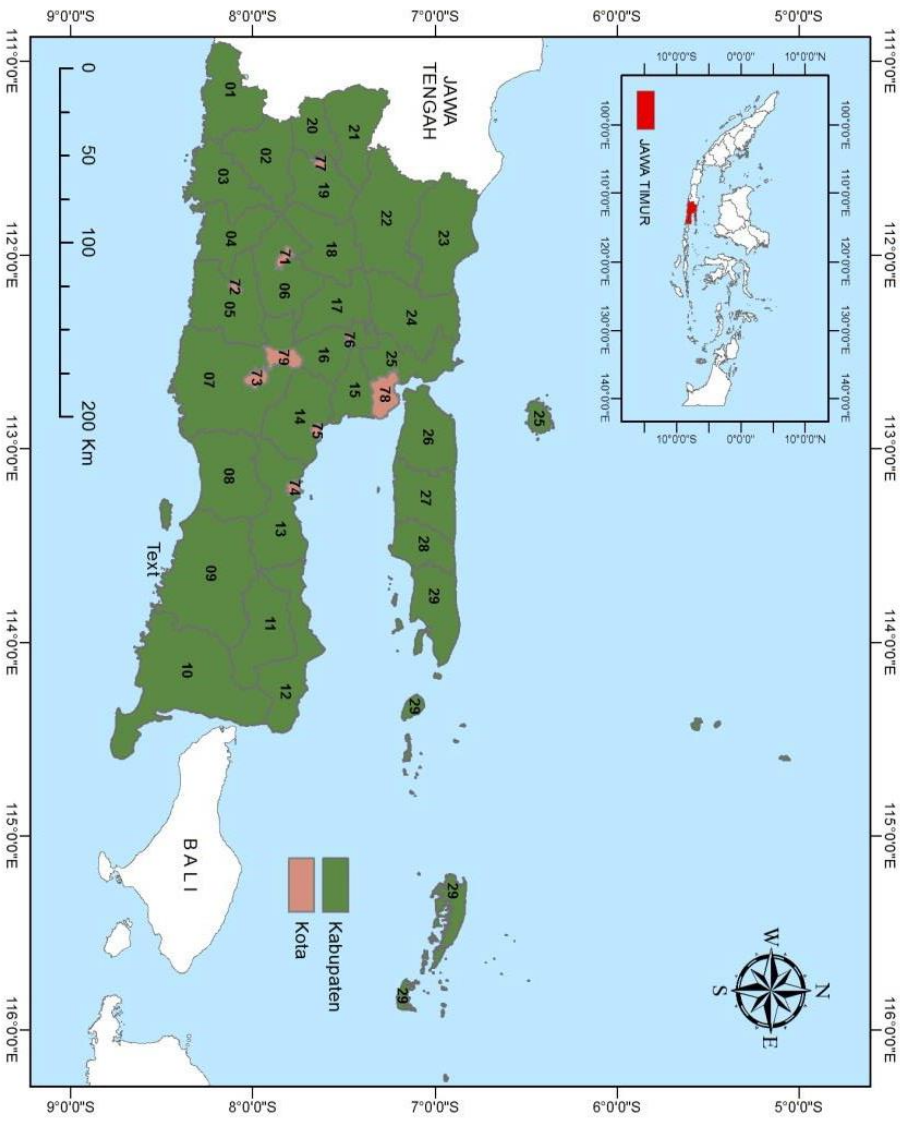
Tabel 2.3 menjelaskan kode wilayah kabupaten/kota untuk provinsi Jawa Timur.

Tabel 2.3 Daftar Kode Kabupaten/Kota di Jawa Timur

Kode	Kabupaten/Kota	Ibu Kota Kabupaten
01	Kabupaten Pacitan	Pacitan
02	Kabupaten Ponorogo	Ponorogo
03	Kabupaten Trenggalek	Trenggalek
04	Kabupaten Tulungagung	Tulungagung
05	Kabupaten Blitar	Kanigoro
06	Kabupaten Kediri	Ngasem
07	Kabupaten Malang	Kepanjen
08	Kabupaten Lumajang	Lumajang
09	Kabupaten Jember	Jember
10	Kabupaten Banyuwangi	Banyuwangi
11	Kabupaten Bondowoso	Bondowoso
12	Kabupaten Situbondo	Situbondo
13	Kabupaten Probolinggo	Kraksaan
14	Kabupaten Pasuruan	Bangil
15	Kabupaten Sidoarjo	Sidoarjo
16	Kabupaten Mojokerto	Mojosari
17	Kabupaten Jombang	Jombang
18	Kabupaten Nganjuk	Nganjuk
19	Kabupaten Madiun	Caruban
20	Kabupaten Magetan	Magetan
21	Kabupaten Ngawi	Ngawi
22	Kabupaten Bojonegoro	Bojonegoro
23	Kabupaten Tuban	Tuban
24	Kabupaten Lamongan	Lamongan

Tabel 2.3 Daftar Kode Kabupaten/Kota di Jawa Timur (Lanjutan)

Kode	Kabupaten/Kota	Ibu Kota Kabupaten
25	Kabupaten Gresik	Gresik
26	Kabupaten Bangkalan	Bangkalan
27	Kabupaten Sampang	Sampang
28	Kabupaten Pamekasan	Pamekasan
29	Kabupaten Sumenep	Sumenep
71	Kota Kediri	-
72	Kota Blitar	-
73	Kota Malang	-
74	Kota Probolinggo	-
75	Kota Pasuruan	-
76	Kota Mojokerto	-
77	Kota Madiun	-
78	Kota Surabaya	-
79	Kota Batu	-



(Sumber:Provinsi Jawa Timur Dalam Angka Tahun 2017)

Gambar 2.12 Peta Provinsi Jawa Timur

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data status desa tertinggal 5 Kabupaten di Jawa Timur tahun 2014. Data desa tertinggal didapat dari Potensi Desa (Podes) Provinsi Jawa Timur 2014 yang dikeluarkan oleh Badan Pusat Statistik. Jumlah desa tertinggal di 5 Kabupaten terdiri dari 1.122 desa yang ada di Jawa Timur dengan jumlah data pada kelas minoritas sebanyak 115 desa tertinggal, sedangkan kelas mayoritas sebanyak 1.007 desa tidak tertinggal, serta rasio *imbalanced* sebesar 1:8,75.

3.2 Variabel Penelitian

Variabel yang digunakan pada penelitian ini terdiri atas variabel respon dan variabel prediktor. Variabel prediktor (X) merupakan variabel yang diduga mempengaruhi variabel respon berdasarkan lima dimensi yang dibentuk oleh BPS Indonesia. Variabel respon (Y) merupakan variabel yang berisi kelas Status desa yang terdiri dari dua kategori yaitu {0} untuk desa tidak tertinggal dan {1} untuk desa tertinggal. Pengukuran variabel respon didapat dari Indeks Pembangunan Desa 2014 yang dikeluarkan oleh BPS dan Kementrian Perencaan Pembangunan Nasional/ Bappenas. Desa yang termasuk tidak tertinggal adalah kategori desa berkembang dan desa mandiri. Sedangkan variabel prediktor didapatkan dari hasil pendataan Podes Provinsi Jawa Timur tahun 2014. Berikut ini adalah penjelasan variabel penelitian yang akan digunakan.

Variabel respon:

Y : Status ketertinggalan desa ({0} untuk desa tidak tertinggal, {1} desa tidak tertinggal)

Variabel prediktor:

1. Pelayanan Dasar
   - X_1 : Rasio sekolah dasar terhadap total murid
   - X_2 : Rasio tempat pos kesehatan desa (poskesdes) terhadap total penduduk
   - X_3 : Rasio tempat praktik bidan terhadap total penduduk

2. Kondisi Infrastruktur
   X_4 : Rasio keluarga pakai listrik terhadap total keluarga
   X_5 : Rasio toko kelontong terhadap total penduduk
3. Aksesibilitas/Transportasi
   X_6 : Jarak tempuh dari kantor kepala desa ke kantor kecamatan (km)
4. Pelayanan Umum
   X_7 : Rasio jumlah warga penderita gizi buruk terhadap total penduduk
5. Penyelenggaraan Pemerintah
   X_8 : Rasio Pendapatan Asli Desa (PAD) terhadap total penduduk

Konsep dan definisi yang digunakan pada data desa tertinggal mengacu pada BPS yaitu:

1. Status ketertinggalan desa. Desa tertinggal adalah desa-desa yang kondisinya tertinggal dibandingkan desa yang lain dan nilai Indeks Pembangunan Desanya kurang dari 50.
2. Rasio banyaknya SD, yaitu jumlah sekolah SD/MI baik negeri maupun swasta dibagi total murid.
3. Rasio tempat pos kesehatan desa (poskesdes), yaitu sarana kesehatan dalam menyediakan pelayanan kesehatan dasar bagi masyarakat desa. Jumlah poskesdes dibagi total penduduk setiap desa.
4. Rasio tempat praktik bidan, yaitu sarana kesehatan yang digunakan untuk tempat praktik bidan dalam memberikan pelayanan ibu hamil dan bayi dibanding total penduduk desa.
5. Rasio keluarga pakai listrik, adalah keluarga yang sumber penerangannya adalah listrik pemerintah dan listrik non pemerintah dibandingkan jumlah keluarga di desa.
6. Rasio banyaknya toko/warung kelontong yaitu jumlah toko/warung kelontong dibagi total penduduk. Toko/warung kelontong adalah bangunan (kedai) yang menjual beraneka barang secara eceran.

7. Jarak tempuh dari kantor kepala desa ke kantor kecamatan, adalah jarak yang sering dilalui dengan kendaraan yang biasa digunakan oleh warga.
8. Rasio jumlah warga penderita gizi buruk gizi buruk, yaitu warga yang kekurangan konsumsi zat gizi disbanding total penduduk.
9. Rasio pendapatan asli desa (PAD) didapat dari hasil usaha (hasil badan usaha milik desa, tanah kas desa); hasil aset (tambatan perahu, pasar desa, tempat pemandian umum, jaringan irigasi); swadaya, partisipasi, dan gotong royong; serta lain-lain pendapatan asli desa (hasil pungutan desa)

Struktur data yang digunakan pada data penelitian ini disajikan pada Tabel 3.1, selengkapnya terlampir di Lampiran 2.

Tabel 3. 1 Struktur Data Penelitian

Desa	Respon	Prediktor			
	Y_i	X_1	X_2	…	X_8
1	0	0,83	0	…	1,073
2	0	1,096	0,022	…	1,172
3	0	0,574	0,025	…	1,043
⋮	⋮	⋮	⋮	⋮	⋮
1120	0	1,115	0	…	0,294
1121	0	1,367	0,025	…	0,627
1122	1	1,333	0	…	1,094

3.3 Langkah Analisis

Langkah yang dilakukan pertama dalam penelitian ini adalah melakukan pemilihan data, perhitungan rasio, dan filtering data. Selanjutnya dalam menjawab tujuan penelitian, langkah analisisnya adalah sebagai berikut:

1. Menganalisis karakteristik desa di 5 Kabupaten Jawa Timur berdasarkan variabel yang diduga mempengaruhi status ketertinggalan desa dengan statistika deskriptif dan *boxplot*.
2. Menganalisis ketepatan klasifikasi data *imbalanced* menggunakan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan bantuan *software R*.
    a. Melakukan klasifikasi data *imbalanced* dengan Regresi Logistik.

i. Mendeteksi adanya multikolinieritas dengan melihat nilai VIF menggunakan persamaan (2.22)
ii. Membagi data menjadi data *training* dan *testing* dengan menggunakan 10-*fold cross validation stratified*.
iii. Melakukan klasifikasi data *training* menggunakan Regresi Logistik.
iv. Menghitung ketepatan klasifikasi pada data *testing* dengan persamaan (2.41) sampai persamaan (2.45).
v. Menguji signifikansi parameter secara serentak dan parsial dengan persamaan (2.15) dan (2.16), berdasarkan pemilihan *fold* terbaik.
vi. Melakukan pemilihan variabel signifikan dengan *backward elimination*.
vii. Mengulangi langkah i, ii, iii, dan iv menggunakan variabel signifikan yang diperoleh dari langkah vi.

b. Melakukan klasifikasi data *imbalanced* 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 dengan parameter *ridge* dipilih otomatis oleh package *R*.
iii. Menghitung ketepatan klasifikasi pada data *testing* dengan persamaan (2.41) sampai persamaan (2.45).
iv. Menghitung ketepatan klasifikasi dengan variabel signifikan yang diperoleh dari langkah 3.a.(vi).

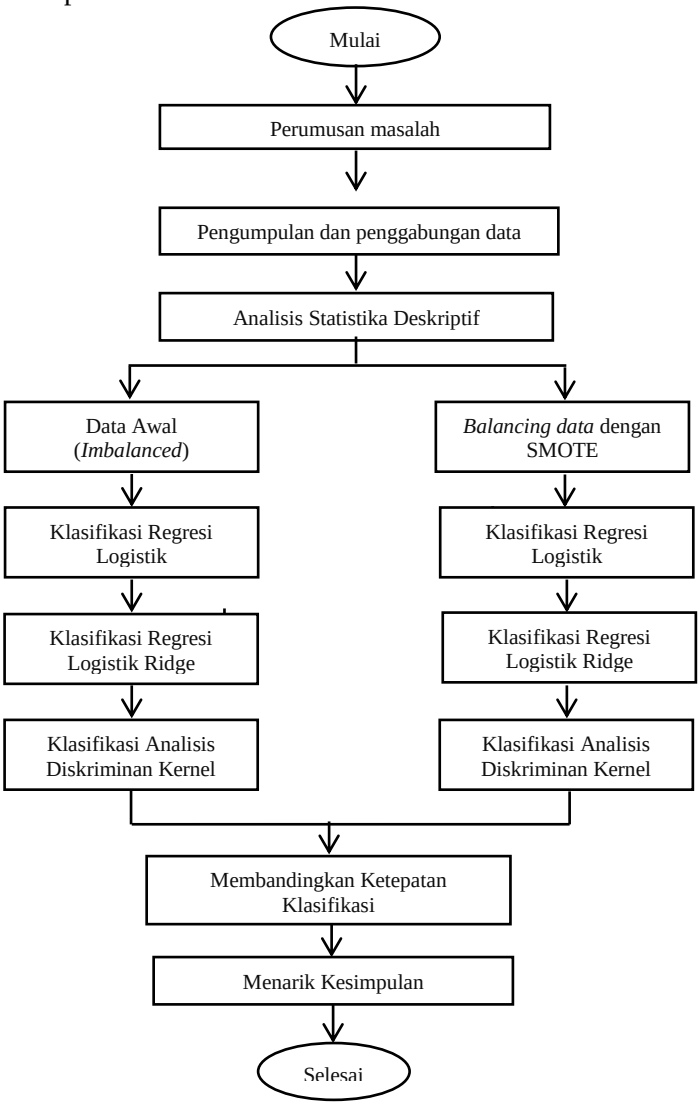
c. Melakukan klasifikasi data *imbalanced* dengan Analisis Diskriminan.
i. Membagi data menjadi data *training* dan *testing* dengan menggunakan 10-*fold cross validation* menggunakan metode stratifikasi.

ii. Melakukan uji asumsi yang meliputi uji asumsi multivariat normal (Bab 2.7a), uji asumsi homogenitas (Bab 2.7b).
iii. Jika seluruh asumsi dipenuhi, maka status desa tertinggal diklasifikasikan menggunakan Analisis Diskriminan Linier. Jika seluruh asumsi tidak dipenuhi, maka dilakukan klasifikasi data *imbalanced* menggunakan Analisis Diskriminan Kernel Gaussian *Radial Basis Function* (RBF).
iv. Melakukan klasifikasi data *training* menggunakan Analisis Diskriminan Kernel.
v. Menghitung ketepatan klasifikasi pada data *testing* dengan persamaan (2.41) sampai persamaan (2.45).
vi. Menghitung ketepatan klasifikasi menggunakan variabel signifikan yang diperoleh pada langkah 3.a.(vi).

3. Menganalisis ketepatan klasifikasi data *balanced* menggunakan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan 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%) seperti persamaan (2.1)
      iii. Menentukan data *k-Nearest Neighbour* (x_{knn}) dengan jarak terdekat dari setiap data minoritas yang akan disintesis menggunakan Persamaan (2.2).
   b. Melakukan klasifikasi data *balanced* dengan Regresi Logistik seperti langkah 3.a.
   c. Melakukan klasifikasi data *balanced* dengan Regresi Logistik Ridge seperti langkah 3.b.
   d. Melakukan klasifikasi data *balanced* dengan Analisis Diskriminan Kernel seperti langkah 3.c.

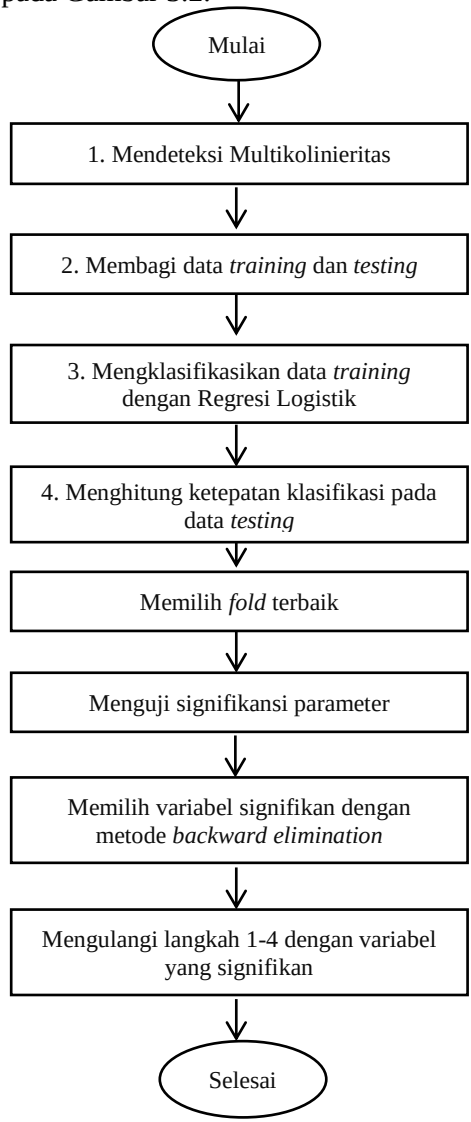
4. Membandingkan nilai ketepatan klasifikasi data *imbalanced* dan data *balanced*.

Gambar 3.1 adalah visualisasi langkah analisis (diagram alir) dalam penelitian ini.



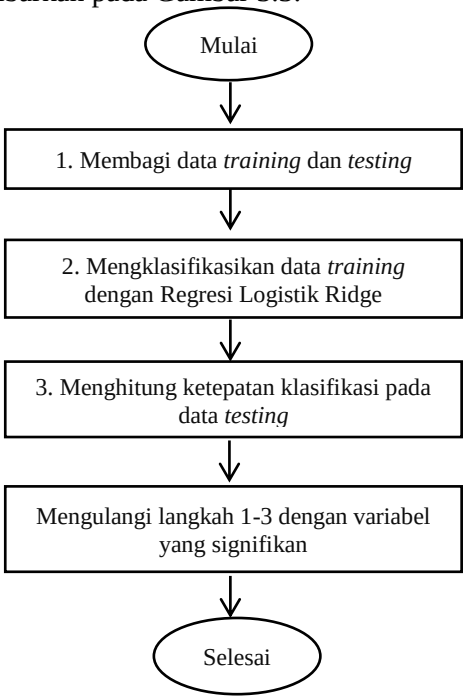
Gambar 3. 1 Diagram Alir Penelitian

Diagram alir untuk proses klasifikasi Regresi Logistik digambarkan pada Gambar 3.2.



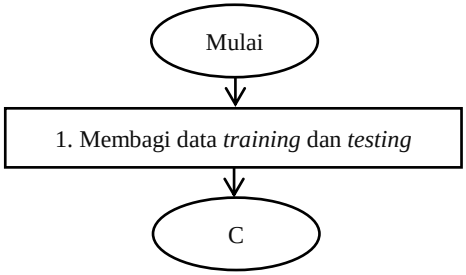
Gambar 3. 2 Diagram Alir Klasifikasi Metode Regresi Logistik

Diagram alir untuk proses klasifikasi Regresi Logistik Ridge digambarkan pada Gambar 3.3.

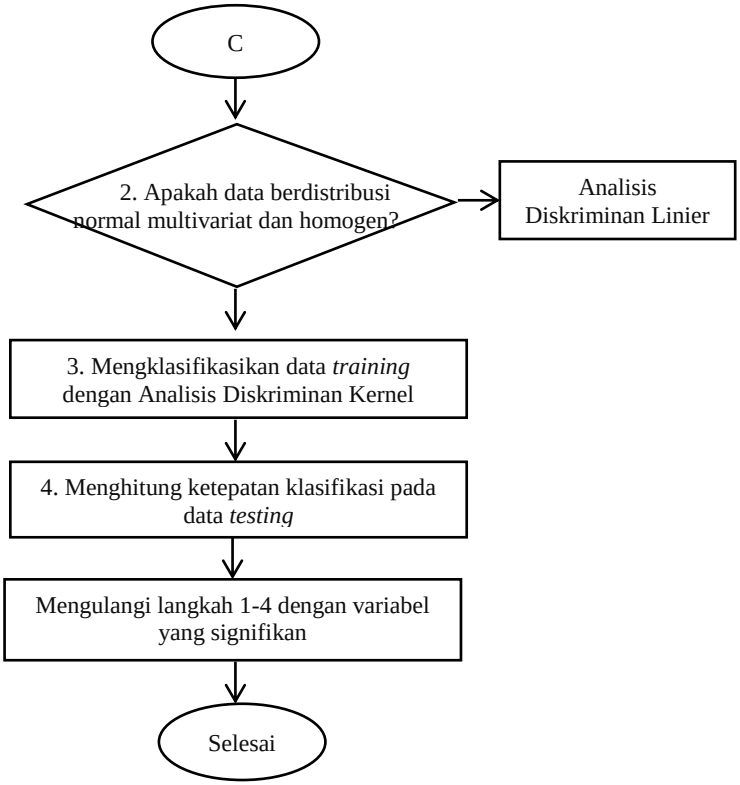


Gambar 3. 3 Diagram Alir Klasifikasi Metode Regresi Logistik Ridge

Diagram alir untuk proses klasifikasi Analisis Diskriminan Kernel digambarkan pada Gambar 3.4.



Gambar 3. 4 Diagram Alir Klasifikasi Metode Analisis Diskriminan Kernel



Gambar 3.4 Diagram Alir Klasifikasi Metode Analisis Diskriminan Kernel (Lanjutan)

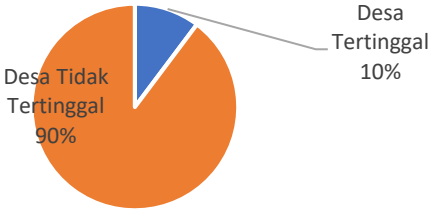
(Halaman ini sengaja dikosongkan)

BAB IV
ANALISIS DAN PEMBAHASAN

Analisa dan pembahasan pada bab ini mencakup karakteristik variabel yang diduga mempengaruhi status ketertinggalan desa di Jawa Timur berdasarkan data Potensi Desa 2014. Setelah mengetahui karakteristik desa tertinggal, dilakukan klasifikasi menggunakan metode Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel dan menghitung nilai ketepatan klasifikasi yang dihasilkan.

4.1 Karakteristik Desa 5 Kabupaten di Jawa Timur

Data yang digunakan analisis selanjutnya adalah kabupaten Bondowoso, kabupaten Situbondo, kabupaten Sumenep, kabupaten Bangkalan, dan kabupaten Sampang dengan jumlah desa tertinggal dan tidak tertinggal sebanyak 115 dan 1007 desa. Perbandingan desa tertinggal dengan desa tidak tertinggal di 5 kabupaten tersebut sebesar 10%:89,75%.



Gambar 4. 1 Proporsi Kelas Status Ketertinggalan Desa

Setiap desa memiliki kualitas pelayanan dasar yang berbeda-beda. Karakteristik desa berdasarkan jumlah SD, Poskesdes, tempat praktik bidan, jumlah warga gizi buruk jarak kantor kepala desa ke kantor kecamatan disajikan pada Tabel 4.1.

Tabel 4.1 Karakteristik Pelayanan Dasar Desa Menurut Kelompok

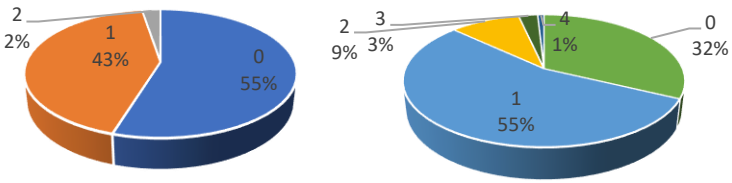
Kelompok Desa Tidak Tertinggal (Kelas: 0)					
Variabel	**Mean**	**StDev**	**Min**	**Maks**	**Modus**
SD/MI	3,89	2,85	0	27	2
Poskesdes	0,60	0,61	0	11	1
Praktik Bidan	1,21	1,11	0	16	1
Gizi Buruk	0,849	3,898	0	98	0
Jarak	5,34	7,54	1	197	1

Tabel 4.1 Karakteristik Pelayanan Dasar Desa Menurut Kelompok (Lanjutan)

Kelompok Desa Tertinggal (Kelas: 1)					
Variabel	**Mean**	**StDev**	**Min**	**Maks**	**Modus**
SD/MI	3,17	1,81	1	9	2
Poskesdes	0,48	0,55	0	2	0
Praktik Bidan	0,85	0,76	0	4	1
Gizi Buruk	0,583	1,821	0	16	0
Jarak	10,86	15,65	1	164	5

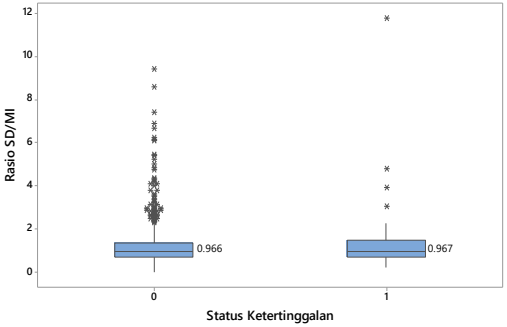
Pada kelompok desa tidak tertinggal, dari 1.007 desa rata-rata jumlah SD/MI setiap desa adalah 4 sekolah. Jumlah desa dengan jumlah SD terbanyak ada di desa Tlambah Kabupaten Sampang. Banyak jumlah SD setiap desa yang sering muncul pada kelompok desa tidak tertinggal adalah 2 sekolah. Jumlah pos kesehatan desa tiap desa memiliki kemiripan jumlah yang sama. Terlihat dari nilai standar deviasi yang kecil yaitu 0,61. Paling banyak desa dengan sarana poskesdes sebanyak 1. Jarak terdekat kantor kepala desa dengan kantor kecamatan adalah 1 km, sedangkan jarak terjauh adalah 197 km yang berada di desa Karamian pulau Karamian Kabupaten Sumenep. Desa dengan gizi buruk terbanyak terdapat di desa Sapeken Kabupaten Sumenep.

Kelompok desa tertinggal di 5 Kabupaten Jawa Timur sebanyak 115 desa. Jumlah sekolah dasar yang ada di desa minimal 1 sekolah. Banyak desa yang tidak memiliki sarana poskesdes. Untuk tempat pos kesehatan desa terdapat 63 desa yang tidak memiliki sarana kesehatan tersebut. Desa yang tidak memiliki tempat praktik bidan sebanyak 37 desa, sedangkan yang memiliki 1 tempat praktik bidan sebanyak 63 desa dan yang memiliki lebih dari 1 tempat praktik bidan sebanyak 15 desa. Desa pada kelompok tertinggal dengan jarak kantor desa dengan kantor kecamatan terjauh adalah desa Masakambing yang terdapat di kabupaten Sumenep. Jumlah warga penderita gizi buruk pada kelompok desa tertinggal terbanyak terdapat di desa Kembangsari kabupaten Situbondo. Berikut ini adalah persentase poskesdes dan persentase keberadaan tempat praktik bidan pada desa tertinggal di 5 Kabupaten:



Gambar 4.2 Persentase Poskesdes (kiri) dan Tempat Praktik Bidan Kelompok Desa Tertinggal di 5 Kabupaten

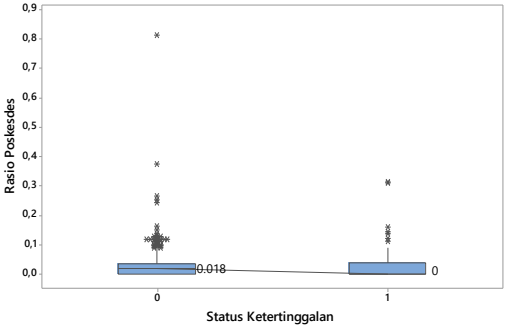
Selain menggunakan tabel, deskripsi status desa tertinggal berdasarkan nilai rasio dapat dilihat melalui boxplot. Nilai yang digunakan berskala rasio, karena pelayanan yang ada di desa juga mempertimbangkan jumlah penduduk yang berdomisili di desa tersebut. Berikut ini adalah *boxplot* dari rasio Sekolah Dasar/MI:



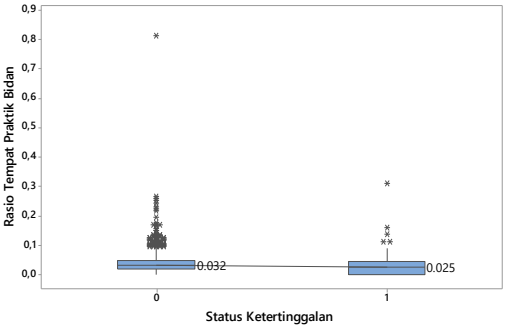
Gambar 4.3 Boxplot Rasio SD/MI

Sekolah Dasar atau Madrasah Ibtidaiyah merupakan sarana penting dalam menunjang pendidikan anak. Berdasarkan Gambar 4.3 tidak terdapat perbedaan antara rasio SD/MI di kelompok desa tertinggal maupun desa tidak tertinggal. Terlihat dari nilai median kedua kelompok tersebut hampir sama yaitu 0,966 dan 0,967. Interval rasio pada kelompok desa tertinggal lebih pendek daripada kelompok desa tidak tertinggal. Keragaman atau variasi rasio SD/MI tertinggi terjadi pada kelompok desa tertinggal, dikarenakan luas *box* yang lebih lebar.

Selain itu, dari segi pelayanan kesehatan terdapat rasio Poskesdes dan rasio tempat praktik bidan terhadap jumlah penduduk. Berikut ini adalah *boxplot* dari kedua rasio.



(a)

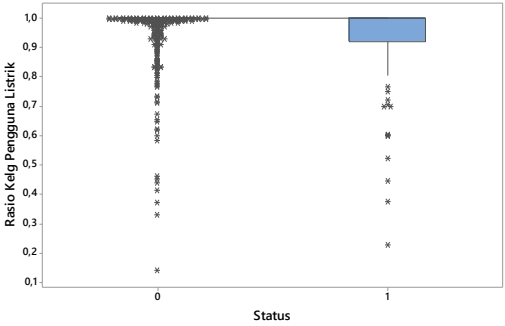


(b)

Gambar 4.4 Boxplot Rasio Poskesdes (a) dan Tempat Praktik Bidan (b)

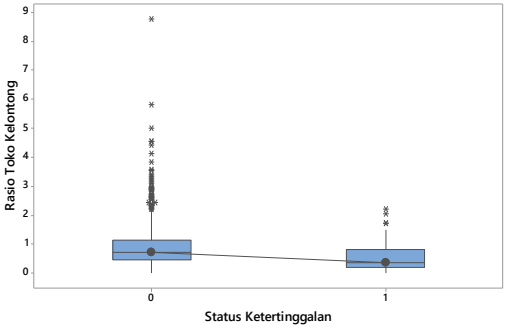
Gambar 4.4 (a) menunjukkan rasio Poskesdes. Terdapat perbedaan median rasio antara kelompok desa tertinggal dan kelompok desa tidak tertinggal yaitu 0 dengan 0,018. Rasio poskesdes di kelompok desa tertinggal cenderung lebih rendah. Ini menunjukkan masih banyak desa yang tidak memiliki pelayanan kesehatan poskesdes, terlebih pada kelompok desa tertinggal. Sedangkan pada rasio tempat praktik bidan, nilai tengah pada kedua kelompok hampir sama, yaitu 0,032 dan 0,025. Pada kedua rasio terdapat nilai *extreme* yang terdapat pada kelompok desa tidak tertinggal.

Salah satu fasilitas infrastruktur yang harus ada adalah tersedianya jaringan listrik pada desa. Listrik sangat diperlukan guna menunjang kehidupan. Berikut ini adalah rasio keluarga pengguna listrik:



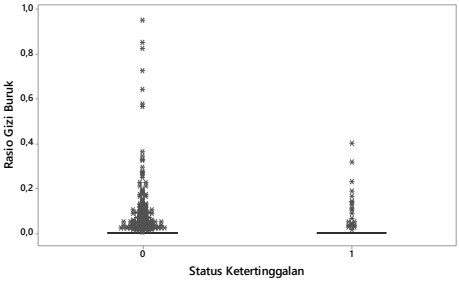
Gambar 4.5 Boxplot Rasio Keluarga Pengguna Listrik

Untuk infrastruktur ketersediaan listrik tidak ada perbedaan pada median rasio keluarga pengguna listrik. Hal ini menunjukkan bahwa fasilitas listrik sudah menjangkau pada sebagian besar masyarakat. Walaupun terdapat beberapa desa yang memiliki persentase kurang dari 40% masih terdapat di desa tertinggal dan ada yang terendah dari desa tidak tertinggal. Infrastruktur lain yang selalu ada di setiap desa adalah adanya toko kelontong yang mendukung perkonomian desa.



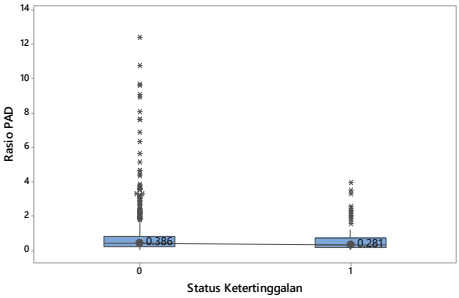
Gambar 4.6 Boxplot Rasio Toko Kelontong

Terlihat dari Gambar 4.6, rasio toko kelontong terhadap jumlah penduduk desa pada kelompok desa tertinggal cenderung lebih kecil. Pada kedua kelompok tidak simeteris (miring) ke arah kanan (*skewness* positif), dikarenakan median tidak berada di tengah box, adanya *outlier* di bagian atas boxplot serta Panjang *whisker* (garis) bagian atas yang lebih panjang.



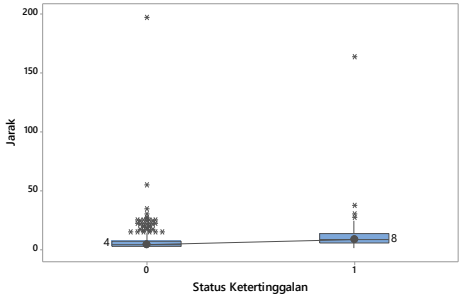
Gambar 4.7 Boxplot Rasio Gizi Buruk

Median rasio gizi buruk pada kedua kelompok adalah sama yaitu 0, menunjukkan 50% dari pengamatan berada di rasio 0. Lebar box nya tidak terlihat, yang berarti rasio gizi buruk tidak menyebar. Dari segi penyelenggaraan pemerintah, terdapat rasio Pendapatan Asli Desa (PAD) terhadap jumlah penduduk dan jarak kantor desa ke kantor kecamatan (km).



Gambar 4.8 Boxplot Rasio Pendapatan Asli Desa

Interval rasio PAD pada kelompok desa tidak tertinggal lebih lebar daripada kelompok desa tertinggal. Median rasio kelompok desa tertinggal juga lebih rendah yaitu 0,281 dibandingkan median kelompok desa tidak tertinggal yang berada di 0,386. Sedangkan pada jarak kantor desa ke kantor kecamatan, median kelompok desa tertinggal lebih tinggi daripada kelompok desa tidak tertinggal yaitu 8 dan 4. Pada desa kelompok tidak tertinggal terdapat banyak nilai *outlier* dan terdapat satu titik *extreme*, sedangkan kelompok desa tertinggal terdapat 3 *outlier* dan 1 titik *extreme*.



Gambar 4.9 Boxplot Jarak Kantor Desa ke Kantor Kecamatan

Ringkasan karakteristik desa tertinggal 5 Kabupaten di Jawa Timur secara lengkap pada Tabel 4.2:

Tabel 4.2 Karakteristik Desa Menurut Variabel Penelitian

Nama Variabel	Karakteristik
Rasio sekolah dasar terhadap total murid (X_1)	Median kelompok desa tertinggal dan kelompok tidak tertinggal **sejajar**
Rasio tempat pos kesehatan desa (poskesdes) terhadap total penduduk (X_2)	Median kelompok desa tertinggal **lebih rendah** daripada kelompok tidak tertinggal sejajar
Rasio tempat praktik bidan terhadap total penduduk (X_3)	Median kelompok desa tertinggal **lebih rendah** daripada kelompok tidak tertinggal sejajar
Rasio keluarga pakai listrik terhadap total keluarga (X_4)	Median kelompok desa tertinggal dan kelompok tidak tertinggal **sejajar**
Rasio toko kelontong terhadap total penduduk (X_5)	Median kelompok desa tertinggal **lebih rendah** daripada kelompok tidak tertinggal sejajar
Jarak tempuh dari kantor kepala desa ke kantor kecamatan (km) (X_6)	Median kelompok desa tertinggal **lebih tinggi** daripada kelompok tidak tertinggal sejajar
Rasio jumlah warga penderita gizi buruk terhadap total penduduk (X_7)	Median kelompok desa tertinggal dan kelompok tidak tertinggal **sejajar**
Rasio Pendapatan Asli Desa (PAD) terhadap total penduduk (X_8)	Median kelompok desa tertinggal **lebih rendah** daripada kelompok tidak tertinggal sejajar

4.2 Klasifikasi pada Data *Imbalanced*

Klasifikasi pada data *imbalanced* menggunakan tiga metode, yaitu Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel serta masing-masing metode dilakukan klasifikasi dengan seluruh variabel dan variabel yang signifikan.

4.2.1 Klasifikasi dengan Regresi Logistik pada Data *Imbalanced*

Metode pertama yang akan digunakan adalah Regresi Logistik. Pengecekan awal sebelum dilakukan proses klasifikasi adalah mendeteksi adanya multikolinieritas dalam model. Selain itu, akan dilakukan klasifikasi pada seluruh variabel dan pada variabel yang signifikan berdasarkan analisis *backward elimination*.

A. Regresi Logistik dengan Semua Variabel

Dalam menganalisis data, perlu dilakukan deteksi multikolinieritas untuk melihat hubungan antar variabel. Salah satu ukuran yang dapat digunakan untuk menguji adanya multikolinieritas pada regresi adalah *Variance Inflation Factor* (VIF). Adanya multikolinieritas dinilai dari nilai VIF yang dihasilkan. Nilai VIF yang melebihi 2,5 akan menunjukkan adanya multikolinieritas dan sebaliknya. Berikut ini adalah hasil dari VIF dari data *imbalanced*.

Tabel 4.3 Nilai VIF Variabel Bebas Data *Imbalanced*

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1,01	1,51	1,54	1,08	1,04	1,06	1,01	1,02

Dari Tabel 4.3, dapat dilihat bahwa tidak terdapat variabel bebas yang memiliki angka VIF lebih besar dari 5. Nilai VIF tertinggi terdapat pada variabel rasio tempat praktik bidan. Oleh karena itu, pada data *imbalanced* tidak terdeteksi adanya multikolinieritas.

Setelah melakukan pengecekan multikolinearitas pada data *imbalanced*, selanjutnya dilakukan pengklasifikasian status ketertinggalan desa di 5 Kabupaten Jawa Timur menggunakan *stratified 10 fold cross validation*. Variabel yang digunakan adalah seluruh variabel prediktor. Berikut ini adalah rincian hasil ketepatan klasifikasi pada data *imbalanced*.

Tabel 4.4 Ketepatan Klasifikasi Regresi Logistik Data *Imbalanced* dengan Semua Variabel

Fold	Training					Testing				
	G-mean	AUC	AT	Sens	Spes	G-mean	AUC	AT	Sens	Spes
1	0,20	0,52	0,90	0,04	1,00	0,29	0,54	0,90	0,08	1,00
2	0,22	0,52	0,90	0,05	0,99	0,29	0,54	0,90	0,08	1,00
3	0,26	0,53	0,90	0,07	1,00	0,00	0,50	0,89	0,00	0,99
4	0,24	0,53	0,90	0,06	1,00	0,42	0,58	0,90	0,18	0,98
5	0,24	0,53	0,90	0,06	0,99	0,00	0,50	0,90	0,00	1,00
6	0,29	0,54	0,90	0,09	0,99	0,00	0,50	0,89	0,00	0,99
7	0,24	0,53	0,90	0,06	0,99	0,00	0,50	0,90	0,00	1,00
8	0,24	0,53	0,90	0,06	0,99	0,00	0,50	0,89	0,00	1,00
9	0,22	0,52	0,90	0,05	1,00	0,00	0,50	0,89	0,00	1,00
10	0,26	0,53	0,90	0,07	0,99	0,29	0,54	0,89	0,08	0,99
Mean	0,24	0,53	0,90	0,06	0,99	0,13	0,52	0,90	0,04	1,00
St-dev	0,03	0,01	0,00	0,01	0,00	0,16	0,03	0,00	0,06	0,01

Dapat diketahui bahwa model terbaik yang didapat adalah model pada *fold* ke-4, dimana nilai G-*mean*, akurasi total (AT), sensitivitas (sens) dan spesifisitas (spes) yang dihasilkan pada data *testing* paling tinggi dibandingkan *fold* yang lain. Berdasarkan Tabel 4.3 diketahui nilai rata-rata akurasi total pada data *training* dan *testing* menghasilkan ketepatan yang bagus yaitu sama-sama 90%. Tetapi rata-rata nilai G-*mean* sangat kecil karena nilai G-mean pada beberapa *fold* di data *testing* banyak bernilai 0 karena model regresi logistik pada *fold* tersebut tidak dapat mengklasifikasikan pada kelas positif (minoritas) dimana nilai sensitivitas menunjukkan nilai yang rendah. Rata-rata nilai AUC sebesar 0,52, ini menunjukkan metode Regresi Logistik pada data *imbalanced* termasuk kategori kebaikan model klasifikasi yang salah (*failure*). Nilai rata-rata spesifisitas pada data *training* dan *testing* menunjukkan nilai 1, berarti klasifikasi dengan metode Regresi Logistik menghasilkan ketepatan akurasi pada kelas negatif sebesar 100%. Standar deviasi antar *fold* menghasilkan nilai kurang dari 1, ini menunjukkan sebaran nilai terhadap rata-rata sedikit.

Berdasarkan Tabel 4.4, model terbaik untuk memprediksi data status desa tertinggal terdapat di *fold* ke-4. Oleh karena itu akan dilakukan pengujian signifikansi parameter pada *fold* ke-4. Pengujian pertama adalah uji serentak dengan *Likelihood Ratio*

Test. Dari hasil uji serentak didapatkan nilai G sebesar 594,63 dengan df =9, tingkat kesalahan (α) 10%, serta nilai kritis sebesar 13,36. Dapat diambil keputusan untuk menolak H_0 karena nilai G > nilai kritis. Artinya, model pada *fold* ke-4 terdapat minimal 1 variabel yang berpengaruh signifikan terhadap model.

Pada pengujian parameter secara serentak diputuskan tolak H_0 sehingga dapat dilanjutkan pengujian secara parsial untuk mengetahui variabel mana saja yang tidak berpengaruh signifikan terhadap model. Berikut ini adalah hasil pengujian secara parsial:

Tabel 4.5 Hasil Uji Parsial Data *Imbalanced*

Variabel	Koefisien	S.E	Z*	*p-value*
Konstanta	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

dengan menggunakan tingkat kesalahan (α) sebesar 10% didapat nilai kritis ($Z_{\alpha/2}$) sebesar 1,645. Variabel dengan nilai |Z*| lebih besar daripada nilai kritis atau nilai *p-value* kurang dari α akan diambil keputusan tolak H_0 yang berarti variabel tersebut berpengaruh signifikan terhadap pembentukan model. Variabel yang berpengaruh signifikan terhadap pemodelan data *imbalanced* adalah variabel rasio keluarga pakai listrik, rasio toko kelontong, dan jarak tempuh kantor kepala desa ke kantor kecamatan.

Berdasarkan pengujian parameter secara parsial diketahui terdapat variabel yang signifikan. Maka akan dilakukan pemilihan variabel pada data *imbalanced* menggunakan *backward elimination* dengan cara mengeliminasi variabel yang paling tidak signifikan secara bertahap.

Pada Tabel 4.5, dapat diketahui bahwa variabel yang berpengaruh signifikan dalam pengelompokkan status desa tertinggal pada data *imbalanced* adalah variabel rasio keluarga

pakai listrik, rasio toko kelontong, dan jarak tempuh kantor kepala desa ke kantor kecamatan. Untuk selanjutnya variabel tersebut akan digunakan pengklasifikasian menggunakan Regresi Logistik Ridge dan Analisis Diskriminan Kernel dengan variabel yang signifikan.

Hasil akhir yang didapat dari *backward elimination* adalah sebagai berikut:

Tabel 4.6 Hasil *Backward Elimination* Data *Imbalanced*

Variabel	Koefisien	S.E	Z*	*p-value*
Konstanta	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

B. Regresi Logistik dengan Variabel Signifikan

Setelah dilakukan pemilihan variabel yang signifikan, selanjutnya akan dilakukan deteksi multikolinieritas dan dilakukan pengklasifikasian metode Regresi Logistik dengan variabel yang signifikan. Salah satu ukuran yang dapat digunakan untuk mendeteksi adanya multikolinieritas pada regresi adalah *Variance Inflation Factor* (VIF). Berikut ini adalah hasil dari VIF dari data *imbalanced* dengan variabel signifikan.

Tabel 4.7 Nilai VIF Variabel Bebas Data *Imbalanced* Variabel Signifikan

X_4	X_5	X_6
1,07	1,02	1,05

Tabel 4.7, dapat dilihat bahwa tidak terdapat variabel bebas yang memiliki angka VIF lebih besar dari 5. Nilai VIF tertinggi terdapat pada variabel rasio keluarga pengguna listrik. Oleh karena itu, pada data *imbalanced* menggunakan variabel signifikan tidak terdeteksi adanya multikolinieritas.

Selanjutnya akan dilakukan pengklasifikasian variabel yang signifikan dengan metode Regresi Logistik. Berdasarkan Tabel 4.8 nilai rata-rata G-*mean* pada data testing yang kecil yaitu 10%, ketepatan akurasi total sebesar 89,7%. Nilai G-mean lebih kecil daripada akurasi total dikarenakan G-mean memperhitungkan ketepatan klasifikasi data kelas positif dan kelas negatif. Titik sebaran nilai ketepatan klasifikasi tidak jauh

dari rata-ratanya, menunjukkan tahapan *stratified* 10-*fold* CV memberikan pembagian data yang baik.

Tabel 4.8 Rata-rata Ketepatan Klasifikasi Regresi Logistik Data *Imbalanced* dengan Variabel Signifikan

	Training	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,522	0,010	0,515	0,027
G-*mean*	0,216	0,047	0,100	0,156
Akurasi Total	0,898	0,002	0,897	0,005
Sensitivitas	0,049	0,021	0,035	0,059
Spesifisitas	0,995	0,002	0,995	0,007

Nilai AUC yang dihasilkan pada klasifikasi Regresi Logistik data *imbalanced* dengan variabel yang signifikan juga memberikan kebaikan model klasifikasi yang salah. Jika dibandingkan dengan klasifikasi Regresi Logistik sebelumnya (Tabel 4.3) diketahui nilai ketepatan klasifikasi dengan seluruh variabel lebih baik daripada klasifikasi dengan variabel yang signifikan, tetapi nilai ketepatan yang dihasilkan hampir sama dan tidak berbeda jauh.

4.2.2 Klasifikasi Regresi Logistik Ridge pada Data *Imbalanced*

Metode klasifikasi Regresi Logistik Ridge dilakukan sebagai pembanding metode Regresi Logistik dimana penambahan parameter *ridge* dapat mengakomodasi adanya multikolinieritas dalam data dan *robust* terhadap Regresi Logistik, nilai lambda secara otomatis dipilih oleh *package* (Lampiran 5. J).

Tabel 4.9 menunjukkan rata-rata ketepatan klasifikasi dari data *imbalanced* menggunakan Regresi Logistik Ridge dengan partisi data *stratified* 10-*fold* CV pada seluruh variabel dan pada variabel yang signifikan. Rata-rata nilai G-*mean* data *testing* baik pengklasifikasian dengan seluruh variabel maupun variabel yang signifikan memberikan nilai sangat kecil yaitu 10%. Akurasi menggunakan Regresi Logistik Ridge sangat tinggi yaitu 89,8% dibandingkan nilai G-mean. Nilai akurasi pada kelas negatif (spesifisitas) lebih besar daripada akurasi kelas positif (sensitivitas). Nilai AUC pada klasifikasi Regresi Logistik Ridge baik dengan seluruh variabel maupun variabel signifikan memberikan nilai rata-rata sebesar 0,515,

menunjukkan klasifikasi dengan Regresi Logistik Ridge pada data *imbalanced* memberikan kebaikan klasifikasi yang salah.

Tabel 4.9 Rata-rata Ketepatan Klasifikasi Regresi Logistik Ridge Data *Imbalanced*

Seluruh Variabel	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,506	0,005	0,515	0,027
G-*mean*	0,105	0,059	0,100	0,156
Akurasi Total	0,896	0,001	0,898	0,004
Sensitivitas	0,014	0,011	0,035	0,059
Spesifisitas	0,997	0,001	0,996	0,007
Variabel Signifikan	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,512	0,009	0,515	0,028
G-mean	0,141	0,065	0,100	0,156
Akurasi Total	0,897	0,001	0,898	0,005
Sensitivitas	0,026	0,016	0,035	0,059
Spesifisitas	0,996	0,001	0,996	0,007

4.2.3 Klasifikasi Analisis Diskriminan Kernel pada Data *Imbalanced*

Berikut ini akan dibahas tentang klasifikasi desa tertinggal pada data *imbalanced* menggunakan metode Analisis Diskriminan Kernel dengan fungsi kernel *Radial Basis Function* (RBF). Pengklasifikasian dengan Analisis Diskriminan Kernel dirujuk dari hasil pengujian distribusi normal multivariat (Lampiran 6. A) dan uji homogenitas (Lampiran 6. B) yang menunjukkan data tidak memenuhi asumsi dalam Analisis Diskriminan Linier.

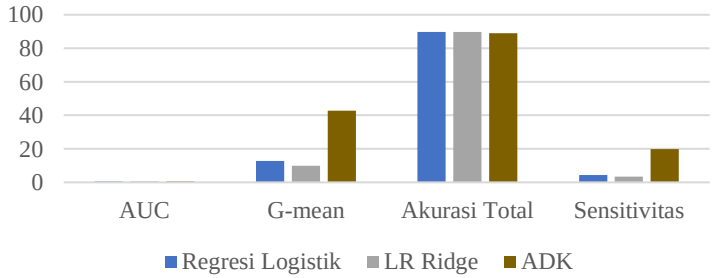
Tabel 4.10 Rata-rata Ketepatan Klasifikasi Analisis Diskriminan Kernel Data *Imbalanced*

Seluruh Variabel	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,604	0,009	0,584	0,042
G-*mean*	0,479	0,020	0,428	0,099
Akurasi Total	0,895	0,002	0,890	0,015
Sensitivitas	0,237	0,020	0,199	0,085
Spesifisitas	0,970	0,003	0,969	0,018
Variabel Signifikan	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,603	0,017	0,584	0,050
G-mean	0,479	0,020	0,428	0,099
Akurasi Total	0,895	0,002	0,891	0,016
Sensitivitas	0,237	0,020	0,199	0,085
Spesifisitas	0,970	0,003	0,969	0,018

Berdasarkan Tabel 4.10 diketahui rata-rata nilai AUC pada data *testing* masih dibawah 0,6. Rata-rata nilai G-*mean* menghasilkan ketepatan sebesar 42,8%. Akurasi total menunjukkan nilai yang tinggi yaitu 89%. Hasil sensitivitas pada analisis diskriminan kernel pada data *imbalanced* juga menghasilkan nilai yang kecil yaitu 20%. Berbanding terbalik dengan akurasi kelas negatif yang ditunjukkan oleh nilai spesifisitas yang sangat tinggi. Ini membuktikan bahwa pada data *imbalanced* memberikan ketepatan klasifikasi yang cenderung mengklasifikasikan ke kelas mayoritas.

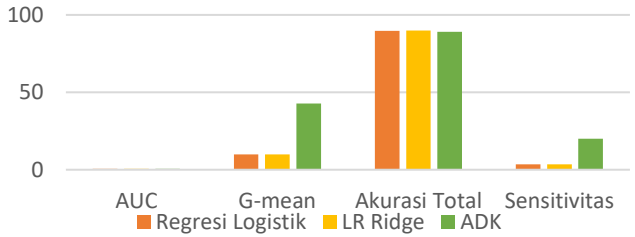
4.2.4 Analisis Ketepatan Klasifikasi Data *Imbalanced*

Telah dilakukan klasifikasi data *imbalanced* menggunakan tiga metode klasifikasi dengan menggunakan seluruh variabel dan variabel yang signifikan. Berikut ini akan dianalisis perbandingan rata-rata ketepatan klasifikasi pada data *testing* dengan menggunakan seluruh variabel.



Gambar 4.10 Rata-rata Ketepatan Klasifikasi Data *Imbalanced* dengan Seluruh Variabel

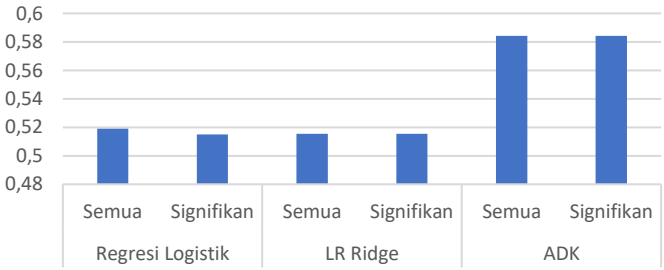
Rata-rata ketepatan klasifikasi G-mean data *imbalanced* dengan seluruh variabel memberikan hasil yang sangat kecil. Dengan Regresi Logistik Ridge memberikan nilai G-mean paling kecil, sedangkan tertinggi dihasilkan oleh metode Analisis Diskriminan Kernel (ADK). Nilai akurasi total dan spesifisitas klasifikasi pada masing-masing metode memberikan hasil yang bagus. Namun, tidak diikuti dengan ketepatan nilai sensitivitas. Sehingga dapat dikatakan pada data *imbalanced* memberikan ketepatan klasifikasi yang semu karena cenderung mengklasifikasikan pada kelas mayoritas.



Gambar 4.11 Rata-rata Ketepatan Klasifikasi Data *Imbalanced* dengan Variabel Signifikan

Klasifikasi menggunakan variabel yang signifikan (Gambar 4.11) juga memberikan hasil yang tidak berbeda jauh dengan menggunakan seluruh variabel. Terjadi sedikit peningkatan ketepatan klasifikasi nilai akurasi total pada metode Regresi Logistik Ridge. Secara keseluruhan, klasifikasi data *imbalanced* pada kasus status ketertinggalan desa 5 Kabupaten di Jawa Timur terbaik dihasilkan pada metode Analisis Diskriminan Kernel.

Viasualiasi AUC (Gambar 4.12) menunjukkan klasidikasi pada data *imbalanced* dengan menggunkan klasifikasi Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel didapatkan ketiga metode tersebut tergolong kategori kebaikan klasifikasi yang salah, dikarenakan nilai AUC yang dibawah 0,6.



Gambar 4.12 Rata-rata AUC pada Data *Imbalanced*

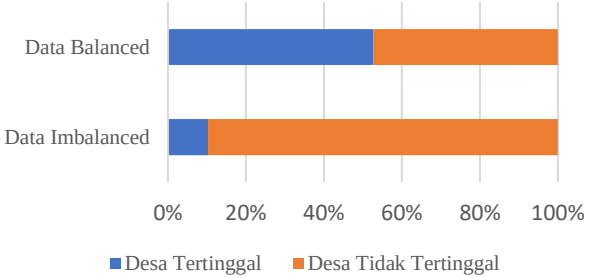
4.3 Klasifikasi pada Data *Balanced*

Data status desa tertinggal akan dilakukan penambahan observasi pada kelas minoritas dengan pendekatan SMOTE. Setelah data *balanced*, dilakukan analisis klasifikasi

menggunakan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel.

4.3.1 Penerapan Metode *Synthetic Minority Oversampling Technique*

Jumlah data *imbalanced* di 5 Kabupaten terdapat 1.122 data dengan perbandingan kelompok desa tertinggal dan tidak tertinggal sebesar 1:8,756. Penerapan metode *oversampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE) bekerja dengan membuat replikasi dari data minoritas. Berikut ini adalah hasil replikasi SMOTE.



Gambar 4.13 Proporsi Kelompok Desa

Berdasarkan Gambar 4.13, diketahui proporsi kelompok desa tertinggal dan desa tidak tertinggal setelah dilakukan *balancing* dengan SMOTE dengan persentase replikasi *oversampling* sebanyak 900% dengan k=5 menghasilkan proporsi kelompok yang hampir sama yaitu 53% : 47% dengan jumlah observasi sebanyak 2.185. Jumlah desa tertinggal yang awalnya 115 desa direplikasi menjadi 1.150 desa. Sedangkan jumlah desa tidak tertinggal menjadi 1.035 desa.

4.3.2 Klasifikasi Regresi Logistik pada Data *Balanced*

Berikut ini adalah pembahasan perhitungan performa klasifikasi berdasarkan klasifikasi pada data *balanced* menggunakan metode Regresi Logistik. Data *balanced* didapatkan dengan penerapan metode SMOTE. Pertama, dilakukan pendeteksian multikolinieritas antar variabel prediktor. Selain itu, akan dilakukan klasifikasi pada seluruh variabel dan pada variabel yang signifikan berdasarkan analisis *backward elimination*.

A. Regresi Logistik dengan Seluruh Variabel

Mendeteksi multikolinieritas dengan melihat nilai VIF yang didapat dari analisis regresi. Berikut ini adalah hasil dari VIF setiap variabel dari data *balanced*.

Tabel 4.11 Nilai VIF Variabel Bebas Data *Balanced* Seluruh Variabel

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1,01	1,30	1,35	1,15	1,09	1,07	1,02	1,06

Dari Tabel 4.11, dapat dilihat bahwa tidak terdapat variabel prediktor yang memiliki angka VIF lebih besar dari 5. Oleh karena itu, pada data *balanced* tidak terdeteksi adanya kasus multikolinieritas.

Setelah melakukan pengecekan multikolinearitas pada data *balanced*, selanjutnya dilakukan pengklasifikasian status ketertinggalan desa di 5 Kabupaten Jawa Timur menggunakan *stratified 10 fold cross validation*. Variabel yang digunakan adalah seluruh variabel prediktor. Berikut ini adalah rincian hasil ketepatan klasifikasi pada data *balanced*.

Tabel 4.12 Ketepatan Klasifikasi Regresi Logistik Data *Balanced* dengan Semua Variabel

Fold	Training					Testing				
	AUC	G-mean	AT	Sens	Spes	AUC	G-mean	AT	Sens	Spes
1	0,76	0,76	0,76	0,75	0,77	0,78	0,78	0,78	0,74	0,82
2	0,76	0,76	0,76	0,76	0,76	0,78	0,78	0,78	0,78	0,78
3	0,76	0,76	0,76	0,76	0,76	0,80	0,80	0,79	0,78	0,81
4	0,77	0,77	0,77	0,77	0,77	0,76	0,76	0,77	0,82	0,71
5	0,77	0,77	0,77	0,77	0,77	0,74	0,74	0,74	0,77	0,70
6	0,76	0,76	0,76	0,76	0,77	0,76	0,76	0,76	0,76	0,77
7	0,77	0,77	0,77	0,76	0,77	0,77	0,77	0,77	0,76	0,79
8	0,77	0,77	0,77	0,76	0,77	0,72	0,72	0,72	0,67	0,77
9	0,77	0,77	0,77	0,77	0,77	0,75	0,75	0,75	0,74	0,76
10	0,76	0,76	0,76	0,76	0,76	0,79	0,78	0,78	0,77	0,80
Mean	0,77	0,77	0,76	0,76	0,77	0,76	0,76	0,76	0,76	0,77
St-dev	0,00	0,00	0,00	0,00	0,00	0,02	0,02	0,02	0,04	0,04

Berdasarkan Tabel 4.12 diketahui rata-rata nilai AUC pada data training dan data testing menunjukkan hasil sekitar 0,70-0,80, ini menunjukkan klasifikasi Regresi Logistik data *balanced* termasuk dalam kategori kebaikan model klasifikasi yang cukup (*fair classification*). Nilai rata-rata akurasi total pada

data *training* dan *testing* menghasilkan ketepatan yang bagus yaitu sama-sama 76% serta rata-rata nilai G-*mean* juga memiliki persentase yang bagus yaitu 76%. Nilai rata-rata spesifisitas dan sensitivitas pada data *training* dan *testing* menunjukkan nilai yang seimbang yaitu disekitar 76%. Ini menunjukkan pada data *balanced*, Regresi Logistik dapat mengklasifikasikan dengan baik tanpa ada kecenderungan pengklasifikasian pada kelas mayoritas. Standar deviasi antar *fold* menghasilkan nilai hampir mendekati 0, ini menunjukkan sebaran nilai terhadap rata-rata tidak jauh berbeda atau dapat dikatakan stabil.

Dapat diketahui bahwa model terbaik yang didapat pada data *balanced* adalah model pada *fold* ke-3, dimana nilai G-*mean*, akurasi total yang dihasilkan pada data *testing* paling tinggi dibandingkan *fold* yang lain. Oleh karena itu akan dilakukan pengujian signifikansi parameter pada *fold* ke-3. Pengujian pertama adalah uji serentak dengan *Likelihood Ratio Test*. Dari hasil uji serentak didapatkan nilai G sebesar 206,228 dengan df = 9 serta nilai kritis pada tingkat kesalahan 10% sebesar 13,36. Dapat diambil keputusan untuk menolak H_0 karena nilai G > nilai kritis. Artinya, model pada *fold* ke-3 terdapat minimal 1 variabel yang berpengaruh signifikan terhadap model.

Pada pengujian secara serentak diputuskan tolak H_0 sehingga dapat dilanjutkan pengujian secara parsial untuk mengetahui variabel mana saja yang tidak berpengaruh signifikan terhadap model. Berikut ini adalah hasil pengujian secara parsial:

Tabel 4.13 Hasil Uji Parsial Data *Balanced*

Variabel	Koefisien	S.E	Z*	*p-value*
Konstanta	0,968	0,812	1,193	0,233
X_1	-0,152	0,065	-2,352	0,019
X_2	8,019	1,914	4,190	0,000
X_3	-9,721	1,981	-4,908	0,000
X_4	-1,523	0,821	-1,855	0,064
X_5	-0,965	0,118	-8,206	0,000
X_6	0,259	0,017	15,220	0,000
X_7	-1,044	0,960	-1,088	0,277
X_8	-0,170	0,061	-2,805	0,005

dengan menggunakan tingkat kesalahan (α) sebesar 10% didapat nilai kritis ($Z_{\alpha/2}$) sebesar 1,645. Variabel dengan nilai |Z*| lebih besar daripada nilai kritis atau nilai *p-value* kurang dari α akan diambil keputusan tolak H_0 yang berarti variabel tersebut berpengaruh signifikan terhadap pembentukan model. Variabel yang tidak berpengaruh signifikan terhadap pemodelan data *balanced* adalah rasio jumlah gizi buruk.

Berdasarkan pengujian parsial diketahui terdapat tujuh variabel yang signifikan. Maka akan dilakukan pemilihan variabel pada data *balanced* menggunakan *backward elimination* dengan cara mengeliminasi variabel yang paling tidak signifikan secara bertahap.

Hasil akhir yang didapat dari *backward elimination* adalah sebagai berikut:

Tabel 4.14 Hasil *Backward Elimination* Data *Balanced*

Variabel	Koefisien	S.E	Z*	*p-value*
Konstanta	1,008	0,814	1,238	0,216
X_1	-0,152	0,065	-2,351	0,019
X_2	7,898	1,909	4,137	0,000
X_3	-9,633	1,977	-4,873	0,000
X_4	-1,585	0,822	-1,929	0,054
X_5	-0,963	0,117	-8,200	0,000
X_6	0,259	0,017	15,228	0,000
X_8	-0,173	0,061	-2,862	0,004

Berdasarkan hasil *backward elimination* dan dengan tingkat kesalahan sebesar 10%, didapat variabel-variabel yang berpengaruh signifikansi terhadap model. Variabel tersebut adalah rasio SD/MI, rasio poskesdes, rasio tempat praktik bidan, rasio keluarga pengguna listrik, rasio toko kelontong, jarak tempuh kantor kepala desa ke kecamatan, dan rasio PAD. Untuk selanjutnya variabel tersebut akan digunakan pengklasifikasian menggunakan Regresi Ridge dan Analisis Diskriminan Kernel dengan variabel yang signifikan pada data *balanced*.

B. Regresi Logistik dengan Variabel Signifikan

Sebelum dilakukan klasifikasi dengan Regresi Logistik, dilakukan pendeteksian multikolinieritas menggunakan nilai VIF yang didapat dari Regresi Linier.

Tabel 4.15 Nilai VIF Variabel Bebas Data *Balanced* Variabel Signifikan

X_1	X_2	X_3	X_4	X_5	X_6	X_8
1,01	1,29	1,35	1,14	1,09	1,07	1,05

Dari Tabel 4.15, dapat dilihat bahwa tidak terdapat variabel bebas yang memiliki angka VIF lebih besar dari 5. Oleh karena itu, pada data *imbalanced* menggunakan variabel yang signifikan tidak terdeteksi adanya multikolinieritas. Setelah didapat variabel yang signifikan, selanjutnya dilakukan pengklasifikasian variabel yang signifikan dengan metode Regresi Logistik. Berikut ini adalah hasil rata-rata ketepatan klasifikasi dengan partisi data *training* dan *testing stratified 10-fold CV*.

Tabel 4.16 Rata-rata Ketepatan Klasifikasi Regresi Logistik Data *Balanced* dengan Variabel Signifikan

	Training	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,762	0,003	0,763	0,024
G-*mean*	0,762	0,003	0,763	0,024
Akurasi Total	0,762	0,003	0,763	0,024
Sensitivitas	0,762	0,004	0,761	0,039
Spesifisitas	0,762	0,003	0,765	0,036

Rata-rata nilai AUC sebesar 0,7, yang menunjukkan ketepatan klasifikasi data *balanced* dengan variabel yang signifikan menunjukkan kebaikan klasifikasi yang cukup. Nilai rata-rata G-*mean* pada data *testing* sudah bagus yaitu 76,3%, serta ketepatan akurasi total yang tinggi sebesar 76,3%. Nilai G-mean dan nilai akurasi total bernilai sama besar karena pada data *balanced*, metode Regresi Logistik mampu mengklasifikasikan dengan baik dan tidak ada kecenderungan klasifikasi. Titik sebaran nilai ketepatan klasifikasi tidak jauh dari rata-ratanya ditunjukkan dengan hasil standar deviasi pada masing-masing nilai menghasilkan angka yang kecil dan menunjukkan tahapan *stratified* 10-*fold* CV memberikan partisi data yang baik.

4.3.3 Klasifikasi Regresi Logistik Ridge pada Data *Balanced*

Berikut ini adalah ketepatan klasifikasi dari data *balanced* menggunakan metode Regresi Logistik Ridge dengan tahapan

stratified 10-*fold* CV pada seluruh variabel dan pada variabel yang signifikan.

Tabel 4.17 Rata-rata Ketepatan Klasifikasi Regresi Logistik Ridge Data *Balanced*

Seluruh Variabel	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,765	0,003	0,764	0,021
G-*mean*	0,765	0,003	0,763	0,021
Akurasi Total	0,765	0,003	0,763	0,021
Sensitivitas	0,764	0,004	0,760	0,033
Spesifisitas	0,767	0,003	0,767	0,038
Variabel Signifikan	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,762	0,003	0,760	0,024
G-mean	0,762	0,003	0,759	0,024
Akurasi Total	0,762	0,003	0,760	0,024
Sensitivitas	0,763	0,004	0,759	0,038
Spesifisitas	0,761	0,003	0,760	0,035

Rata-rata nilai G-*mean* dan akurasi total pada data *testing* baik pengklasifikasian dengan seluruh variabel maupun variabel yang signifikan memberikan nilai yang bagus yaitu sekitar 76%. Nilai sensitivitas dan spesifisitas klasifikasi pada data *balanced* telah memberikan hasil yang seimbang, ini menunjukkan data *balanced* memberikan ketepatan yang lebih akurat pada kelas positif. Ketepatan klasifikasi dengan seluruh variabel prediktor dengan pembagian *stratified* 10-*fold CV* memberikan ketepatan klasifikasi yang stabil, dilihat dari nilai standar deviasi yang kecil (kurang dari 1).

4.3.4 Klasifikasi Analisis Diskriminan Kernel pada Data *Balanced*

Berikut ini akan dibahas tentang klasifikasi desa tertinggal pada data *balanced* menggunakan metode Analisis Diskriminan Kernel dengan fungsi kernel *Radial Basis Function* (RBF). Pengklasifikasian dengan Analisis Diskriminan Kernel dirujuk dari hasil pengujian distribusi normal multivariat (Lampiran 6. A) dan uji homogenitas (Lampiran 6. B) yang menunjukkan data tidak memenuhi asumsi dalam Analisis Diskriminan Linier.

Berikut ini adalah hasil ketepatan klasifikasi data training dan testing untuk data *imbalanced* dengan tahapan *stratified* 10-*fold CV*.

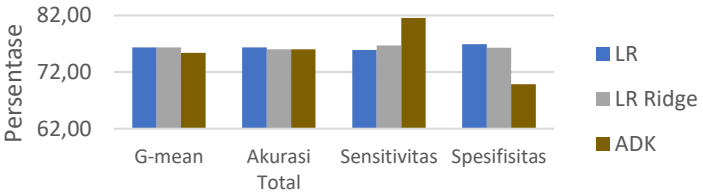
Tabel 4.18 Rata-rata Ketepatan Klasifikasi Analisis Diskriminan Kernel Data *Balanced*

Seluruh Variabel	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,757	0,001	0,757	0,001
G-*mean*	0,754	0,002	0,754	0,013
Akurasi Total	0,760	0,001	0,760	0,010
Sensitivitas	0,814	0,003	0,816	0,027
Spesifisitas	0,700	0,004	0,699	0,039
Variabel Signifikan	***Training***	***(Stdev)***	***Testing***	***(Stdev)***
AUC	0,757	0,001	0,757	0,001
G-mean	0,755	0,002	0,754	0,013
Akurasi Total	0,760	0,001	0,760	0,010
Sensitivitas	0,814	0,003	0,816	0,027
Spesifisitas	0,700	0,004	0,699	0,039

Berdasarkan Tabel 4.18, rata-rata nilai AUC sebesar 0,7, yang menunjukkan ketepatan klasifikasi data *balanced* dengan metode Analisis Diskriminan Kernel menunjukkan kebaikan klasifikasi yang cukup. Rata-rata nilai G-*mean* menghasilkan ketepatan sebesar 75,4%. Akurasi total menunjukkan nilai yang tinggi yaitu 76%. Hasil sensitivitas pada analisis diskriminan kernel pada data *balanced* juga menghasilkan nilai yang bagus yaitu 81,6%. Sedangkan nilai spesifisitas sedikit lebih kecil daripada nilai sensitivitas. Perbedaan ketepatan klasifikasi data *balanced* menggunakan metode Analisis Diskriminan Kernel dengan variabel yang signifikan maupun keseluruhan variabel memberikan hasil yang sama.

4.3.5 Analisis Ketepatan Klasifikasi Data *Balanced*

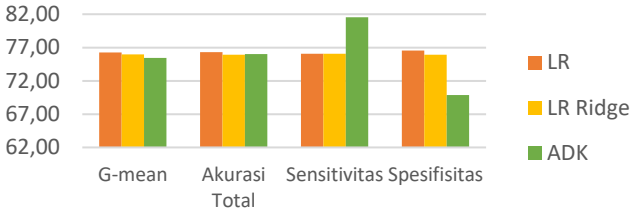
Klasifikasi data *balanced* setelah diterapkan metode resampling SMOTE memberikan ketepatan klasifikasi yang lebih bagus. Berikut ini adalah ketepatan klasifikasi data *balanced* dengan seluruh variabel.



Gambar 4.14 Rata-rata Ketepatan Klasifikasi pada Data *Balanced* dengan Semua Variabel

Nilai G-mean tertinggi pada klasifikasi data *balanced* dengan seluruh variabel dihasilkan pada metode Regresi Logistik. Nilai akurasi total pada seluruh metode telah menunjukkan ketepatan klasifikasi yang bagus dengan nilai sekitar 76%. Nilai sensitivitas tertinggi dihasilkan oleh metode Analisis Diskriminan Kernel. Secara keseluruhan, pada klasifikasi data *balanced* dengan semua variabel, Regresi Logistik memberikan hasil terbaik dengan nilai G-mean dan Akurasi Total yang lebih tinggi daripada metode Regresi Logistik Ridge dan Analisis Diskriminan Kernel.

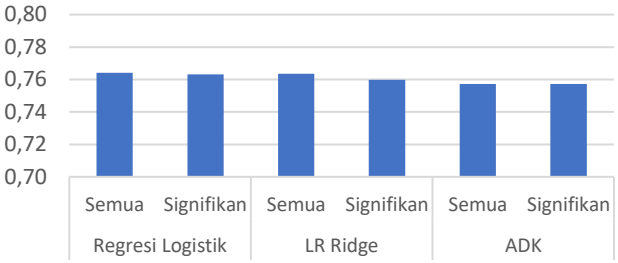
Pada klasifikasi data *balanced* dengan variabel yang signifikan, dihasilkan perbandingan rata-rata ketepatan klasifikasi pada data testing sebagai berikut.



Gambar 4.15 Rata-rata Ketepatan Klasifikasi pada Data *Balanced* dengan Variabel yang Signifikan

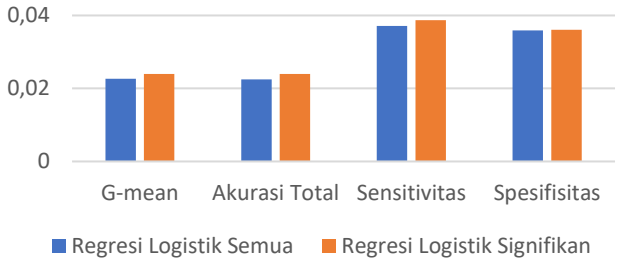
Rata-rata ketepatan klasifikasi data *balanced* dengan variabel yang signifikan memberikan nilai yang hampir sama dengan menggunakan seluruh variabel. Metode yang mengalami perubahan adalah Regresi Logistik Ridge, dimana mengalami sedikit penurunan nilai G-mean dan akurasi total setelah digunakan variabel yang signifikan. Secara keseluruhan, ketepatan klasifikasi terbaik pada data *balanced* dengan variabel yang signifikan dihasilkan oleh metode Regresi Logistik.

Kebaikan model klasifikasi dapat dilihat dari rata-rata nilai AUC yang dihasilkan (Gambar 4.15). Pada data *balanced*, dengan menggunakan metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel menghasilkan nilai AUC antara 0,7 sampai 0,8. Ini menunjukkan kebaikan model klasifikasi pada data *balanced* menggunakan ketiga metode tersebut termasuk kategori klasifikasi yang cukup (*fair classification*).

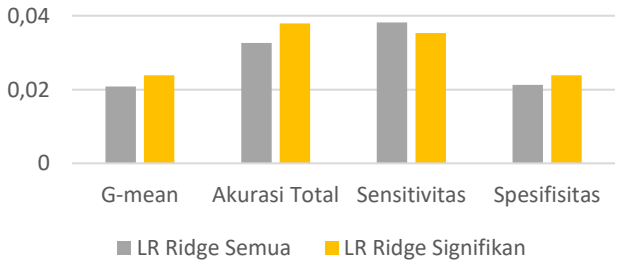


Gambar 4.16 Rata-rata AUC pada Data *Balanced*

Perbandingan lain yang dilakukan dengan melihat nilai standar deviasi yang dihasilkan dari setiap indikator ketepatan klasifikasi. Nilai standar deviasi yang diharapkan pada saat menggunakan variabel yang signikan menghasilkan nilai yang lebih kecil atau sama dengan standar deviasi saat menggunakan seluruh variabel. Berikut ini adalah standar deviasi setiap indikator berdasarkan metode klasifikasi.

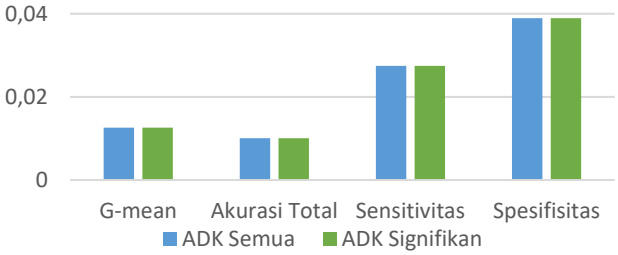


(a)



(b)

Gambar 4.17 Standar Deviasi Ketepatan Klasifikasi Data Balanced (a) Regresi Logistik (b) Regresi Logistik Ridge (c) Analisis Diskriminan Kernel



(c)

Gambar 4.17 Standar Deviasi Ketepatan Klasifikasi Data *Balanced* (a) Regresi Logistik (b) Regresi Logistik Ridge (c) Analisis Diskriminan Kernel (Lanjutan)

Perubahan nilai standar deviasi ketepatan klasifikasi pada metode Regresi Logistik dari semua variabel ke seluruh variabel mengalami peningkatan. Pada metode Regresi Logistik Ridge, nilai G-mean, akurasi total, dan spesifisitas mengalami peningkatan. Tetapi nilai sensitivitas pada metode Regresi Logistik Ridge mengalami sedikit penurunan. Pada metode Analisis Diskriminan Kernel tidak ada perbedaan standar deviasi antara semua variabel dengan variabel yang signifikan.

4.4 Efektivitas SMOTE

Setelah dilakukan klasifikasi menggunakan tiga metode klasifikasi yaitu Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel dengan masing-masing dilakukan pada data *imbalanced* dan data *balanced* didapatkan nilai ketepatan akurasi klasifikasi di Tabel 4.19. Klasifikasi dilakukan dengan membagi data dengan *stratified 10-fold cross validation*.

Tabel 4.19 Hasil Rata-rata Ketepatan Klasifikasi Data *Imbalanced* dan Data *Balanced*

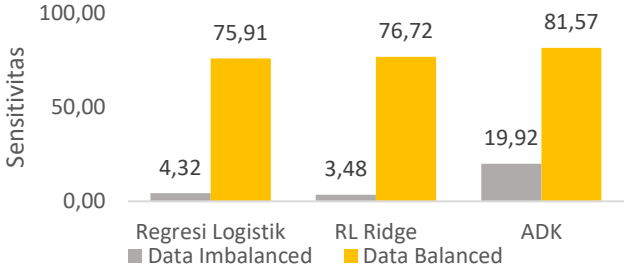
Metode	Variabel	Data *Imbalanced*			Data *Balanced*		
		AUC	G-mean	Sens	AUC	G-mean	Sens
Regresi Logistik	Seluruh	0,52	12,87	4,32	0,76	76,35	75,91
	Signifikan	0,51	9,98	3,48	0,76	76,25	76,09
LR Ridge	Seluruh	0,52	9,98	3,48	0,76	76,34	76,72
	Signifikan	0,52	9,98	3,48	0,76	75,97	76,04
ADK	Seluruh	0,58	42,79	19,92	0,76	75,42	81,57
	Signifikan	0,58	42,78	19,93	0,76	75,42	81,57

Tabel 4.19, menunjukkan terjadi peningkatan ketepatan klasifikasi nilai AUC, G-mean dan sensitivitas dari data

imbalanced ke data *balanced*. Nilai AUC pada data *imbalanced* berada berada di sekitar 0,5, yang menunjukkan ketepatan klasifikasi pada data *imbalanced* termasuk kategori klasifikasi yang salah. Pada data *imbalanced* nilai rata-rata G-mean kecil dikarenakan nilai rata-rata sensitivitas kecil. Pada data *imbalanced* terjadi kecenderungan klasifikasi ke kelas mayoritas atau kelas desa tidak tertinggal. Nilai sensitivitas pada data *imbalanced* dengan metode Regresi Logistik maupun Regresi Logistik Ridge sangat kecil disebabkan oleh klasifikasi tidak dapat mengklasifikasikan kelas minoritas, sehingga model klasifikasi yang didapat dari data *training* tidak mampu mengklasifikasikan kelas minoritas pada data *testing*.

Berbeda saat proporsi kelas seimbang (*balanced*), rata-rata nilai AUC, G-mean dan sensitivitas menjadi lebih besar dari rata-rata ketepatan klasifikasi data *imbalanced*. Ini dikarenakan klasifikasi data *balanced* dapat dengan tepat memprediksi kelas dengan benar. Berikut ini visualisasi perbandingan nilai rata-rata G-mean dan sensitivitas dari data *imbalanced* dan data *balanced*.

Gambar 4.18 Rata-rata (a) AUC, (b) G-mean dan (c) Sensitivitas pada Semua Variabel

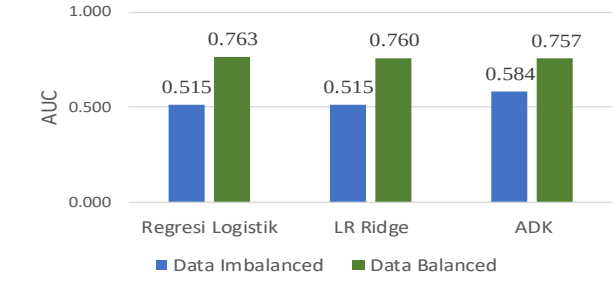


(c)

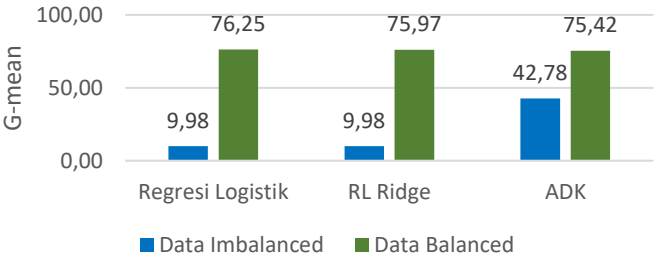
Gambar 4.18 Rata-rata (a) AUC, (b) G-mean dan (c) Sensitivitas pada Semua Variabel (Lanjutan)

Peningkatan nilai rata-rata dari data *imbalanced* ke data *balanced* terjadi pada semua kriteria ketepatan, yaitu AUC, G-mean, dan sensitivitas. Peningkatan nilai AUC paling banyak terjadi pada metode Regresi Logistik Ridge sebesar 1,48 kali, disusul oleh Regresi Logistik (1,47 kali) dan Analisis Diskriminan Kernel (1,29 kali). Peningkatan nilai G-mean terbanyak dari data *imbalanced* ke data *balanced* dengan menggunakan semua variabel (Gambar 4.18) terjadi pada metode Regresi Logistik Ridge dengan peningkatan sebesar 7,6 kali. Sedangkan peningkatan Regresi Logistik dan Analisis Diskriminan Kernel masing-masing sebesar 5,9 kali dan 1,76 kali. Peningkatan pada ketepatan kelas positif atau minoritas terbanyak terjadi pada metode Regresi Logistik Ridge, disusul oleh metode Regresi Logistik dan Analisis Diskriminan Kernel.

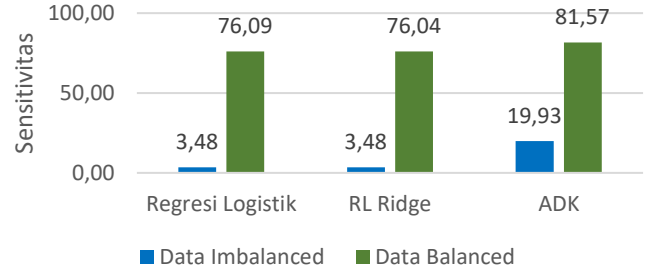
Sedangkan saat menggunakan variabel yang signifikan (Gambar 4.19) antara metode Regresi Logistik dan Regresi Logistik Ridge peningkatan G-mean dan sensitivitas hampir sama yaitu sekitar 7,6 kali dan 21,8 kali. Peningkatan G-mean dan sensitivitas pada metode Analisis Diskriminan Kernel sebanyak 1,8 kali dan 4,1 kali. Peningkatan terbanyak nilai AUC juga berturut-turut adalah Regresi Logistik (1,48 kali), Regresi Logistik Ridge (1,47 kali), dan Analisis Diskriminan Kernel (1,29 kali).



(a)



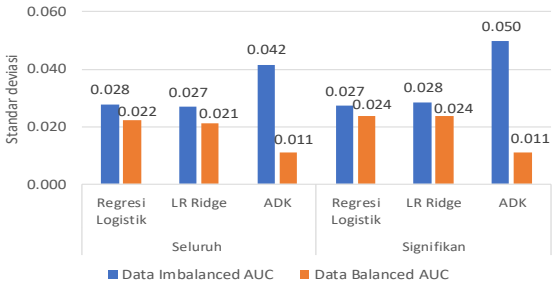
(b)



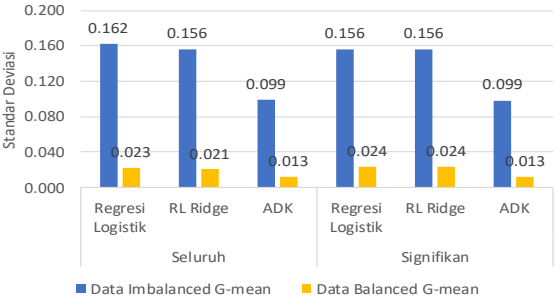
(c)

Gambar 4.19 Rata-rata (a) AUC, (b) G-mean dan (c) Sensitivitas pada Variabel Signifikan

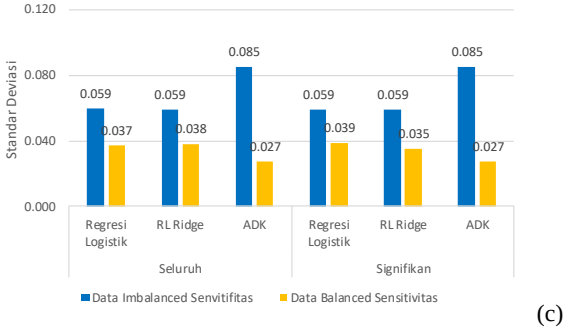
Setelah dilihat nilai rata-rata, selanjutnya akan diliat dari nilai standar deviasi yang dihasilkan dari nilai AUC, G-mean dan sensitivitas (Gambar 4.20). Apabila menunjukkan penurunan nilai standar deviasi, ini menunjukkan nilai ketepatan yang dihasilkan menjadi lebih stabil.



(a)



(b)



(c)

Gambar 4.20 Hasil Standar Deviasi (a) AUC, (b) G-mean dan (c) Sensitivitas pada Data *Imbalanced* dan Data *Balanced*

Penurunan standar deviasi nilai AUC dari data *imbalanced* ke data *balanced* paling besar terjadi pada Analisis Diskriminan Kernel, dengan penurunan menggunakan seluruh variabel sebesar 3,8 kali dan dengan variabel yang signifikan penurunannya sebesar 4,57 kali. Penurunan nilai standar deviasi G-mean dari data *imbalanced* ke data *balanced* pada metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis

Diskriminan Kernel dengan seluruh variabel berturut turut adalah 7,1 kali, 7,5 kali, dan 7,8 kali. Sedangkan bila menggunakan variabel yang signifikan terjadi penurunan G-mean sebesar 6,5 kali, 6,5 kali, dan 7,8 kali. Penurunan nilai sensitivitas dari data *imbalanced* ke data *balanced* pada metode Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel dengan seluruh variabel berturut turut adalah 1,6 kali, 1,5 kali, dan 3,1 kali. Sedangkan bila menggunakan variabel yang signifikan terjadi penurunan sensitivitas sebesar 1,5 kali, 1,7 kali, dan 3,1 kali.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Setelah dilakukan analisis dan pembahasan mengenai klasifikasi status desa tertinggal di Jawa Timur menggunakan metode regresi logistik dan analisis diskriminan kernel serta penerapan metode replikasi data minoritas (SMOTE), dapat diambil kesimpulan sebagai berikut.

1. Data observasi desa tertinggal di 5 Kabupaten di Jawa Timur merupakan data *imbalanced* dengan proporsi kelas desa tertinggal sebesar 10% dan kelas desa tidak tertinggal 90%. Perbedaan nilai median kelompok desa tertinggal lebih rendah daripada median kelompok desa tidak tertinggal terdapat pada rasio Poskesdes, tempat praktik bidan, toko kelontong, dan PAD. Sedangkan yang nilainya median antar kelompok sama terdapat pada rasio Sekolah Dasar/MI, keluarga pengguna listrik, dan warga penderita gizi buruk. Untuk variabel jarak, median desa tertinggal lebih tinggi daripada median kelompok desa tidak tertinggal.
2. Perbandingan data *imbalanced* kelas desa tertinggal:desa tidak tertinggal sebesar 10%:90%. Pada data *imbalanced* baik dengan seluruh variabel dan variabel signifikan tidak terdeteksi adanya multikolinieritas, tidak berdistribusi normal multivariat, dan tidak homogen. Ketepatan klasifikasi menggunakan Regresi Logistik, Regresi Logistik Ridge, dan Analisis Diskriminan Kernel menghasilkan ketepatan akurasi total yang tinggi (semu), tetapi nilai G-mean dan sensitivitas (akurasi kelas minoritas) yang rendah yaitu 10% dan 3,5%. Ketepatan akurasi pada data *imbalanced*, akan cenderung mengklasifikasikan pada kelas mayoritas (spesifisitasnya tinggi).
3. Proporsi kelas hasil resampling dengan SMOTE menjadi seimbang yaitu 53%:47%. Data *balanced* baik dengan seluruh variabel dan variabel signifikan tidak terdeteksi adanya multikolinieritas, tidak berdistribusi normal

multivariat, dan tidak homogen. Ketepatan klasifikasi menggunakan metode Regresi Logistik, Regresi Logistik Ridge dan Analisis Diskriminan Kernel menghasilkan ketepatan klasifikasi yang cukup bagus. Ketepatan rata-rata nilai AUC, G-mean, dan sensitivitas pada seluruh variabel dan variabel yang signifikan menghasilkan nilai yang tidak jauh berbeda, begitu pula antar metodenya. Metode Analisis Regresi Logistik menghasilkan nilai AUC (0,764) dan G-mean (76,35%) yang paling tinggi.

4. Efektifitas kombinasi SMOTE dengan tiga metode menghasilkan peningkatan nilai AUC, G-mean dan sensitivitas. Serta nilai standar deviasi yang dihasilkan dari penerapan SMOTE menjadi lebih kecil dibandingkan data yang *imbalanced*, yang menunjukkan kestabilan ketepatan klasifikasi. Metode Regresi Logistik dan Regresi Logistik Ridge menghasilkan peningkatan nilai AUC terbesar baik pada seluruh variabel maupun variabel signifikan. Peningkatan G-mean dan sensitivitas terbesar terjadi pada kombinasi SMOTE dengan Regresi Logistik Ridge menggunakan seluruh variabel (7,6 kali dan 22 kali).

5.2 Saran

Saran yang dapat diberikan peneliti untuk penelitian selanjutnya adalah sebagai berikut:

1. Menambahkan variabel dan jenis data yang lain, seperti jenis data kategorik.
2. Menggunakan variabel respon yang *multiclass*.
3. Menggunakan metode resampling secara umum yang lain, seperti yang terdapat pada teknik *oversampling*, *undersampling*, maupun gabungan.
4. Menggunakan metode klasifikasi *imbalanced* yang lain.

Saran untuk pemerintah yaitu sebaiknya menambah fasilitas sarana kesehatan pada setiap desa serta pemberdayaan sumber daya manusia agar dapat meningkatkan Pendapatan Asli Desa.

DAFTAR PUSTAKA

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (Second ed.). United State of America: A John Wiley & Sons, Inc.

Antipov, E., & Pokryshevskaya, E. (2010). Applying CHAID for Logistic Regression Diagnostics and Classification Accuracy Improvement. *Journal of Targeting, Measurement and Analysis for Marketing, 18*, 109-117.

Aswa, R., Saleh AF, M., & Talangko, L. (2015). *Analisis Diskriminan Kernel dengan Metode Cross Validation (Studi Kasus : Faktor-Faktor yang Berhubugan dengan Kejadian Hipertensi pada Puskesmas Usuku Wakatobi Sulawesi Tenggara Tahun 2013).* Makassar: Program Sarjana, Universitas Hasanuddin.

Badan Pusat Statistik Provinsi Jawa Timur. (2017). *Provinsi Jawa Timur Dalam Angka 2017.* Surabaya: Badan Pusat Statistik Provinsi Jawa Timur.

BAPPEDA PROVINSI JATIM. (2017). *Potret Pembangunan Jawa Timur 2008-2017.* Surabaya: BAPPEDA Provinsi Jawa Timur.

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

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

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

Castellanos, F. J., Valero-Mas, J. J., Calvo-Zaragoza, J., & Rico, J. (2018). Oversampling Imbalanced Data in the String Space. *Elsevier*, 1-10.

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

Choi, M. J. (2010). *A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines.* Iowa: Graduate Theses. Iowa State University.

Direktorat Jenderal Pembangunan Daerah Tertinggal. (2016). *Profil & Potensi Daerah Tertinggal*. Dipetik Juni 8, 2018, dari Provinsi Jawa Timur: http://ditjenpdt.kemendesa.go.id/potensi/province/10-provinsi-jawa-timur#

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

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

Gorunescu, F. (2011). *Data Mining Concepts and Techniques.* Berlin: Springer .

Gujarati, D. N. (2004). *Basic Econometrics.* New York: Tata McGraw Hill.

Hair, J. F., Black, W., Babin, J., & Anderson, R. (2006). *Multivariate Data Analysis* (7th ed.). Pearson Education Prentice Hall, Inc.

Hairani, H., Setiawan, N. A., & Adji, T. B. (2016). Metode Klasifikasi Data Mining dan Teknik Sampling SMOTE Menangani Class Imbalance Untuk Segmentasi Customer Pada Industri Perbankan. *SNST Fakultas Teknik* (hal. 168-172). Semarang: Universitas Wahid Hasyim.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts & Techniques* (3rd ed.). United States of America: Elsevier Inc.

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

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., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). New Jersey: John Wiley & Sons, Inc.

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

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

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 137-163.

Li, Y., Gong, S., & Liddell, H. (2001, 10 14). *Kernel Discriminant Analysis.* Diambil kembali dari http://homepages.inf.ed.ac.uk: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/LI1/kda/

Liu, S., Kawamoto, T., Morita, O., Yoshinari, K., & Honda, H. (2017). Discriminating Between Adaptive And Carcigonic liver Hyperthropy In Rat Studies Using Logistic Ridge Regression Analysis of Toxicogenomic Data. The Mode of Action And Predictive Models.

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

Maalouf, M., Homouz, D., & Trafalis, T. B. (2018). Logistic Regression in Large Rare Events and Imbalanced Data: A Performance Comparison of Prior Correction and Weighting Methods. *Computational Intelligence, 34*, 161-174.

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

Novritasari, A. A., & Purnami, S. W. (2015). *Klasifikasi Kerentanan Seseorang Terserang Stroke di Jawa Timur Menggunakan Synthetic Minoryty Oversampling Technique (SMOTE) dan Support Vector Machine (SVM).* Surabaya: Tugas Akhir. ITS.

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.

Republik Indonesia. (2014). *Undang-Undang Nomor 6 Tahun 2014.* Jakarta: Kementrian Hukum dan Hak Asasi Manusia.

Ryan, T. P. (2009). *Modern Regression Methods* (2nd ed.). New York: Wiley.

Sain, H., & Purnami, S. W. (2013). Combine Sampling Support Vector Machine Untuk Klasifikasi Data Imbalance. *Procedia Computer Science*, 59-66.

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

Sharma, S. (1996). *Applied Multivariate Technique.* United States: John Wiley & Sons, Inc.

Sulasih, D. E., Purnami, S. W., & Rahayu, S. P. (2016). *Rare Event Weighted Logistic Regression Untuk Klasifikasi Imbalanced*

Data (Studi Kasus: Klasifikasi Desa Tertinggal di Provinsi Jawa Timur). Surabaya: Thesis. ITS.

Sungkono, J., & Nugrahaningsih, T. (2017). Simulasi Dampak Multikolinearitas pada Kondisi Penyimpangan Asumsi Normalitas. *Magistra*, 45-50.

Sunyoto, Setiawan, & Zain, I. (2009). *Regresi Logistik Ridge: Pada Keberhasilan Siswa Negeri 1 Kediri Diterima di Perguruan Tinggi Negeri.* Surabaya: Thesis, Statistika, Institut Teknologi Sepuluh Nopember.

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

You, D., Hamsici, O. C., & Matinez, A. M. (2010). Kernel Optimization in Discriminant Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 631 - 638.

Yuan, G.-X., Ho, C.-H., & Lin, C.-J. (2012). Recent Advances of Large-Scale Linear Classification. *Proceedings of The IEEE, 100*, 2584-2603.

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

LAMPIRAN

Lampiran 1 Persentase Desa Tertinggal di Jawa Timur

Kabupaten/Kota	Desa Tertinggal	Persentase Desa Tertinggal
Kabupaten Pacitan	0	0
Kabupaten Ponorogo	4	1,423
Kabupaten Trenggalek	6	3,947
Kabupaten Tulungagung	4	1,556
Kabupaten Blitar	1	0,455
Kabupaten Kediri	2	0,583
Kabupaten Malang	7	1,852
Kabupaten Lumajang	10	5,051
Kabupaten Jember	1	0,442
Kabupaten Banyuwangi	0	0
Kabupaten Bondowoso	17	8,134
Kabupaten Situbondo	12	9,091
Kabupaten Probolinggo	9	2,769
Kabupaten Pasuruan	12	3,519
Kabupaten Sidoarjo	1	0,312
Kabupaten Mojokerto	2	0,669
Kabupaten Jombang	1	0,331
Kabupaten Nganjuk	7	2,652
Kabupaten Madiun	2	1,010
Kabupaten Magetan	0	0
Kabupaten Ngawi	2	0,939
Kabupaten Bojonegoro	4	0,955
Kabupaten Tuban	8	2,572
Kabupaten Lamongan	2	0,433
Kabupaten Gresik	0	0,000
Kabupaten Bangkalan	44	16,117
Kabupaten Sampang	14	7,778
Kabupaten Pamekasan	7	3,933
Kabupaten Sumenep	28	8,537
Kota Batu	0	0

Lampiran 2 Data *Imbalanced* Desa Tertinggal

Desa	**X_1**	**X_2**	**X_3**	⋯	**X_8**	**Status**
1	0,83	0	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	0,032	⋯	0,6	0
8	1,087	0,016	0,016	⋯	1,749	0
9	0,606	0	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
13	1,093	0	0,052	⋯	1,098	0
14	0,948	0,064	0,064	⋯	10,737	0
15	1,389	0,129	0	⋯	3,737	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	⋯	0,432	1
1115	1,015	0	0	⋯	0,323	1
1116	0,734	0,031	0,031	⋯	0,092	0
1117	0,907	0,034	0	⋯	0,101	0
1118	0,803	0,081	0	⋯	0,244	1
1119	0,953	0,019	0,029	⋯	0,242	0
1120	1,115	0	0,012	⋯	0,294	0
1121	1,367	0,025	0	⋯	0,627	0
1122	1,333	0	0	⋯	1,094	1

Lampiran 3 Data *Balanced* Desa Tertinggal

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
13	0,322	0,000	0,035	⋯	0,349	0
14	1,136	0,000	0,000	⋯	0,649	1
15	1,583	0,000	0,046	⋯	0,388	1
⋮	⋮	⋮	⋮	⋱	⋮	⋮
2169	0,619	0	0,08	⋯	0,197	0
2170	1,056	0,035	0,035	⋯	0,615	0
2171	1,193898	0	0	⋯	1,966	0
2172	1,918	0	0	⋯	0,899032	1
2173	1,105	0,035	0,035	⋯	1,291	0
2174	0,589	0,000	0,014	⋯	0,043	0
2175	0,870	0,048	0,096	⋯	1,78	0
2176	0,651	0,000	0,000	⋯	0,133	1
2177	0,773	0,007	0,059	⋯	0,118	0
2178	1,501	0,000	0,061	⋯	0,43269	1
2179	0,873	0,000	0,034	⋯	0,401	0
2180	0,974	0,009	0,027	⋯	0,027	0
2181	1,320	0,022	0,044	⋯	0,111	0
2182	1,333	0,000	0,095	⋯	0,473	0
2183	0,817	0,000	0,025	⋯	0,256416	1
2184	1,511	0,060	0,003	⋯	2,261037	1
2185	1,740	0,000	0,011	⋯	0,407402	1

Lampiran 4 Syntax Penelitian di *software R*

A. Syntax SMOTE

```
library(unbalanced)
data<-read.csv("F:/data5.csv", header=TRUE, sep=";")
head(data)
print(table(data$Status))
Y<-as.factor(data$Status)
X=data[,-9]
databaru=ubSMOTE(X,Y,perc.over=900,k=5, perc.under = 100,
     verbose=TRUE)
newdata=cbind(databaru$X,databaru$Y)
print(table(databaru$Y))

write.csv(newdata,file="F:/ Data SMOTE.csv")
```

B. Syntax Klasifikasi Regresi Logistik, Regresi Logistik Ridge dan Analisis DIskriminan Kernel

```
library(data.table)
library(caret)
library(MASS)
library(MXM)
library(glmnet)
library(e1071)
library(pROC)
library(ROCR)

#membaca data
data<-read.csv("F:/rasio.csv", header=TRUE, sep=";")
data1<-read.csv("F:/SMOTE.csv", header=TRUE, sep=";")
Y<-as.factor(data$Status)

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

Lampiran 4 Syntax Penelitian di R (Lanjutan)

```
GmeanTrain=rep(0,r)
GmeanTest=rep(0,r)

#Regresi Logistik
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]=as.numeric(roc(train$Status, predtrain)$auc)
  AUCTest[i]=as.numeric(roc(test$Status, predtest)$auc)

  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)
summary(model)
#export csv
hasiltotal=data.frame(TotalAccuracyTrain,
                      TotalAccuracyTest,
```

Lampiran 4 Syntax Penelitian di R (Lanjutan)

```
                SensTrain,
                SensTest,
                SpesTrain,
                SpesTest,
                GmeanTrain,
                GmeanTest)
hasilmean=data.frame(mean(TotalAccuracyTrain),
                mean(TotalAccuracyTest),
                mean(SensTrain),
                mean(SensTest),
                mean(SpesTrain),
                mean(SpesTest),
                mean(GmeanTrain),
                mean(GmeanTest))
hasilmean
write.csv(hasiltotal,file= "F:/Reglog_total.csv")
write.csv(hasilmean,file="F:/ Reglog_total_mean.csv")
```

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

#COBA
data<-read.csv("F:/data5.csv", header=TRUE, sep=";")
Y<-as.factor(data$Status)

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

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

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

Lampiran 4 Syntax Penelitian di R (Lanjutan)

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

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

  AUCTrain[i]=as.numeric(roc(train$Status, predtrain)$auc)
  AUCTest[i]=as.numeric(roc(test$Status, predtest)$auc)

  GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
  GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
```

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

#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.vektor(predtrain2)
predtrain=as.numeric(predtrain3)

predtest1=kfda.predict(model, test)
predtest2=predtest1$class
predtest3=as.vektor(predtest2)
predtest=as.numeric(predtest3)

  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]))
```

Lampiran 4 Syntax Penelitian di R (Lanjutan)

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

  AUCTrain[i]=as.numeric(roc(train$Status, predtrain)$auc)
  AUCTest[i]=as.numeric(roc(test$Status, predtest)$auc)

  GmeanTrain[i]=sqrt((SensTrain[i])*(SpesTrain[i]))
  GmeanTest[i]=sqrt((SensTest[i])*(SpesTest[i]))
}
```

C. Syntax Pengujian Analisis Diskriminan

```
data=read.csv("F:/SMOTE.csv",header=T,sep=";")

#Uji Mardia's Test
Library(MVN)
Hasil=mvn(data=data[-9],mvnTest = c("mardia"),covariance = TRUE,
multivariatePlot = "qq")
Hasil

#Uji Homogenitas
library(biotools)
Uji_Homogen<-boxM(data.matrix(data)[,-9], data[,9])
Uji_Homogen
```

Lampiran 5 Output Nilai Ketepatan Klasifikasi

A. Output Klasifikasi Data *Imbalanced* Metode Regresi Logistik

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,898	0,039	0,996	0,197	0,903	0,083	1,000	0,289
2	0,898	0,049	0,994	0,220	0,903	0,083	1,000	0,289
3	0,900	0,067	0,996	0,259	0,893	0,000	0,990	0,000
4	0,900	0,058	0,997	0,240	0,902	0,182	0,980	0,422
5	0,898	0,058	0,994	0,240	0,902	0,000	1,000	0,000
6	0,901	0,087	0,994	0,293	0,893	0,000	0,990	0,000
7	0,898	0,058	0,994	0,240	0,902	0,000	1,000	0,000
8	0,899	0,058	0,994	0,241	0,893	0,000	1,000	0,000
9	0,899	0,049	0,996	0,220	0,893	0,000	1,000	0,000
10	0,897	0,068	0,991	0,260	0,893	0,083	0,990	0,287
mean	0,899	0,059	0,995	0,241	0,897	0,043	0,995	0,129
stdev	0,001	0,012	0,001	0,025	0,005	0,059	0,007	0,162

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,898	0,039	0,996	0,197	0,902	0,000	1,000	0,000
2	0,896	0,019	0,996	0,139	0,903	0,083	1,000	0,289
3	0,900	0,067	0,996	0,259	0,893	0,000	0,990	0,000
4	0,901	0,067	0,997	0,259	0,902	0,182	0,980	0,422
5	0,898	0,048	0,996	0,219	0,902	0,000	1,000	0,000
6	0,902	0,087	0,996	0,294	0,893	0,000	0,990	0,000
7	0,896	0,029	0,996	0,169	0,902	0,000	1,000	0,000
8	0,898	0,039	0,996	0,197	0,893	0,000	1,000	0,000
9	0,898	0,029	0,997	0,170	0,893	0,000	1,000	0,000
10	0,896	0,068	0,990	0,259	0,893	0,083	0,990	0,287
mean	0,898	0,049	0,995	0,216	0,897	0,035	0,995	0,100
stdev	0,002	0,021	0,002	0,047	0,005	0,059	0,007	0,156

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

B. Output Klasifikasi Data *Imbalanced* Metode Regresi Logistik Ridge

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,896	0,019	0,996	0,139	0,894	0,000	1,000	0,000
2	0,896	0,000	0,998	0,000	0,903	0,083	1,000	0,289
3	0,896	0,010	0,998	0,098	0,902	0,000	1,000	0,000
4	0,897	0,019	0,998	0,139	0,902	0,182	0,980	0,422
5	0,895	0,019	0,996	0,138	0,902	0,000	1,000	0,000
6	0,897	0,010	0,999	0,098	0,893	0,000	0,990	0,000
7	0,896	0,010	0,998	0,098	0,902	0,000	1,000	0,000
8	0,896	0,019	0,996	0,139	0,893	0,000	1,000	0,000
9	0,896	0,000	0,998	0,000	0,893	0,000	1,000	0,000
10	0,898	0,039	0,996	0,197	0,893	0,083	0,990	0,287
mean	0,896	0,014	0,997	0,105	0,898	0,035	0,996	0,100
stdev	0,001	0,011	0,001	0,059	0,004	0,059	0,007	0,156

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,896	0,019	0,996	0,000	0,902	0,000	1,000	0,000
2	0,895	0,010	0,996	0,098	0,903	0,083	1,000	0,289
3	0,897	0,038	0,996	0,196	0,903	0,000	1,000	0,000
4	0,896	0,019	0,997	0,138	0,902	0,182	0,980	0,422
5	0,895	0,019	0,996	0,138	0,902	0,000	1,000	0,000
6	0,899	0,058	0,996	0,240	0,893	0,000	0,990	0,000
7	0,895	0,019	0,996	0,138	0,902	0,000	1,000	0,000
8	0,896	0,019	0,996	0,139	0,893	0,000	1,000	0,000
9	0,897	0,010	0,998	0,098	0,893	0,000	1,000	0,000
10	0,899	0,049	0,996	0,220	0,893	0,083	0,990	0,287
mean	0,897	0,026	0,996	0,141	0,898	0,035	0,996	0,100
stdev	0,001	0,016	0,001	0,065	0,005	0,059	0,007	0,156

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

C. Output Klasifikasi Data *Imbalanced* Metode Analisis Diskriminan Kernel

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,895	0,214	0,972	0,456	0,885	0,250	0,960	0,490
2	0,897	0,223	0,974	0,466	0,885	0,250	0,960	0,490
3	0,897	0,279	0,968	0,520	0,866	0,273	0,931	0,504
4	0,897	0,231	0,974	0,474	0,902	0,273	0,970	0,514
5	0,897	0,260	0,970	0,502	0,893	0,091	0,980	0,299
6	0,893	0,250	0,967	0,492	0,884	0,182	0,960	0,418
7	0,891	0,240	0,966	0,482	0,875	0,091	0,960	0,295
8	0,892	0,214	0,969	0,455	0,920	0,333	0,990	0,574
9	0,897	0,233	0,972	0,476	0,911	0,167	1,000	0,408
10	0,896	0,223	0,972	0,466	0,884	0,083	0,980	0,286
mean	0,895	0,237	0,970	0,479	0,890	0,199	0,969	0,428
stdev	0,002	0,020	0,003	0,020	0,015	0,085	0,018	0,099

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-means	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,895	0,214	0,972	0,456	0,885	0,250	0,960	0,490
2	0,897	0,223	0,974	0,466	0,885	0,250	0,960	0,490
3	0,897	0,279	0,968	0,520	0,866	0,273	0,931	0,504
4	0,897	0,231	0,974	0,474	0,902	0,273	0,970	0,514
5	0,897	0,260	0,970	0,502	0,893	0,091	0,980	0,299
6	0,893	0,250	0,967	0,492	0,884	0,182	0,960	0,418
7	0,891	0,240	0,966	0,482	0,875	0,091	0,960	0,295
8	0,892	0,214	0,969	0,455	0,920	0,333	0,990	0,574
9	0,897	0,233	0,972	0,476	0,911	0,167	1,000	0,408
10	0,896	0,223	0,972	0,466	0,884	0,083	0,980	0,286
mean	0,895	0,237	0,970	0,479	0,891	0,199	0,969	0,428
stdev	0,002	0,020	0,003	0,020	0,016	0,085	0,018	0,099

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

D. Output Klasifikasi Data *Balanced* Metode Regresi Logistik

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-mean	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,760	0,755	0,766	0,760	0,776	0,739	0,817	0,777
2	0,760	0,756	0,765	0,760	0,781	0,783	0,779	0,781
3	0,762	0,761	0,764	0,763	0,795	0,783	0,808	0,795
4	0,769	0,768	0,769	0,769	0,767	0,817	0,712	0,763
5	0,769	0,765	0,773	0,769	0,740	0,774	0,702	0,737
6	0,763	0,757	0,768	0,763	0,761	0,757	0,767	0,762
7	0,767	0,764	0,770	0,767	0,771	0,757	0,786	0,771
8	0,767	0,764	0,770	0,767	0,716	0,670	0,767	0,717
9	0,772	0,769	0,775	0,772	0,748	0,739	0,757	0,748
10	0,761	0,759	0,763	0,761	0,784	0,774	0,796	0,785
mean	0,765	0,762	0,768	0,765	0,764	0,759	0,769	0,764
stdev	0,004	0,005	0,004	0,004	0,022	0,037	0,036	0,023

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-mean	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,760	0,757	0,764	0,760	0,776	0,739	0,817	0,777
2	0,758	0,756	0,760	0,758	0,781	0,800	0,760	0,780
3	0,759	0,761	0,757	0,759	0,790	0,783	0,798	0,790
4	0,763	0,764	0,762	0,763	0,767	0,817	0,712	0,763
5	0,763	0,765	0,760	0,763	0,740	0,774	0,702	0,737
6	0,760	0,757	0,762	0,760	0,766	0,765	0,767	0,766
7	0,762	0,760	0,764	0,762	0,771	0,748	0,796	0,772
8	0,768	0,766	0,769	0,768	0,706	0,670	0,748	0,707
9	0,767	0,768	0,765	0,767	0,748	0,739	0,757	0,748
10	0,761	0,761	0,761	0,761	0,784	0,774	0,796	0,785
mean	0,762	0,762	0,762	0,762	0,763	0,761	0,765	0,763
stdev	0,003	0,004	0,003	0,003	0,024	0,039	0,036	0,024

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

E. Output Klasifikasi Data *Balanced* Metode Regresi Logistik Ridge

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitifitas	Spesifisitas	G-mean	Akurasi Total	Sensitifitas	Spesifisitas	G-mean
1	0,760	0,757	0,765	0,761	0,776	0,739	0,817	0,777
2	0,762	0,758	0,767	0,763	0,781	0,783	0,779	0,781
3	0,760	0,760	0,760	0,760	0,795	0,783	0,808	0,795
4	0,768	0,767	0,769	0,768	0,767	0,817	0,712	0,763
5	0,768	0,767	0,769	0,768	0,735	0,774	0,692	0,732
6	0,766	0,764	0,768	0,766	0,757	0,748	0,767	0,757
7	0,766	0,766	0,766	0,766	0,771	0,757	0,786	0,771
8	0,770	0,769	0,770	0,770	0,725	0,687	0,767	0,726
9	0,768	0,768	0,768	0,768	0,748	0,748	0,748	0,748
10	0,763	0,762	0,763	0,763	0,780	0,765	0,796	0,781
mean	0,765	0,764	0,767	0,765	0,763	0,760	0,767	0,763
stdev	0,003	0,004	0,003	0,003	0,021	0,033	0,038	0,021

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitifitas	Spesifisitas	G-mean	Akurasi Total	Sensitifitas	Spesifisitas	G-mean
1	0,758	0,758	0,758	0,758	0,772	0,739	0,808	0,773
2	0,759	0,757	0,763	0,760	0,781	0,800	0,760	0,780
3	0,757	0,760	0,754	0,757	0,790	0,783	0,798	0,790
4	0,766	0,768	0,763	0,765	0,767	0,817	0,712	0,763
5	0,764	0,765	0,764	0,764	0,744	0,783	0,702	0,741
6	0,762	0,761	0,762	0,762	0,761	0,757	0,767	0,762
7	0,761	0,761	0,760	0,761	0,761	0,739	0,786	0,762
8	0,767	0,767	0,766	0,767	0,706	0,678	0,738	0,707
9	0,766	0,768	0,764	0,766	0,734	0,730	0,738	0,734
10	0,762	0,764	0,760	0,762	0,780	0,765	0,796	0,781
mean	0,762	0,763	0,761	0,762	0,760	0,759	0,760	0,759
stdev	0,003	0,004	0,003	0,003	0,024	0,038	0,035	0,024

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

F. Output Klasifikasi Data *Balanced* Metode Analisis Diskriminan Kernel

Semua Variabel

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-mean	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,757	0,816	0,691	0,751	0,776	0,800	0,750	0,775
2	0,760	0,807	0,708	0,756	0,758	0,826	0,683	0,751
3	0,759	0,810	0,702	0,754	0,758	0,791	0,721	0,755
4	0,761	0,817	0,699	0,756	0,758	0,887	0,615	0,739
5	0,759	0,813	0,700	0,754	0,753	0,817	0,683	0,747
6	0,760	0,814	0,698	0,754	0,766	0,826	0,699	0,760
7	0,758	0,814	0,695	0,753	0,771	0,817	0,718	0,766
8	0,761	0,815	0,701	0,756	0,771	0,783	0,757	0,770
9	0,760	0,816	0,697	0,755	0,748	0,809	0,680	0,741
10	0,761	0,813	0,703	0,756	0,743	0,800	0,680	0,737
mean	0,760	0,814	0,700	0,754	0,760	0,816	0,699	0,754
stdev	0,001	0,003	0,004	0,002	0,010	0,027	0,039	0,013

Variabel Signifikan

	Training				Testing			
Fold	Akurasi Total	Sensitivitas	Spesifisitas	G-mean	Akurasi Total	Sensitivitas	Spesifisitas	G-means
1	0,757	0,816	0,691	0,751	0,776	0,800	0,750	0,775
2	0,760	0,808	0,708	0,756	0,758	0,826	0,683	0,751
3	0,759	0,811	0,702	0,755	0,758	0,791	0,721	0,755
4	0,760	0,815	0,699	0,755	0,758	0,887	0,615	0,739
5	0,759	0,813	0,700	0,754	0,753	0,817	0,683	0,747
6	0,760	0,814	0,698	0,754	0,766	0,826	0,699	0,760
7	0,758	0,814	0,695	0,753	0,771	0,817	0,718	0,766
8	0,762	0,816	0,701	0,756	0,771	0,783	0,757	0,770
9	0,762	0,817	0,700	0,756	0,748	0,809	0,680	0,741
10	0,761	0,813	0,703	0,756	0,743	0,800	0,680	0,737
mean	0,760	0,814	0,700	0,755	0,760	0,816	0,699	0,754
stdev	0,001	0,003	0,004	0,002	0,010	0,027	0,039	0,013

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

G. Output AUC Metode Regresi Logistik

Data *Imbalanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,517	0,542	0,517	0,500
2	0,522	0,542	0,508	0,542
3	0,531	0,495	0,531	0,495
4	0,527	**0,581**	0,532	0,581
5	0,526	0,500	0,522	0,500
6	0,541	0,495	0,541	0,495
7	0,526	0,500	0,512	0,500
8	0,526	0,500	0,517	0,500
9	0,522	0,500	0,513	0,500
10	0,530	0,537	0,529	0,537
Mean	0,527	0,519	0,522	0,515
Stdev	0,006	0,028	0,010	0,027

Data *Balanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,760	0,778	0,760	0,778
2	0,760	0,781	0,758	0,780
3	0,763	**0,795**	0,759	0,790
4	0,769	0,764	0,763	0,764
5	0,769	0,738	0,763	0,738
6	0,763	0,762	0,760	0,766
7	0,767	0,771	0,762	0,772
8	0,767	0,718	0,768	0,709
9	0,772	0,748	0,767	0,748
10	0,761	0,785	0,761	0,785
Mean	0,765	0,764	0,762	0,763
Stdev	0,004	0,022	0,003	0,024

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

H. Output AUC Metode Regresi Logistik Ridge

Data *Imbalanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,508	0,500	0,508	0,500
2	0,499	0,542	0,503	0,542
3	0,504	0,500	0,522	0,500
4	0,509	0,581	0,508	0,581
5	0,507	0,500	0,507	0,500
6	0,504	0,495	0,531	0,495
7	0,504	0,500	0,512	0,500
8	0,508	0,500	0,508	0,500
9	0,499	0,500	0,509	0,500
10	0,517	0,537	0,517	0,537
Mean	0,506	0,515	0,512	0,515
Stdev	0,005	0,027	0,009	0,028

Data *Balanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,761	0,778	0,758	0,773
2	0,763	0,781	0,760	0,780
3	0,760	0,795	0,757	0,790
4	0,768	0,764	0,765	0,764
5	0,768	0,733	0,764	0,742
6	0,766	0,757	0,762	0,762
7	0,766	0,771	0,761	0,763
8	0,770	0,727	0,767	0,708
9	0,768	0,748	0,766	0,734
10	0,763	0,781	0,762	0,781
Mean	0,765	0,764	0,762	0,760
Stdev	0,003	0,021	0,003	0,024

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

I. Output AUC Metode Analisis Diskriminan Kernel

Data *Imbalanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,593	0,605	0,594	0,642
2	0,598	0,605	0,597	0,605
3	0,623	0,602	0,637	0,566
4	0,602	0,622	0,598	0,571
5	0,615	0,536	0,600	0,541
6	0,608	0,571	0,634	0,576
7	0,603	0,526	0,603	0,480
8	0,591	0,662	0,592	0,662
9	0,603	0,583	0,588	0,625
10	0,598	0,532	0,588	0,573
Mean	0,604	0,584	0,603	0,584
Stdev	0,009	0,042	0,017	0,050

Data *Balanced*

Fold	Semua Variabel		Variabel Signifikan	
	Training	Testing	Training	Testing
1	0,754	0,775	0,754	0,775
2	0,757	0,754	0,758	0,754
3	0,756	0,756	0,757	0,756
4	0,758	0,751	0,757	0,751
5	0,756	0,750	0,756	0,750
6	0,756	0,763	0,756	0,763
7	0,755	0,768	0,755	0,768
8	0,758	0,770	0,759	0,770
9	0,757	0,744	0,758	0,744
10	0,758	0,740	0,758	0,740
Mean	0,757	0,757	0,757	0,757
Stdev	0,001	0,011	0,001	0,011

Lampiran 5 Output Nilai Ketepatan Klasifikasi (Lanjutan)

J. Output Model Regresi Logistik Ridge

Data *Imbalanced* (Fold ke-4)

```
Coefficients:
    Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept) -0.060168        NA          NA          NA         NA
SD           0.001877  0.056308    1.263211      0.045   0.964446
Poskesdes    1.730847  2.242696    1.177842      1.904   0.056901 .
Bidan       -1.631530 -2.257588    1.151217     -1.961   0.049874 *
Listrik     -2.009829 -5.309797    1.277337     -4.157   3.23e-05 ***
Toko        -0.278917 -6.439358    1.161189     -5.545   2.93e-08 ***
Jarak        0.017372  4.322265    1.284413      3.365   0.000765 ***
Gizi.Buruk   0.092438  0.225734    1.263923      0.179   0.858254
PAD         -0.037904 -1.299025    1.216436     -1.068   0.285568
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge paramter: 0.07400673, chosen automatically, computed using 1 PCs

Degrees of freedom: model 4.093 , variance 2.445
```

Data *Balanced* (Fold ke-3)

```
Coefficients:
    Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept)   0.8646            NA          NA       NA        NA
SD           -0.1399       -5.7566      2.6014   -2.213    0.0269 *
Poskesdes     8.5155       15.2856      3.3527    4.559  5.14e-06 ***
Bidan        -9.5107      -16.3219      3.2515   -5.020  5.17e-07 ***
Listrik      -1.3788       -6.6516      3.6907   -1.802    0.0715 .
Toko         -0.9674      -27.3512      3.2244   -8.483   < 2e-16 ***
Jarak         0.2477       89.9842      5.6730   15.862   < 2e-16 ***
Gizi.Buruk   -0.4293       -1.1911      2.5001   -0.476    0.6338
PAD          -0.1523       -6.3389      2.4902   -2.546    0.0109 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge paramter: 0.0008114524, chosen automatically, computed using 7 PCs

Degrees of freedom: model 8.837 , variance 15.92
```

Lampiran 6 Output Pengujian Asumsi Analisis Diskriminan Kernel

A. Pengujian Mardia's Test

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : data5[, -9]

$`multivariateNormality`
             Test        Statistic p value Result
1 Mardia Skewness 89166.7928796931       0     NO
2 Mardia Kurtosis 908.757694417918       0     NO
3             MVN             <NA>    <NA>     NO
---------------------------------------
```

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : data5sign[, -4]
$`multivariateNormality`
             Test        Statistic p value Result
1 Mardia Skewness 52967.4556953527       0     NO
2 Mardia Kurtosis 913.934769139669       0     NO
3             MVN             <NA>    <NA>     NO
```

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : SMOTE[, -9]

   $`multivariateNormality`
             Test        Statistic p value Result
1 Mardia Skewness 107454.364977483       0     NO
2 Mardia Kurtosis 956.517738866302       0     NO
3             MVN             <NA>    <NA>     NO
---------------------------------------
```

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : SMOTEsign[, -8]
$`multivariateNormality`
             Test        Statistic p value Result
1 Mardia Skewness 102842.865339924       0     NO
2 Mardia Kurtosis  839.74786260028       0     NO
3             MVN             <NA>    <NA>     NO
```

Lampiran 6 Output Pengujian Asumsi Analisis Diskriminan Kernel (Lanjutan)

B. Pengujian Homogenitas

```
Box's M-test for Homogeneity of Covariance Matrices

data:  data.matrix(data5kab)[, -9]
Chi-Sq (approx.) = 469.55, df = 36, p-value < 2.2e-16
```

```
Box's M-test for Homogeneity of Covariance Matrices

data:  data.matrix(data5sign)[, -4]
Chi-Sq (approx.) = 366.56, df = 6, p-value < 2.2e-16
```

```
Box's M-test for Homogeneity of Covariance Matrices

data:  data.matrix(SMOTE)[, -9]
Chi-Sq (approx.) = 2785, df = 36, p-value < 2.2e-16
```

```
Box's M-test for Homogeneity of Covariance Matrices

data:  data.matrix(SMOTEsign)[, -8]
Chi-Sq (approx.) = 2555.9, df = 28, p-value < 2.2e-16
```

Lampiran 7 Surat Pernyataan

BADAN PUSAT STATISTIK
PROVINSI JAWA TIMUR

SENSUS EKONOMI

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	:	Canggih Shoffi Imanwardhani
Fakultas/Program Studi	:	Fakultas Matematika, Komputasi dan Sains Data / Statistika
N.R.P	:	06211440000051
Alamat Rumah	:	Kalikepiting Jaya 8 No.54, 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 :

"Pendekatan Teknik Sampling SMOTE dalam Menangani Imbalanced Data pada Klasifikasi 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