SKRIPSI

ANALISIS HASIL KLASIFIKASI REGRESI LOGISTIK DENGAN BERBAGAI METODE UNTUK MENANGANI DATA TIDAK SEIMBANG PADA KASUS PENYAKIT KRITIS

ALEXANDER S

NPM: 6161901096

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN

2023

FINAL PROJECT

ANALYSIS OF LOGISTIC REGRESSION CLASSIFICATION WITH MULTIPLE METHODS TO HANDLE IMBALANCED DATASET ON CRITICAL DISEASE CASES

ALEXANDER S

NPM: 6161901096

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY

2023

LEMBAR PENGESAHAN

ANALISIS HASIL KLASIFIKASI REGRESI LOGISTIK DENGAN BERBAGAI METODE UNTUK MENANGANI DATA TIDAK SEIMBANG PADA KASUS PENYAKIT KRITIS

Alexander S

NPM: 6161901096

Bandung, 18 Agustus 2023

Menyetujui,

Pembimbing 1	Pembimbing 2
Maria Anestasia, M.Si., M.Act.Sc.	Dr. Andreas Parama Wijaya

Ketua Penguji	Anggota Penguji
Benny Yong, Ph.D.	Jonathan Hoseana, Ph.D.

Mengetahui,

Ketua Program Studi

Dr. Livia Owen

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS HASIL KLASIFIKASI REGRESI LOGISTIK DENGAN BERBAGAI METODE UNTUK MENANGANI DATA TIDAK SEIMBANG PADA KASUS PENYAKIT KRITIS

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
18 Agustus 2023

Alexander S
NPM: 6161901096

ABSTRAK

Model-model pembelajaran mesin tersupervisi yang digunakan untuk melakukan klasifikasi umumnya dibentuk dengan data yang seimbang, padahal banyak masalah klasifikasi di dunia ini memiliki data yang tidak seimbang (*imbalanced dataset*), misalnya dalam kasus penyakit. Data yang digunakan pada eksperimen ini adalah data penyakit jantung koroner dan kanker paru-paru. Data yang tidak seimbang dapat menyebabkan model klasifikasi mengalami kesulitan dalam memprediksi kelas minoritas yang dapat menyebabkan model yang diperoleh tidak berperforma baik dalam matriks kebingungan (*confusion matrix*), yang terdiri dari ukuran *recall*, *precision*, *accuracy*, dan *F1-score*. Beberapa metode untuk mengatasi masalah tersebut adalah mengganti nilai ambang batas (*threshold value*) pada model tertentu, menggunakan teknik pengambilan ulang sampel (*resampling*) pada data *training*, dan menggunakan model berbobot (*weighted model*). Eksperimen dilakukan untuk melihat efek dari berbagai metode terhadap hasil klasifikasi ketika menggunakan data tidak seimbang dengan harapan dapat menyeimbangkan nilai *precision* dan *recall*.

Kata-kata kunci: data tidak seimbang, *resampling*, deteksi penyakit, pembelajaran mesin tersupervisi, klasifikasi

ABSTRACT

Supervised machine learning models for classification are usually trained with a balanced dataset, while many classification problems in reality deal with imbalanced datasets, as in disease cases. Datasets used in this experiment are those of coronary heart and lung cancer diseases. Imbalanced datasets could cause classification models to have difficulty in predicting the minority class, resulting in the model obtained not performing well in confusion matrix, which consists of metrics such as recall, precision, accuracy, and F1-score. Some methods to deal with such a problem are changing the threshold value on certain models, applying a resampling on the training data, and using weighted models. An experiment is carried out to examine the effect of multiple methods towards the classification results while using imbalanced dataset with the hope of balancing the precision and recall score.

Keywords: imbalanced dataset, resampling, disease detection, supervised machine learning, classification

KATA PENGANTAR

Puji syukur diberikan kepada Tuhan Yang Maha Esa, atas berkat dan rahmat yang diberikan, penulis dapat menyelesaikan menulis skripsi ini. Skripsi dengan judul "Analisis Hasil Klasifikasi Regresi Logistik dengan Berbagai Metode untuk Menangani Data Tidak Seimbang pada Kasus Penyakit Kritis" merupakan salah satu syarat untuk memperoleh gelar Strata-1 Program Studi Matematika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan, Bandung.

Dalam penulisan skripsi ini hingga akhir, penulis mendapatkan bantuan dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih yang tulus kepada:

- Anggota-anggota keluarga yang selalu memberikan doa dan dukungan dalam proses penulisan skripsi.
- Ibu Maria Anestasia, M.Si., M.Act.Sc. dan Bapak Dr. Andreas Parama Wijaya selaku dosen pembimbing yang memberikan bantuan, arahan, dan saran dalam membimbing penulis menyelesaikan skripsi.
- Bapak Benny Yong, Ph.D. dan Bapak Jonathan Hoseana, Ph.D. selaku dosen penguji yang memberikan saran dan komentar sehingga skripsi ini menjadi lebih baik.
- Dosen-dosen Progam Studi Matematika Universitas Katolik Parahyangan yang memberikan ilmu dan wawasan selama proses pembelajaran penulis.
- Teman-teman penulis yang memberikan dukungan dan membantu dalam proses penulisan skripsi.
- Pihak-pihak lain yang membantu dalam proses penulisan skripsi.

Akhir kata, semoga skripsi ini bermanfaat bagi orang-orang yang membacanya.

Bandung, 18 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Di dunia ini, terdapat banyak sekali penyakit, baik yang mematikan maupun tidak. Contoh dari penyakit yang memakan banyak korban adalah penyakit jantung koroner [1] dan kanker paru-paru [2]. Penyakit jantung koroner, yang disebut juga penyakit jantung iskemik, adalah penyakit yang disebabkan terjadinya penyumbatan pembuluh arteri koroner atau penyempitan karena endapan lemak, yang secara bertahap menumpuk di dinding arteri. Penyakit jantung koroner termasuk penyakit yang berbahaya karena dapat menyebabkan serangan jantung dan komplikasi serius lainnya, seperti nyeri dada, aritmia, dan gagal jantung yang dapat berakibat fatal seperti menyebabkan kematian. Di lain pihak, kanker paru-paru dapat menyebabkan berbagai masalah pernapasan seperti penyumbatan saluran udara utama yang dapat menyebabkan cairan menumpuk di sekitar paru-paru yang dapat menyebabkan kematian. Terdapat banyak faktor yang dapat menyebabkan kedua penyakit tersebut. Beberapa faktor yang mempengaruhi kemungkinan seseorang terkena penyakit jantung koroner atau kanker paru-paru adalah kebiasaan merokok, umur, dan jenis kelamin.

Pemeriksaan penyakit jantung dan kanker memerlukan biaya yang relatif mahal serta berbagai alat untuk memberikan hasil yang akurat, sehingga tidak semua orang dapat memperoleh akses untuk pemeriksaan kedua penyakit tersebut. Alternatif lain yang berfungsi sebagai pemeriksaan awal adalah penggunaan suatu model untuk memprediksi risiko seseorang terkena penyakit berdasarkan faktor-faktor yang telah disebutkan di atas. Cara ini bukan untuk menggantikan diagnosis dokter, tetapi untuk pemeriksaan awal yang diharapkan dapat mengefisienkan biaya pemeriksaan sekaligus menghasilkan prediksi yang akurat [3]. Pada skripsi ini, dibentuk suatu model klasifikasi untuk memprediksi risiko seseorang akan terkena kedua penyakit tersebut dari berbagai faktor, dengan tujuan menginformasikan kepada orang tersebut hal-hal apa saja yang dapat meningkatkan risiko terkena kedua penyakit tersebut, sehingga orang tersebut dapat berusaha untuk mencegah atau mengurangi risiko masing-masing.

Model regresi logistik adalah salah satu model yang paling populer dalam ilmu kesehatan [4], dan digunakan pada penelitian ini. Model regresi logistik memberikan prediksi peluang apakah suatu kejadian terjadi (kelas 1) atau tidak (kelas 0), bukan mengklasifikasi langsung kelas dari suatu data. Umumnya hasil prediksi peluang yang lebih dari sama dengan 50% akan diklasifikasi sebagai kelas 1 dan peluang yang kurang dari 50% akan diklasifikasi sebagai kelas 0 [5]. Regresi logistik populer dalam melakukan klasifikasi karena secara umum regresi logistik mudah untuk diimplementasikan dan model yang diperoleh juga mudah untuk diinterpretasikan. Akan tetapi,

terdapat juga kelemahan dari regresi logistik, yaitu tidak ada asumsi yang dibuat mengenai distribusi variabel bebas, dan data yang kecil dapat memberikan hasil yang kurang tepat [4]. Variabel-variabel bebas sebaiknya tidak berkorelasi satu sama lain karena dapat menyebabkan masalah dalam estimasi model. Oleh karena itu, akan dilakukan juga analisis komponen utama (PCA) dengan tujuan untuk mengurangi korelasi antara variabel-variabel bebas (multikolinearitas) tersebut, sehingga akan memberikan estimasi yang lebih baik atau lebih stabil untuk parameter-parameter regresinya, dengan harapan dapat meningkatkan performa model dalam melakukan klasifikasi [6].

Salah satu permasalahan utama lainnya adalah data penyakit, terutama penyakit yang berbahaya, umumnya merupakan data yang tidak seimbang, di mana banyaknya observasi untuk salah satu kelas jauh lebih sedikit dibandingkan kelas lainnya. Hal tersebut bermasalah karena banyak model prediksi yang ada berasumsi bahwa data yang dimiliki seimbang, sedangkan banyak data di kehidupan nyata yang tidak seimbang [7]. Sebagai contoh, data kanker paru-paru yang digunakan pada skripsi ini telah dibentuk model regresi logistik pada penelitian sebelumnya, tetapi pada penelitian tersebut hanya digunakan nilai akurasi (*accuracy*) sebagai ukuran dalam menentukan kebaikan model yang diperoleh [5]. Jika dilihat hanya dari nilai akurasi, dapat dikatakan bahwa model tersebut memberikan hasil yang bagus. Akan tetapi, pada skripsi ini akan ditunjukkan bahwa akurasi tidak dapat dijadikan ukuran yang baik dalam menentukan kebaikan suatu model ketika digunakan data tidak seimbang. Oleh karena itu, akan dilakukan beberapa cara seperti *resampling* pada data *training*, penggantian ambang batas model klasifikasi, dan konstruksi model berbobot (*weighted model*) untuk menangani masalah data tidak seimbang. Jadi, pada penelitian ini akan digunakan model regresi logistik dan dibantu oleh PCA untuk meningkatkan hasil model serta beberapa cara untuk menangani data tidak seimbang.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan sebelumnya, berikut rumusan masalah pada skripsi ini.

1. Bagaimana aplikasi regresi logistik dalam memprediksi data kasus penyakit?
2. Bagaimana cara menangani masalah multikolinearitas dengan analisis komponen utama dalam membangun model prediksi?
3. Bagaimana cara menangani masalah data tidak seimbang dalam membangun model prediksi?
4. Bagaimana cara menentukan apakah model yang dimiliki sudah baik?

1.3 Tujuan

Berdasarkan rumusan masalah yang telah diberikan di atas, tujuan dari makalah skripsi ini adalah

1. mengaplikasikan dan menganalisis hasil model regresi logistik dalam memprediksi data kasus penyakit,
2. mengetahui cara menangani masalah multikolinearitas dengan analisis komponen utama dalam membangun model prediksi,

3. mengetahui cara menangani masalah data tidak seimbang dalam membangun model prediksi,
4. menentukan apakah model yang dimiliki sudah baik atau belum dengan *confusion matrix*.

1.4 *State of the Art*

Regresi logistik merupakan salah satu model klasifikasi yang paling populer, karena mudah untuk diimplementasikan, serta hasil yang diperoleh juga mudah untuk diinterpretasikan. Akan tetapi, regresi logistik memiliki beberapa kelemahan, salah satunya adalah variabel-variabel bebas sebaiknya tidak berkorelasi tinggi karena dapat menyebabkan masalah dalam estimasi. Salah satu cara untuk mengatasi hal tersebut adalah melakukan analisis komponen utama dengan tujuan mengurangi multikolinearitas antara variabel-variabel bebas [6, hlm. 381].

Penentuan kebaikan suatu model klasifikasi tidak cukup dari satu atau dua ukuran saja. Salah satu ukuran yang dapat digunakan adalah ukuran *accuracy* [5]. Akan tetapi, ukuran *accuracy* tidak selalu baik digunakan, terutama ketika datanya tidak seimbang antara banyaknya observasi kelas 0 dan kelas 1 [8]. Data yang tidak seimbang membawa banyak masalah dalam masalah klasifikasi karena dapat menyebabkan banyak model klasifikasi menjadi cenderung mengklasifikasi observasi sebagai kelas mayoritas. Hal tersebut dapat menyebabkan nilai *accuracy* yang tinggi, tetapi jika dilihat dari hasil klasifikasi, sering sekali banyak kesalahan dalam mengklasifikasi kelas minoritas. Ukuran lain yang dapat digunakan untuk mengatasi masalah tersebut adalah ukuran *F1-score* serta dengan bantuan *confusion matrix*.

Pada penelitian sebelumnya, data yang digunakan hanyalah data kanker paru-paru [5], dan pada skripsi ini digunakan juga data penyakit jantung koroner untuk membandingkan hasil dari model untuk kedua data. Pada penelitian tersebut juga tidak diatasi masalah data tidak seimbang. Oleh karena itu, pada skripsi ini dilakukan berbagai metode untuk mengatasi masalah data tidak seimbang, yaitu menggunakan ambang batas yang berbeda [9], menggunakan teknik *resampling* pada data *training* [10], serta dengan menggunakan model regresi logistik berbobot [7].