**SKRIPSI**

# IMPLEMENTASI REGRESI LOGISTIK BINER DENGAN REGULARISASI LASSO DALAM MENANGANI MASALAH OVERFITTING (STUDI KASUS: KANKER PAYUDARA WISCONSIN)

**SINDI LESTARI**

**21106010036**



**PROGRAM STUDI MATEMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA**

**YOGYAKARTA**

**2025**

# IMPLEMENTASI REGRESI LOGISTIK BINER DENGAN REGULARISASI LASSO DALAM MENANGANI MASALAH OVERFITTING
# (STUDI KASUS: KANKER PAYUDARA WISCONSIN)

Skripsi

Untuk memenuhi sebagian persyaratan

mencapai derajat Sarjana S-1

Program Studi Matematika



diajukan oleh

**SINDI LESTARI**

**21106010036**

Kepada

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA

YOGYAKARTA

2025

# SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi / Tugas Akhir
Lamp :

Kepada
Yth. Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta
di Yogyakarta

*Assalamu'alaikum wr. wb.*

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka kami selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama : Sindi Lestari
NIM : 21106010036
Judul Skripsi : Implementasi Regresi Logistik Biner dengan Regularisasi LASSO dalam menangani Masalah Overfitting
Studi Kasus: Kanker Payudara Wisconsin

sudah dapat diajukan kembali kepada Program Studi Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Matematika.

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara tersebut di atas dapat segera dimunaqasyahkan. Atas perhatiannya kami ucapkan terima kasih.

*Wassalamu'alaikum wr. wb.*



Yogyakarta, 6 Mei 2025
Pembimbing

Sri Utami Zuliana, S.Si., M.Sc., Ph.D.
NIP. 19741003 200003 2 002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

# PENGESAHAN TUGAS AKHIR

Nomor : B-1040/Un.02/DST/PP.00.9/06/2025

Tugas Akhir dengan judul : Implementasi Regresi Logistik Biner dengan Regularisasi LASSO dalam Menangani Masalah Overfitting
Studi Kasus: Kanker Payudara Wisconsin

yang dipersiapkan dan disusun oleh:

Nama : SINDI LESTARI
Nomor Induk Mahasiswa : 21106010036
Telah diujikan pada : Selasa, 20 Mei 2025
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

## TIM UJIAN TUGAS AKHIR

Ketua Sidang

Sri Utami Zuliana, S.Si., M.Sc., Ph.D.
SIGNED

Valid ID: 68492de263937

Penguji I

Lilih Deva Martias, M.Sc.
SIGNED

Valid ID: 683d10f4469b3

Penguji II

Sri Istiyarti Uswatun Chasanah, M.Si.
SIGNED

Valid ID: 68474ee047be0

Yogyakarta, 20 Mei 2025
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 684a3c1710b16

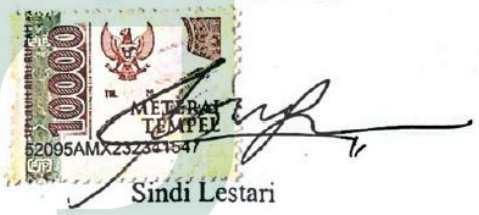
## SURAT PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

| | |
|---|---|
| Nama | : Sindi Lestari |
| NIM | : 21106010036 |
| Program Studi | : Matematika |
| Fakultas | : Sains dan Teknologi |

Dengan ini menyatakan bahwa isi skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu Perguruan Tinggi dan sesungguhnya skripsi ini merupakan hasil pekerjaan penulis sendiri sepanjang pengetahuan penulis, bukan duplikasi atau saduran dari karya orang lain kecuali bagian tertentu yang penulis ambil sebagai bahan acuan. Apabila terbukti pernyataan ini tidak benar, sepenuhnya menjadi tanggung jawab penulis.

Yogyakarta, 7 Mei 2025




Sindi Lestari

# HALAMAN PERSEMBAHAN

Tiada lembar yang paling indah dalam laporan skripsi ini kecuali lembar persembahan,

Skripsi ini saya persembahkan sebagai bentuk tanda cinta dan terima kasih yang mendalam kepada orang tua, Kakak dan keluarga besar yang selalu memberikan dukungan, semangat, serta doa yang tak henti mengiringi proses penyelesaian skripsi ini.

Tak lupa, persembahan ini juga untuk **Almamater tercinta Universitas Islam Negeri Sunan Kalijaga Yogyakarta** yang telah menjadi ruang tumbuh dan belajar selama masa studi.

# MOTTO

*Orang lain tidak akan bisa paham struggle dan masa sulitnya kita, yang mereka tahu hanya sebagian success stories-nya saja. Jadi, berjuanglah untuk diri sendiri meskipun tidak akan ada yang tepuk tangan. Kelak diri kita di masa depan akan sangat bangga dengan apa yang kita perjuangkan hari ini.*

# PRAKATA

Puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan karunianya yang tak ternilai harganya berupa keimanan, kesabaran, kekuatan dan kelancaran. Shalawat serta salam semoga selalu tercurahkan kepada Nabi Muhammad SAW sehingga penulis dapat menyelesaikan skripsi ini yang berjudul "Implementasi Regresi Logistik Biner dengan Regularisasi LASSO dalam Menangani Masalah Overfitting".

Penulis menyadari bahwa proses penyusunan skripsi ini tidak lepas dari berbagai tantangan. Oleh karena itu, penulis menyampaikan rasa terima kasih yang mendalam kepada semua pihak yang telah memberikan dukungan dan motivasi hingga skripsi ini dapat diselesaikan.

Penulis dengan tulus menyampaikan rasa terima kasih yang mendalam serta penghargaan yang sebesar-besarnya kepada:

1. Bapak Prof. Noorhaidi, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor UIN Sunan Kalijaga Yogyakarta.
2. Ibu Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
3. Ibu Dr. Epha Diana Supandi, S.Si., M.Sc., selaku Ketua Program Studi Matematika Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta dan selaku Dosen Penasihat Akademik.
4. Ibu Sri Utami Zuliana, S.Si., M.Sc., Ph.D., selaku pembimbing, atas waktu dan kesempatan yang telah diberikan untuk membimbing dan berbagi ilmu.

Arahan serta petunjuk dari Ibu sangat membantu dalam proses penulisan ini dan menjadi pengalaman berharga untuk perjalanan ke depan. Penulis juga menyampaikan permohonan maaf atas segala kekhilafan yang mungkin terjadi selama masa bimbingan.

5. Seluruh dosen program studi Matematika dan staf fakultas Sains dan Teknologi yang senantiasa memberikan ilmu dan layanan terbaik kepada penulis dari awal hingga akhir perkuliahan.

6. Cinta Pertamaku, Ayahanda Alm. Sugiman. Meskipun ayah telah tiada, harapan terakhir yang ayah titipkan kini telah terpenuhi. Semoga hasil dari apa yang penulis capai ini dapat menjadi kebanggaanmu di alam sana, dan semoga ayah selalu tenang di sisi-Nya. Terima kasih telah menjadi ayah yang luar biasa, dan dedikasi ini adalah wujud kecil dari cinta dan penghargaan yang tidak akan pernah cukup untuk membalas segalanya.

7. Pintu Surgaku, Ibunda Linlin, terima kasih atas segala cinta tanpa syarat yang selalu diberikan, terima kasih atas doa-doa yang tak pernah putus, yang menjadi pengiring langkah penulis dalam setiap perjuangan. Terima kasih telah menjadi tempat berpulang, tempat berbagi cerita, dan sumber ketenangan di tengah perjalanan hidup ini. Segala pengorbanan dan kasih sayang yang telah tercurah tidak akan pernah bisa penulis balas sepenuhnya, tetapi penulis berharap apa yang telah penulis capai dapat menjadi alasan untuk tersenyum bangga.

8. Kakak-kakak tersayang, Dimas Nur Syahbani, Nur Alimansyah dan Putriana, terima kasih atas segala cinta, perhatian, dan dukungan yang tak pernah putus kalian berikan. Penulis sangat bersyukur memiliki kakak-kakak yang penuh

kasih dan peduli. Semoga apa yang telah penulis capai saat ini dapat menjadi salah satu bentuk rasa syukur atas segala kebaikan kalian. Terima kasih karena selalu percaya pada penulis.

9. Mas Tujiran, Mba Sumini, dan Novi Rahmadinika, penulis mengucapkan terima kasih yang sebesar-besarnya karena telah menjadi bagian penting dalam pendidikan penulis. Kalian telah menyayangi penulis seperti anak dan saudara sendiri, meskipun tidak ada ikatan darah di antara kita yang merupakan anugerah luar biasa dalam hidup penulis. Semoga Allah SWT. membalas segala kebaikan kalian dengan keberkahan dan kebahagiaan yang melimpah.

10. Keluarga Besar Udiardjo, terima kasih atas dukungan, dan kebersamaan yang selalu kalian berikan. Semoga kebaikan dan kehangatan yang selalu kalian berikan menjadi berkah yang terus melimpah dalam kehidupan kalian.

11. Muhammad Fauzi Nur Robiultsani, terima kasih telah menjadi bagian yang tak terpisahkan dalam perjalanan ini. Terima kasih untuk setiap waktu yang telah diluangkan, setiap dukungan yang diberikan, dan setiap kata penyemangat yang selalu menguatkan penulis untuk tetap menikmati proses. Hadirmu menjadi sumber semangat di tengah kesibukan, menjadi tempat berbagi di saat penuh tekanan, dan menjadi pengingat untuk terus maju meski lelah menghadang.

12. Fita Dwi Aryani, terima kasih untuk selalu memberikan semangat, motivasi dan kesediaannya menjadi tempat berbagi cerita maupun keluh kesah dengan penuh perhatian. Terima kasih juga atas bantuan kecil maupun besar yang telah diberikan, baik dalam bentuk tindakan nyata maupun dukungan moral yang tak ternilai harganya.

13. Kak Linggar, terimakasih atas bantuan yang telah diberikan dalam penyusunan skripsi ini.

14. Lailatul Ulla, terima kasih atas semangat, bantuan, dan berjuang bersama dalam menyelesaikan skripsi ini.

15. Teman seperbimbingan, Muna, Laili dan Zaza, terima kasih atas dukungan, semangat, dan kebersamaan sepanjang proses penyusunan skripsi ini.

16. Semua pihak yang telah memberikan dukungan, bantuan, dan doa, meskipun tidak dapat disebutkan satu per satu. Kebaikan, perhatian, dan bantuan kalian sangat berarti dalam perjalanan penulis menyelesaikan studi ini. Semoga segala kebaikan yang diberikan dibalas dengan keberkahan dan kebahagiaan.

Penulis menyadari bahwa skripsi ini masih belum sempurna, sehingga penulis sangat mengharapkan kritik dan saran yang membangun. Akhir kata, penulis berharap semoga Allah SWT membalas semua kebaikan dari pihak-pihak yang telah membantu menyelesaikan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi semua orang, khususnya dalam pengembangan ilmu pengetahuan. *Aamiin*.



Yogyakarta, 20 Mei 2025

Penulis

# DAFTAR ISI

# DAFTAR TABEL

# DAFTAR GAMBAR

# DAFTAR LAMBANG

| | | |
|---|---|---|
| $\pi(x)$ | = | probabilitas kejadian sukses |
| $y$ | = | variabel respon |
| $x$ | = | variabel prediktor |
| $\beta_0$ | = | kontanta |
| $\beta_p$ | = | paramater sebanyak $p$ |
| $N$ | = | jumlah total percobaan |
| $p$ | = | jumlah variabel prediktor |
| $n$ | = | jumlah observasi |
| $j$ | = | indeks variabel prediktor |
| $i$ | = | indeks observasi |
| $\eta$ | = | prediktor linier |
| $g$ | = | *link function* |
| $\mu$ | = | rata-rata distribusi |
| $Var$ | = | varians |
| $\lambda$ | = | parameter penalti |
| $L$ | = | fungsi likelihood |
| $l$ | = | fungsi log likelihood |
| $\sum_{i=1}^{n}$ | = | penjumahan dari semua observasi $i$ dari 1 hingga $n$ |
| $\prod_{i=1}^{n}$ | = | perkalian dari semua observasi $i$ dari 1 hingga $n$ |
| $R^2$ | = | koefisien determinasi |
| $\partial$ | = | turunan |

| | | |
|---|---|---|
| $R^2$ | = | koefisien determinasi |
| $P$ | = | regularisasi |
| $\psi$ | = | fungsi penalti yang diterapkan pada koefisien $\beta_j$ |
| $l_p$ | = | fungsi likelihood penalti |
| $S(z_j, \lambda)$ | = | fungsi *soft-thresholding* untuk memperbarui koefisien $\beta_j$ |
| $t$ | = | indeks iterasi dalam coordinate descent |
| $w_i$ | = | bobot untuk observasi ke-$i$ |
| $z_j$ | = | nilai kuasi-gradien |
| $CV$ | = | *cross validation* |
| $F$ | = | *fold* dalam pembagian data |
| $K$ | = | jumlah *fold* dalam *k-fold cross validation* |
| $k$ | = | indeks *fold* dalam *k-fold cross validation* |

# INTISARI

**Implementasi Regresi Logistik Biner dengan Regularisasi LASSO dalam menangani Masalah Overfitting**

**(Studi Kasus: Kanker Payudara Wisconsin)**

SINDI LESTARI

21106010036

Regresi logistik biner sering digunakan untuk memprediksi variabel dengan dua kemungkinan hasil. Dalam penerapannya, overfitting dapat terjadi ketika model terlalu sesuai dengan data latih dan gagal memberikan prediksi yang akurat pada data baru. Untuk mengatasi overfitting, dapat digunakan teknik regularisasi seperti metode LASSO yang menambahkan fungsi penalti terhadap besar koefisien regresi. Penelitian ini bertujuan untuk mengetahui bagaimana penerapan metode LASSO membantu mengurangi overfitting dan meningkatkan kemampuan model dalam menghadapi data baru. Dengan metode LASSO, beberapa parameter regresi akan diberikan penalti, sehingga nilai koefisiennya menjadi lebih kecil atau bahkan nol, yang membuat model lebih sederhana dan tidak terlalu rumit. Hasil penelitian menunjukkan bahwa penerapan metode LASSO pada model regresi logistik biner meningkatkan akurasi dan kemampuan generalisasi model. Sebelum penerapan metode LASSO, akurasi pada data latih mencapai 100%, sedangkan pada data uji hanya 96%. Setelah diterapkan LASSO, akurasi pada data uji meningkat menjadi 98,23%, sementara akurasi pada data latih menurun menjadi 98,24%, yang menandakan bahwa model menjadi lebih seimbang dan tidak lagi terlalu menyesuaikan diri dengan data latih. Metode LASSO juga menyederhanakan model dengan mengeliminasi variabel yang tidak relevan dan menghasilkan model yang lebih efisien.

**Kata Kunci** : LASSO, Overfitting, Penalti, Regresi Logistik Biner, Regularisasi

# ABSTRACT

**Implementation of Binary Logistic Regression with LASSO Regularization to Handle Overfitting Problem**

**(Case Study: Breast Cancer Wisconsin)**

SINDI LESTARI

21106010036

Binary logistic regression is often used to predict variables with two possible outcomes. In application, overfitting can occur when the model overfits the training data and fails to provide accurate predictions on new data. To overcome overfitting, regularization techniques can be used such as the LASSO method which adds a penalty function to the size of the regression coefficients. This study aims to find out how the application of LASSO method helps to reduce overfitting and improve the ability of model to deal with new data. With the LASSO method, some regression parameters will be penalized, so that their coefficient values become smaller or even zero, which makes the model simpler and less complicated. The results show that applying the LASSO method to the binary logistic regression model improves the accuracy and generalization ability of the model. Before the application of the LASSO method, the accuracy on the training data reached 100%, while on the test data was only 96%. After applying LASSO, the accuracy on the test data increased to 98,23%, while the accuracy on the training data decreased to 98,24%, which indicates that the model becomes more balanced and no longer adjusts too much to the training data. The LASSO method also simplifies the model by eliminating irrelevant variables and producing a more efficient model.

**Keyword**: Binary Logistic Regression, LASSO, Overfitting, Penalty, Regularization

# BAB I

# PENDAHULUAN

## 1.1. Latar Belakang Masalah

Model linier merupakan salah satu metode yang banyak digunakan untuk memahami hubungan antara variabel respon dan variabel prediktor. Model ini mengasumsikan bahwa hubungan antara variabel-variabel tersebut bersifat linier serta bahwa error berdistribusi normal. Namun, dalam banyak kasus, variabel respon tidak selalu berbentuk data kontinu atau tidak memenuhi asumsi distribusi normal. Untuk menangani hal tersebut, dikembangkan *Generalized linear Model* (GLM) sebagai perluasan dari model linier klasik. GLM memungkinkan variabel respon mengikuti berbagai distribusi dari keluarga eksponensial, seperti binomial.

Salah satu bentuk GLM adalah analisis regresi logistik yang digunakan untuk menganalisis hubungan antara variabel respon kategorikal dengan satu atau lebih variabel prediktor. Model ini menghasilkan prediksi dalam bentuk probabilitas terjadinya suatu peristiwa, yang kemudian dapat diubah menjadi nilai kategori menggunakan fungsi logit (Hosmer Jr et al., 2013). Regresi logistik dapat dibagi menjadi tiga jenis utama, yaitu regresi logistik biner, regresi logistik multinomial dan regresi logistik ordinal. Regresi logistik biner digunakan ketika variabel respon hanya memiliki dua kemungkinan hasil. Regresi logistik multinomial digunakan ketika variabel respon memiliki lebih dari dua kategori yang tidak memiliki urutan. Sedangkan regresi logistik ordinal digunakan ketika variabel respon memiliki lebih dari dua kategori yang memiliki urutan atau tingkatan.

Dalam penerapan regresi logistik, overfitting dapat terjadi ketika model terlalu menyesuaikan data pelatihan sehingga hasilnya terlihat sangat akurat pada data tersebut, tetapi tidak mampu menghasilkan prediksi yang baik ketika diterapkan pada data baru. Overfitting sering terjadi ketika model memiliki terlalu banyak parameter atau variabel prediktor yang tidak relevan, yang memperburuk kemampuannya untuk memprediksi dengan akurat di luar data pelatihan. Salah Satu cara untuk menangani masalah overfitting adalah dengan menggunakan regularisasi.

Regularisasi adalah teknik yang menambahkan penalti pada model untuk mengurangi model menjadi terlalu rumit, sehingga mengurangi risiko overfitting. Regularisasi dapat diterapkan pada regresi logistik sehingga disebut regresi logistik terpenalti, yang menggabungkan konsep regresi logistik dengan penalti pada parameter untuk mengontrol besar koefisien-koefisien model. Teknik regularisasi yang umum digunakan dalam regresi logistik terpenalti, yaitu LASSO ($L_1$) dan *Ridge Regression* ($L_2$).

Metode *Least Absolute Shrinkage and Selection Operator* (LASSO) diperkenalkan pada tahun 1996 yang merupakan pengembangan dari regresi Ridge. LASSO adalah metode regresi yang dapat mengecilkan koefisien regresi hingga mencapai nol (Tibshirani 1996). Perbedaan utama antara analisis LASSO dan Ridge terletak pada jenis penalti yang digunakan. Pada LASSO, penalti diterapkan pada nilai absolut koefisien regresi, sedangkan pada Ridge, penalti dikenakan pada kuadrat koefisien regresi. Akibatnya, LASSO dapat mengecilkan beberapa koefisien regresi hingga tepat nol, sementara Ridge hanya dapat mendekatkannya ke nol tanpa benar-benar menjadi nol (Hastie et al. 2015). Salah satu keunggulan utama LASSO adalah kemampuannya untuk melakukan seleksi variabel prediktor, sehingga hanya variabel-variabel yang paling relevan yang dipertahankan dalam model (Robbani

et al. 2019).

Penelitian ini menggunakan data kanker payudara yang merupakan salah satu jenis kanker dengan angka kematian yang cukup tinggi yang menyerang perempuan. Jumlah kasus kanker ini cenderung meningkat setiap tahun, terutama di negara-negara berkembang, di mana keterlambatan dalam diagnosis seringkali menyebabkan keterlambatan pengobatan. Hal ini membuat kanker payudara sering ditemukan pada stadium lanjut (Sofa et al. 2024) . Salah satu kasus kanker payudara adalah kanker payudara wisconsin yang dikumpulkan oleh Dr. William H. Wolberg dan tim dari University of Wisconsin-Madison, Amerika Serikat (*UC Irvine Machine Learning Repository*). Kanker payudara terjadi akibat pertumbuhan sel yang tidak normal pada jaringan payudara. Sebagian besar kasus kanker payudara bermula dari sel-sel yang melapisi saluran (kanker duktal), sebagian lainnya dari lobulus (kanker lobular), dan sejumlah kecil dari jaringan lain (Achmad 2022). Dalam mendiagnosis kanker payudara terdapat banyak variabel yang berperan dalam menentukan keberadaan dan jenis kanker. Namun, dengan banyaknya variabel ini dapat menyebabkan model menjadi terlalu kompleks dan risiko terkena overfitting

Berdasarkan uraian di atas, penelitian ini bertujuan untuk menerapkan metode regularisasi LASSO pada regresi logistik biner yang diaplikasikan pada data Kanker Payudara Wisconsin dalam menangani masalah overfitting pada model dan menemukan variabel prediktor yang paling penting dalam memprediksi kanker payudara.



## 1.2. Rumusan Masalah

1. Bagaimana penerapan regularisasi LASSO terhadap model regresi logistik biner dalam menangani masalah overfitting?

2. Sejauh mana regularisasi LASSO dapat meningkatkan akurasi prediksi model regresi logistik biner dibandingkan dengan model tanpa regularisasi?

3. Bagaimana dampak penerapan regularisasi LASSO terhadap pemilihan variabel dalam model regresi logistik biner?

## 1.3. Batasan Masalah

1. Penelitian dibatasi dengan penggunaan model regresi logistik biner

2. Penelitian hanya berfokus pada metode regularisasi LASSO

3. Penelitian ini hanya menggunakan data kanker payudara wisconsin

4. Analisis dilakukan menggunakan *software R* versi 4.4.3

## 1.4. Tujuan Penelitian

1. Menerapkan regresi logistik biner dengan regularisasi LASSO untuk menangani masalah overfitting

2. Mengetahui sejauh mana regularisasi LASSO dapat meningkatkan akurasi prediksi model regresi logistik biner dibandingkan dengan model tanpa regularisasi

3. Mengetahui dampak penerapan regularisasi LASSO terhadap pemilihan variabel dalam model regresi logistik biner

## 1.5. Manfaat Penelitian

1. Penelitian ini diharapkan dapat memberikan manfaat baik di bidang akademik maupun praktis

2. Meningkatkan pemahaman tentang penerapan regresi logistik biner dengan regularisasi LASSO dalam menangani masalah overfitting pada data kanker payudara wisconsin.

3. Penelitian ini diharapkan dapat membantu mengoptimalkan model prediksi untuk kasus kanker payudara di masa depan, dengan menggunakan teknik LASSO untuk memilih variabel yang relevan.

4. Hasil penelitian ini dapat berguna bagi penelitian-penelitian selanjutnya yang melibatkan analisis regresi logistik biner dengan regularisasi LASSO untuk masalah serupa.

## 1.6. Sistematika Penulisan

Sistematika penulisan pada penelitian ini adalah sebagai berikut:

**BAB 1** : Bab ini membahas tentang latar belakang masalah, batasan masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, sistematika penulisan dan tinjauan pustaka.

**BAB 2** : Bab ini membahas landasan teori yang menjadi dasar acuan, meliputi konsep-konsep dan rancangan penelitian.

**BAB 3** : Bab ini membahas tentang metode penelitian, jenis penelitian dan sumber literatur, variabel data penelitian, langkah-langkah analisis dan diagram analisis data.

**BAB 4** : Bab ini dilakukan proses analisis data yang diikuti dengan interpretasi hasil secara mendalam dan terperinci.

**BAB 5** : Bab ini membahas tentang kesimpulan penelitian dan saran dari penulis.

## 1.7. Tinjauan Pustaka

Sebagai dasar teori dalam penulisan skripsi ini, penulis merujuk pada berbagai sumber seperti buku, makalah, dan jurnal. Berikut ini adalah tinjauan pustaka yang digunakan penelitian ini.

1. Buku karya (Tibshirani 1996) yang berjudul ***Regression Shrinkage and Selection via the LASSO*** membahas metode estimasi baru dalam analisis regresi bernama LASSO, atau *least absolute shrinkage and selection operator*, yang dirancang untuk meningkatkan akurasi prediksi dan interpretabilitas model. LASSO bekerja dengan meminimalkan jumlah kuadrat residual di bawah batasan jumlah nilai absolut dari koefisien, yang sering kali mengakibatkan beberapa koefisien menjadi nol, sehingga menciptakan model yang lebih sederhana dan dapat dipahami. Metode ini menggabungkan keunggulan dari teknik seleksi subset yang menghasilkan model yang mudah diinterpretasikan dan regresi ridge yang stabil, sehingga memberikan solusi yang lebih baik dalam situasi dengan banyak prediktor. Penulis juga membahas penerapan LASSO dalam berbagai model statistik dan menunjukkan potensi keunggulannya dibandingkan metode lainnya.
2. Jurnal karya (Padhilah et al. 2024) dalam judul ***Analisis Regresi Logistik Biner dengan Metode Group LASSO dalam Data Berdimensi Tinggi (Studi Kasus: Indeks Pembangunan Manusia Kota/Kabupaten di Jawa Barat***. Penelitian ini bertujuan untuk menganalisis faktor-faktor yang mempengaruhi Indeks Pembangunan Manusia (IPM) di Kota/Kabupaten Jawa Barat pada tahun 2020 dengan menggunakan regresi logistik biner dan metode Group LASSO. Data yang digunakan terdiri dari 27 amatan dan 40 variabel prediktor yang dikelompokkan menjadi enam kategori. Metode Group LASSO dipilih

karena kemampuannya dalam menangani data berdimensi tinggi dan menangani masalah multikolinieritas yang sering muncul ketika banyak variabel prediktor digunakan. Hasil analisis menunjukkan bahwa kelompok pendidikan, ekonomi, lingkungan, dan kependudukan adalah faktor-faktor yang paling berpengaruh terhadap capaian IPM, dengan akurasi model prediksi mencapai 100%.

3. Jurnal karya (Pak et al. 2025) yang berjudul ***Application of the LASSO Regularisation Technique in Mitigating Overfitting in Air Quality Prediction Models*** membahas penerapan teknik regularisasi LASSO dalam mengurangi overfitting pada model prediksi kualitas udara. Dalam penelitian ini, LASSO digunakan untuk memilih variabel prediktor yang paling relevan dan mengurangi kompleksitas model prediksi kualitas udara, yang seringkali dipengaruhi oleh banyak faktor. Overfitting menjadi masalah utama dalam prediksi kualitas udara, dan LASSO membantu menanganinya dengan menambahkan penalti pada koefisien variabel dalam model, sehingga meningkatkan kemampuan model untuk menggeneralisasi dan memprediksi dengan baik pada data baru. Teknik ini juga membantu dalam seleksi variabel dengan menyusutkan koefisien variabel yang tidak signifikan menjadi nol, membuat model lebih sederhana dan mudah diinterpretasikan. Hasil penerapan LASSO dalam kasus kualitas udara menunjukkan bahwa model prediksi yang menggunakan LASSO lebih akurat dibandingkan dengan model tanpa regularisasi, sehingga teknik ini terbukti efektif dalam meningkatkan prediksi kualitas udara dengan mengurangi overfitting.

4. Jurnal karya (Dwinata et al. 2021) yang berjudul ***Penalized logistic regression model to predict a results of RT-PCR by Using Blood Laboratory Test***

membahas pengembangan model regresi logistik terpenalti (LASSO dan elastic net) untuk memprediksi hasil tes RT-PCR COVID-19 menggunakan data dari 75 pasien di Rumah Sakit Israelita Albert Einstein, Brasil. Dengan 28 variabel prediktor, termasuk hasil tes darah dan usia pasien, penelitian ini bertujuan meningkatkan akurasi prediksi dan memungkinkan identifikasi cepat pasien terinfeksi COVID-19. Analisis menunjukkan bahwa model LASSO memiliki kinerja yang lebih baik dibandingkan elastic net, dengan akurasi 88% dan AUC 93%. Hasil ini menyoroti potensi penggunaan hasil tes laboratorium darah untuk diagnosis yang lebih efisien.

5. Jurnal karya (Rochayani et al. 2020) yang berjudul ***Two-stage Gene Selection and Classification for a High-Dimensional Microarray Data*** membahas tantangan klasifikasi kanker menggunakan data microarray yang memiliki dimensi tinggi, di mana jumlah gen yang dianalisis sangat banyak dibandingkan dengan jumlah observasi yang sedikit. Ketidakseimbangan ini dapat menyebabkan overfitting, sehingga kesulitan dalam model klasifikasi. Untuk menangani masalah ini, peneliti mengembangkan metode seleksi gen dua tahap yang pertama kali menggunakan regresi logistik dengan regulasi LASSO untuk mengurangi jumlah variabel, diikuti dengan penggunaan *Classification and Regression Tree* (CART) untuk seleksi lebih lanjut dan pembuatan model klasifikasi. Proses LASSO dalam studi ini juga menerapkan metode validasi silang (*cross validation*) untuk menentukan parameter regulasi yang optimal, sehingga dapat meminimalkan jumlah gen yang dipilih sekaligus mempertahankan akurasi klasifikasi yang tinggi.

# BAB V

# PENUTUP

## 5.1. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dijelaskan, maka disimpulkan beberapa hal sebagai berikut:

1. Penerapan regularisasi LASSO pada model regresi logistik biner terbukti efektif dalam menangani overfitting dengan menambahkan penalti $L_1$ terhadap koefisien regresi, sehingga beberapa koefisien mengecil atau menjadi nol. Hal ini menyederhanakan model dengan hanya mempertahankan variabel yang benar-benar berpengaruh, sehingga model tidak terlalu menyesuaikan diri dengan data latih dan mampu menggeneralisasi lebih baik terhadap data baru.

2. Regularisasi LASSO dapat meningkatkan akurasi prediksi model regresi logistik biner dengan mengurangi overfitting, sehingga model bisa lebih baik dalam memprediksi data baru. Tanpa regularisasi, model bisa terlalu menyesuaikan diri dengan data latih, seperti terlihat pada akurasi 100% untuk data latih tetapi hanya 96% untuk data uji, yang menunjukkan model kurang baik dalam menghadapi data baru. Setelah menggunakan LASSO, model menjadi lebih sederhana karena hanya mempertahankan variabel yang penting dan membuang yang kurang berpengaruh. Hasilnya, akurasi data uji meningkat menjadi 98.23%, sementara akurasi data latih tetap tinggi di 98.24%, menunjukkan bahwa model lebih stabil dan akurat dalam memprediksi data baru.

Oleh karena itu, LASSO tidak hanya membantu meningkatkan akurasi prediksi, tetapi juga menyederhanakan model.

3. Penerapan regularisasi LASSO dalam regresi logistik biner berdampak pada pemilihan variabel dengan cara menyusutkan koefisien regresi, bahkan menghilangkan beberapa variabel yang kurang penting. Hal ini terjadi karena LASSO mendorong beberapa koefisien menjadi nol, sehingga hanya variabel yang benar-benar berpengaruh terhadap prediksi yang dipertahankan dalam model. Berdasarkan penerapan model dengan metode LASSO pada *Breast Cancer Wisconsin* terdapat tiga belas variabel yang mempengaruhi kanker payudara yaitu *Texture Mean, Concave Point Mean, Fractal Dimension Mean, Radius SE,Smoothness SE, Compactness SE, Fractal Dimension SE, Radius Worst, Texture Worst, Smootness Worst, Concavity Worst, Concav Point Worst dan Symmetry Worst.* Variabel-variabel tersebut dianggap berpengaruh signifikan terhadap klasifikasi kanker jinak atau ganas, karena tetap dipertahankan dalam model. Sementara itu, 17 variabel lainnya memiliki koefisien nol, yang artinya tidak signifikan dalam membantu prediksi model dan otomatis dikeluarkan.

Model akhir hasil regularisasi LASSO dapat dituliskan dalam bentuk logit sebagai berikut.

$$\begin{aligned} g(x) = {} & (-0,3965) + 0,310X_2 + 0,041X_8 + (-0,005)X_{10} + 1,771X_{11} \\ & + 0,003X_{15} + (-0,283)X_{16} + (-0,159)X_{20} + 3,560X_{21} + 0,937X_{22} \\ & + 0,490X_{25} + 0,756X_{27} + 1,404X_{28} + 0,262X_{29} \end{aligned} \tag{5.1}$$

Model ini menunjukkan bahwa masing-masing koefisien merepresentasikan pengaruh dari setiap variabel terhadap log-odds kemungkinan seseorang di-

diagnosis kanker ganas. Koefisien positif menunjukkan bahwa semakin besar nilai variabel tersebut, semakin besar kemungkinan pasien menderita kanker ganas, sedangkan koefisien negatif menunjukkan pengaruh sebaliknya.

Dalam dunia medis, model seperti ini bisa membantu tenaga medis untuk lebih fokus pada variabel-variabel yang paling berpengaruh dalam membedakan kanker jinak dan ganas. Model ini bisa digunakan sebagai alat bantu dalam pengambilan keputusan diagnosis yang lebih cepat dan akurat, khususnya saat melakukan pemeriksaan awal atau skrining terhadap pasien.

## 5.2. Saran

Pada penelitian ini hanya membahas metode LASSO, oleh karena itu saran untuk penelitian selanjutnya adalah dapat menerapkan metode lain pada model regresi logistik seperti ridge dan elastic net atau dapat membandingkan metode-metode tersebut untuk melihat metode mana yang lebih baik dalam menangani masalah overfitting.

# DAFTAR PUSTAKA


Achmad, A. D. (2022), 'Klasifikasi breast cancer menggunakan metode logistic regression', *JTRISTE* **9**(1), 143–148.

Agresti, A. (2007), *An Introduction to Categorical data analysis*, John Wiley & Sons.

Athifaturrofifah, A., Goejantoro, R. & Yuniarti, D. (2020), 'Perbandingan pengelompokan k-means dan k-medoids pada data potensi kebakaran hutan/lahan berdasarkan persebaran titik panas', *Eksponensial* **10**(2), 143–152.

Burhan Nurgiyantoro, Gunawan, M. (2017), *Statistik Terapan untuk Penelitian Ilmu Sosial*, Gadjah Mada University Press.

Dwinata, A., Notodiputro, K. & Sartono, B. (2021), Penalized logistic regression model to predict a results of rt-pcr by using blood laboratory test, *in* 'IOP Conference Series: Materials Science and Engineering', Vol. 1115, IOP Publishing, p. 012087.

Erdianto, M. A. (2023), 'Perancangan model peramalan jangka pendek harga komoditas pertanian di indonesia menggunakan machine learning', *KLIK: Kajian Ilmiah Informatika Dan Komputer* **3**(4), 338–346.

Gill, J., Torres, M. & Pacheco, S. M. T. (2019), *Generalized linear models: a unified approach*, Vol. 134, Sage Publications.

Hasan, I. (2004), *Analisis Data Penelitian Dengan Statistik*, PT Bumi Aksara.

Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. (2005), 'The elements of statistical learning: data mining, inference and prediction', *The Mathematical Intelligencer* **27**(2), 83–85.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical learning with sparsity*, Taylor Francis Group.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied logistic regression*, John Wiley & Sons.

Hutagalung, M. A. I. et al. (2024), 'Penalized maximum likelihood estimation dengan algoritma gradient descent pada model regresi logistik multinomial', *IJM: Indonesian Journal of Multidisciplinary* **2**(6), 673–683.

Kurniawan, D., Wahyudi, M., Pujiastuti, L. & Sumanto, S. (2024), 'Deteksi dan prediksi cerdas penyakit paru-paru dengan algoritma random fores', *Indonesian Journal Computer Science* **3**(1), 51–56.

Latifah, T. & Anggitha, G. D. (2024), 'Implementasi metode random forest, knn (k-nearest neighbour), decision tree classification menggunakan machine learning untuk stroke prediction'.

Leidiana, H. (2013), 'Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor', *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic* **1**(1), 65–76.

McNeish, D. M. (2015), 'Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences', *Multivariate behavioral research* **50**(5), 471–484.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021), *Introduction to linear regression analysis*, John Wiley & Sons.

Padhilah, R., Herrhyanto, N. & Lukman, L. (2024), 'Analisis regresi logistik biner dengan metode group lasso dalam data berdimensi tinggi (studi kasus: Indeks pembangunan manusia kota/kabupaten di jawa barat)', *BIAStatistics Journal of Statistics Theory and Aplication* **18**(1), 1–14.

Pak, A., Rad, A. K., Nematollahi, M. J. & Mahmoudi, M. (2025), 'Application of the lasso regularisation technique in mitigating overfitting in air quality prediction models', *Scientific Reports* **15**(1), 547.

Pekhimenko, G. (2006), 'Penalizied logistic regression for classification', *Dept. Comput. Sci.., Univ. Toronto, Toronto, ON M5S3L1* .

Peng, Y. & Nagata, M. H. (2020), 'An empirical overview of nonlinearity and overfitting in machine learning using covid-19 data', *Chaos, Solitons & Fractals* **139**, 110055.

Pourahmadi, M. (2013), *High-dimensional covariance estimation: with high-dimensional data*, Vol. 882, John Wiley & Sons.

Prastowo, E. Y. (2021), 'Pengenalan jenis kayu berdasarkan citra makroskopik menggunakan metode convolutional neural network', *Jurnal Teknik Informatika Dan Sistem Informasi* **7**(2), 489–497.

Robbani, M., Agustiani, F. & Herrhyanto, N. (2019), 'Regresi least absolute shrinkage and selection operator (lasso) pada kasus inflasi di indonesia tahun 2014-2017', *Jurnal EurekaMatika* **7**(2), 1–16.

Rochayani, M. Y., Sa'adah, U. & Astuti, A. B. (2020), 'Two-stage gene selection and classification for a high-dimensional microarray data', *Jurnal Online Informatika* **5**(1), 9–18.

Salinas Ruíz, J., Montesinos López, O. A., Hernández Ramírez, G. & Crossa Hiriart, J. (2023), *Generalized Linear Mixed Models with Applications in Agriculture and Biology*, Springer Nature.

Sofa, T., Wardiyah, A. & Rilyani, R. (2024), 'Faktor risiko kanker payudara pada wanita', *Jurnal Penelitian Perawat Profesional* **6**(2), 493–502.

Sofiyat, A. I., Tjalla, A. & Mahdiyah, M. (2023), 'Pemodelan regresi logistik biner terhadap penerimaan pegawai di pt xyz jakarta', *Matematika Sains* **1**(1), 1–11.

Talita, A. S. (2016), 'Klasifikasi wisconsin diagnostic breast cancer data dengan menggunakan sequential feature selection dan possibilistic c-means', *Jurnal Ilmiah Komputasi* **15**(1), 47–52.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Usman, H. & Akbar, P. S. (2022), *Metodologi Penelitian Sosial (Edisi Ketiga)*, Bumi Aksara.

Wijiyanto, W., Pradana, A. I., Sopingi, S. & Atina, V. (2024), 'Teknik k-fold cross validation untuk mengevaluasi kinerja mahasiswa', *Jurnal Algoritma* **21**(1), 239–248.