TESIS

PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTIC REGRETION*

Disusun oleh:

Nama : **Winarnie**
NIM : **21.55.2151**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023

TESIS

PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTIC REGRETION*

COMPARISON OF DIMENSION REDUCTION METHODS IN LOGISTIC REGRETION ALGORITHMS

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Winarnie**
NIM : **21.55.2151**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2023

HALAMAN PENGESAHAN

PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTIC REGRETION*

COMPARISON OF DIMENSION REDUCTION METHODS IN LOGISTIC REGRETION ALGORITHMS

Dipersiapkan dan Disusun oleh

Winarnie

21.55.2151

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Sabtu, 5 Agustus 2023

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 5 Agustus 2023
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTIC REGRETION*

COMPARISON OF DIMENSION REDUCTION METHODS IN LOGISTIC REGRETION ALGORITHMS

Dipersiapkan dan Disusun oleh
Winarnie
21.55.2151

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Sabtu, 5 Agustus 2023

Pembimbing Utama	**Anggota Tim Penguji**
Prof. Dr. Kusrini, M.Kom	**Dr. Kumara Ari Yuana, S.T., M.T**
NIK. 190302106	**NIK. 190302575**
	M. Hanafi, S.Kom., M.Eng., Ph.D.
Pembimbing Pendamping	**NIK. 190302024**
Anggit Dwi Hartanto, M.Kom.	**Prof. Dr. Kusrini, M.Kom**
NIK. 190302163	**NIK. 190302106**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 5 Agustus 2023
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Winarnie
NIM : 21.55.2151
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Perbandingan Metode Reduksi Dimensi Pada Algoritma *Logistic Regretion*

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom
Dosen Pembimbing Pendamping : Anggit Dwi Hartanto, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 5 Agustus 2023
Yang Menyatakan,

Winarnie

HALAMAN PERSEMBAHAN

Alhamdulillahirobbil'alamin, puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa, yang telah melimpahkan segala karunia, rahmat dan hidayah-Nya karena dengan seizin-Nyalah penulis, dapat menyelesaikan penyusunan tesis ini dengan judul "PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTIC REGRETION* " dapat diselesaikan dengan baik.

Saya persembahkan karya ini untuk Ibu dan Bapak yang banyak memberikan kasih saying dan doa. Suami dan anak – anak tercinta serta adik – adikku atas semua doa , dukungan dan cinta yang tak terbatas.

Terima kasih yang tak terhingga untuk Ibu prof. Dr. Kusrini. M.Kom dan bapak Anggit Dwi Hartanto, M.Kom selaku Pembimbing Kedua yang dengan penuh kesabaran menuntun dan membimbing penulis hingga selesainya tesis ini dengan baik.

Ucapan terima kasih untuk teman – teman Magister Teknik Informatika 2022 / angkatan 6 untuk setiap canda yang kita lalui dan atas solidaritas yang kompak sehingga masa kuliah selama ini sangat berarti.

HALAMAN MOTTO

MOTTO :

"Dan barangsiapa yang bertakwa kepada Allah, niscaya Allah menjadikan baginya kemudahan dalam urusannya." (Q.S At-Talaq: 4)

"Dan bertawakkallah kepada Allah. Dan cukuplah Allah sebagai Pemelihara." (Q.S Al-Ahzab: 3)

Ali bin Abi Thalib berkata: "Karunia Allah yang paling lengkap adalah kehidupan yang didasarkan pada ilmu pengetahuan."

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Tuhan yang maha esa, yang telah melimpahkan rahmat dan karunia-Nya kepada penulis, sehingga penulis dapat menyelesaikan tesis ini dengan sebaik-baiknya. Penulis mengucapkan terima kasih yang tak terhingga kepada pihak yang telah mendukung diantaranya:

1. Rektor Universitas AMIKOM Yogyakarta, Bapak Prof. Dr. M. Suyanto,M.M. atas kesempatan yang telah diberikan kepada penulis untuk dapat mengikuti dan menyelesaikan pendidikan Program Magister Teknik nformatika Universitas AMIKOM Yogyakarta.
2. Prof. Dr. Kusrini. M.Kom. Sebagai Ketua Program Studi Magister Teknik Informatika sekaligus sebagai pembimbing utama, demikian juga kepada bapak Anggit Dwi Hartanto, M.Kom selaku Pembimbing Kedua yang dengan penuh kesabaran menuntun dan membimbing penulis hingga selesainya tesis ini dengan baik.
3. Dr. Kumara Ari Yuana, S.T., M.T dan M. Hanafi, S.Kom., M.Eng., Ph.D selaku dosen penguji, yang telah memberikan saran dan masukan serta arahan yang baik demi penyelesaian tesis ini.
4. Orang tua tercinta yang telah mendidik dengan penuh kasih sayang dan senantiasa memberikan doa dan semangat hingga selesainya tesis ini dengan baik.

5. Hery Oktafiandi. S.T, M. Eng, suami tercinta dan anak – anak yang memberikan doa dan dorongan sehingga penulis dapat menyelesaikan tesis ini.
6. Seluruh staf pegawai Program Studi Magister S2 Teknik Informatika Fakultas Ilmu Komputer dan Teknik Informatika, serta teman-teman seperjuangan mahasiswa angk 6 magister teknik informatika Amikom.

Penulis menyadari bahwa penelitian ini masih jauh dari kata sempurna, ini dikarenakan oleh keterbatasan, kemampuan dan pengetahuan penulis. Harapan penulis, semoga penelitian ini bermanfaat bagi penulis khususnya dan pembaca pada umumnya. Sekali lagi penulis mengucapkan terima kasih, semoga Tuhan yang maha esa membalas kebaikan yang telah diberikan. Amin.

Yogyakarta, 5 Agustus 2023

Penulis

DAFTAR ISI

DAFTAR TABEL

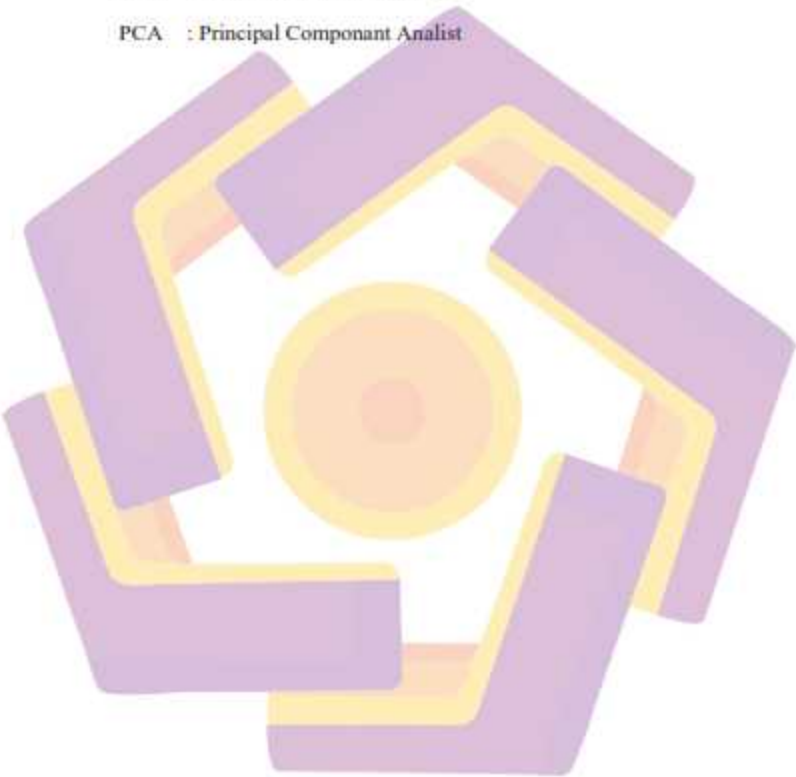
DAFTAR GAMBAR

DAFTAR ISTILAH

LR : Regresi Logistik

LDA : Linear Discriminant Analist

PCA : Principal Componant Analist



INTISARI

Regresi logistik merupakan classifier dengan metode supervised learning, yatiu suatu metode pengklasifikasian dimana setiap pengamatan memiliki variabel prediktor yang berkaitan dengan variabel respon. Regresi logistik sebagai classifier dalam analisis klasifikasi mengklasifikasikan subjek penelitian berdasarkan ambang (threshold) probabilitas, misalnya jika nilai probabilitas lebih dari 0.5, maka akan dibulatkan menjadi 1 yang artinya pengklasifikasian respon adalah di kelas event. Jika nilai probabilitas kurang dari atau sama dengan 0.5, maka akan dibulatkan menjadi 0 yang artinya pengklasifikasian respon adalah di kelas nonevent. Model regresi yang digunakan berdasarkan data training yang kemudian diaplikasikan ke data testing.

Penelitian ini menggunakan metode pengurangan dimensi untuk mengatasi kelemahan kinerja model regresi logistic pada data berdimensi tinggi. *Dataset* dengan sejumlah besar pengamatan menghadirkan tantangan baru dalam data, penambangan, analisis, dan klasifikasi. Metode statistik tradisional tidak bisa mengatasi karena peningkatan jumlah variabel yang terkait dengan setiap pengamatan yang dikenal sebagai data dimensi tinggi. Sebagian besar data sangat berlebihan dan dapat diabaikan untuk mengekstrak fitur kumpulan data. Proses pemetaan data berdimensi tinggi ke ruang berdimensi lebih rendah sedemikian rupa untuk membuang varians yang tidak informatif dari *dataset* atau menemukan subruang di mana data dapat dengan mudah dideteksi dikenal sebagai pengurangan dimensi.

Pada percobaan dataset pertama, hasil akurasi reduksi dimensi dengan LDA sebesar 98% , dengan PCA sebesar 96% dan tanpa reduksi dimensi sebesar 94%. Untuk dataset yang kedua akurasi dengan LDA sebesar 89%, dengan PCA sebesar 91%, dan tanpa reduksi 91%. Selain hasil akurasi, penelitian menggunakan hasil time running yang digunakan pada proses reduksi dimensi. Hasil yang didapat, dengan menggunakan LDA sebesar 19.2 ms dan menggunakan PCA sebesar 29.5 ms. Hasil perbandingan kinerja kedua reduksi dimensi yang digunakan pada regresi logistic menyatakan bahwa LDA lebih baik daripada PCA.

Kata kunci: Regresi Logistik, Data Berdimensi Tinggi, LDA, PCA,

ABSTRACT

Logistic regression is a classifier using the supervised learning method, which is a classification method where each observation has a predictor variable related to the response variable. Logistic regression as a classifier in classification analysis classifies research subjects based on a probability threshold, for example if the probability value is more than 0.5, it will be rounded to 1 which means that the response is classified in the event class. If the probability value is less than or equal to 0.5, it will be rounded to 0, which means that the response is classified in the noneevent class. The regression model used is based on training data which is then applied to data testing

. This study uses the dimension reduction method to overcome the performance weaknesses of the logistic regression model on high-dimensional data. Datasets with a large number of observations present new challenges in data mining, analysis and classification. Traditional statistical methods cannot cope due to the increasing number of variables associated with each observation which is known as high dimensional data. Most of the data is highly redundant and can be neglected to extract features of the data set. The process of mapping high-dimensional data to a lower-dimensional space in such a way as to remove uninformative variances from the dataset or find subspaces where data can be easily detected is known as dimensionality reduction.

In the first dataset experiment, the results of dimension reduction accuracy with LDA were 98%, with PCA of 96% and without dimension reduction of 94%. For the second dataset, the accuracy with LDA is 89%, with PCA is 91%, and without reduction is 91%. In addition to the results of accuracy, the research uses the results of time running which is used in the dimension reduction process. The results obtained, using LDA of 19.2 ms and using PCA of 29.5 ms. The results of the comparison of the performance of the two dimension reductions used in the logistic regression stated that LDA was better than PCA.

Keyword: Logistic regression, high-dimensional data, LDA, PCA

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Dalam data mining terdapat beberapa algortima diantaranya yaitu algortima klasifikasi. Klasifikasi merupakan proses menentukan kesamaan karakteristik dalam satu kelas yang bertujuan untuk memprediksi klas dari data yang belum diketahui labelnya. (Yunial, 2020). Algoritma yang bisa digunakan pada klasifikasi diantaranya adalah regresi logistic, naives bayes, decision tree, random -forest, k- nearest neigboard, dan artificial neural network. (Kaya & Yaganoglu, 2020) . Algoritma regresi logistic merupakan algoritma yang sudah biasa digunakan untuk menyelesaikan permasalahan klasifikasi.

Regresi Logistik adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara peubah respon (*dependent variable*) yang memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas (*independent variable*) berskala kategori atau interval (Hosmer andLemeshow, 2013). Regresi Logistik merupakan regresi non linear, digunakan untuk menjelaskan hubungan antara X dan Y yang bersifat tidak linear, ketidak normalan sebaran Y, keragaman respon tidak konstan yang tidak dapat dijelaskan dengan model regresi linear biasa (Begg, 2009).

Regresi logistik merupakan *classifier* dengan metode *supervised learning*, yatiu suatu metode pengklasifikasian dimana setiap pengamatan memiliki variabel prediktor yang berkaitan dengan variabel respon. Regresi logistik sebagai *classifier*

dalam analisis klasifikasi mengklasifikasikan subjek penelitian berdasarkan ambang (*threshold*) probabilitas, misalnya jika nilai probabilitas lebih dari 0.5, maka akan dibulatkan menjadi 1 yang artinya pengklasifikasian respon adalah di kelas event. Jika nilai probabilitas kurang dari atau sama dengan 0.5, maka akan dibulatkan menjadi 0 yang artinya pengklasifikasian respon adalah di kelas nonevent. Model regresi yang digunakan berdasarkan data training yang kemudian diaplikasikan ke data testing (Sihombing & Yuliati, 2021).

Penelitian ini menggunakaan model regresi logistic karena variabel dependen yang digunakan bersifat dummy. Varibel dummy adalah variabel yang memakai dua kemungkinan nilai sebagai contoh dilambangkan dengan angka 0 dan 1. (Lisnawati & Syafril, 2021). Ada beberapa kelemahan ketika regresi logistic dipakai pada data yang berdimensi tinggi diantaranya yaitu multikolinearitas, (P. Vatcheva & Lee, 2016) overfitting (Dewanta & Abdillah, n.d.), dan kompleksitas komputasi. Data berdimensi tinggi diartikan sebagai banyaknya variabel lebih besar dibandingkan dengan banyaknya observasi. (Bielza et al., 2011). Data dikatakan berdimensi tinggi ketika memiliki lebih dari tiga dimensi (Kantardzic, 2019). Dimensi data adalah atribut dari data. Data penelitian berdimensi tinggi adalah data yang memiliki banyak atribut, seperti pada penelitian ini water quality dataset yang digunakan memiliki 25 atribut dan breast cancer dataset memiliki 31 atribut. Menurut (Agarwal, 2014), data berdimensi tinggi memiliki jumlah dimensi lebih dari sepuluh.

Multikolinearitas merupakan masalah kolerasi antara variabel independen yang besar yang menyebabkan adanya kemungkinan kombinasi linear

antarvariabel. (P. Vatcheva & Lee, 2016) Overfitting pada regresi logistic terjadi karena nilai konstanta kooefisien variabel independen yang besar . (Dewanta & Abdillah, n.d.) Kompleksitas komputasi , contohnya perhitungan estimasi parameter menggunakkan metode numeric membutuhkan komputasi tinggi dalam menyelesaikan perhitungan rumit pada setiap prosesnya (Widhianingsih, 2018). Dari beberapa kekurangan model regresi logistic yang diterapkan pada data berdimensi tinggi maka diperlukan metode yang mengatasi masalah masalah tersebut.

Peneliti menggunakan metode pengurangan dimensi untuk mengatasi kelemahan kinerja model regresi logistic pada data berdimensi tinggi. *Dataset* dengan sejumlah besar pengamatan menghadirkan tantangan baru dalam data, penambangan, analisis, dan klasifikasi. Metode statistik tradisional tidak bisa mengatasi karena peningkatan jumlah variabel yang terkait dengan setiap pengamatan yang dikenal sebagai data dimensi tinggi. Sebagian besar data sangat berlebihan dan dapat diabaikan untuk mengekstrak fitur kumpulan data. Proses pemetaan data berdimensi tinggi ke ruang berdimensi lebih rendah sedemikian rupa untuk membuang varians yang tidak informatif dari *dataset* atau menemukan subruang di mana data dapat dengan mudah dideteksi dikenal sebagai pengurangan dimensi (Lim et al., 2010). Cara untuk menerapkan teknik pengurangan dimensi dapat menggunakan Ekstraksi fitur . Ekstraksi fitur adalah proses transformasi ruang yang mengandung banyak dimensi menjadi ruang dengan dimensi yang lebih sedikit. Pendekatan ini berguna ketika kita ingin menyimpan seluruh informasi

tetapi menggunakan lebih sedikit sumber daya saat memproses informasi. Teknik ekstrasi fitur dapat menggunakan metode PCA dan LDA (Fujiwara et al., 2022).

Linear Discriminant Analysis atau LDA adalah teknik reduksi dimensi. LDA digunakan sebagai langkah pra-pemrosesan dalam Pembelajaran Mesin dan aplikasi klasifikasi pola. Tujuan LDA adalah untuk memproyeksikan fitur-fitur dalam ruang berdimensi lebih tinggi ke ruang berdimensi lebih rendah untuk menghindari mengurangi sumber daya dan biaya dimensi. (Liang et al., 2010). LDA adalah teknik klasifikasi terawasi yang dianggap sebagai bagian dari pembuatan model pembelajaran mesin yang kompetitif. Kategori pengurangan dimensi ini digunakan di bidang seperti pengenalan citra dan analisis prediktif dalam pemasaran.

Analisis Komponen Utama (PCA) adalah teknik pembelajaran tanpa pengawasan yang populer untuk mengurangi dimensi data. Ini meningkatkan interpretabilitas namun, pada saat yang sama, meminimalkan kehilangan informasi. Ini membantu untuk menemukan fitur paling signifikan dalam kumpulan data dan membuat data mudah untuk diplot dalam 2D dan 3D. PCA membantu menemukan urutan kombinasi linier variabel. (Tsoulfidis & Athanasiadis, 2022).

Cahyani Refa dkk. Melakukan penelitian prediksi resiko penyakit diabetes dengan menggunakan *regresi logistic* (Cahyani et al., 2022). Modeliing menggunakan algoritma regresi logistic dengan perbandingan 70% data training dan 30 % data testing. Pada penelitian ini menggunakan 2 metode yaitu dengan normalisasi dan tanpa normalisasi. Hasil yang didapat adalah dengan normalisasi menghasilkan recall sebesar 55%, dan 43% tanpa normalisasi.

Alharthi dkk. Melakukan percobaan menggunakan algoritma *regresi logistic* dengan *dataset* berdimensi tinggi dengan metode *Lasso* dan *Alasso* , menghasilkan peningkatan akurasi sebesar 12.54% dengan metode *lasso* dan 9.58% dengan metode *alasso* (Alharthi et al., 2022). Penelitian menggunakan bahasa pemrograman R dengan metode split data 70% data training dan 30% data testing. Untuk mengetahui data yang hilang menggunakan metode split data yang berbeda-beda yaitu 10%, 20% dan 30%.

Penelitian lainnnya dilakukan oleh Rashid dkk, melakukan penelitian dengan menggunakan PLS-DA dihasilkan kesimpulan bahwa kesalahan klasifkasi lebih sedikit dibanding menggunakan *regresi logistic* biasa (Rashid et al., 2019). Metode yang digunakan berdasarkan jumlah variabel independen yng berbeda beda (p) dan ukuran sample yang berbeda (n), kemudian membandingkan model PLS-DA dengan PCA+LDA. Sehingga diketahui berapa rata-rata kesalahan klasifikasi dari masing-masing model.

Rana Debaradj melakukan penelitian dengan algoritma Random Forest dan *regresi logistic* metode PCA dan LDA pada dataset bunga iris dan mendapatkan kesimpulan bahwa metode LDA memberikan hasil yang lebih baik (Rana et al., 2020). Metode yang digunakan yaitu membagi data training 80% dan data testing 20% kemudian membandingkan dengan PCA dan LDA pada algoritma random forest. Sehingga dapat dilihat perbandingan dari kedua metode.

Penelitian untuk mendeteksi *malware* dengan akurasi tinggi menggunakan analis statistic PCA dan LDA dengan hasil F1 score sebesar 0.857 dengan PCA menggunakan 36 komponent dan nilai F1 scrore dengan LDA sebesar 0.925.

Penelitian ini dilakukan oleh (Şahin et al., 2021). Pada penelitian ini selain nilai akurasi , nilai *running time* pada masing-masing algoritma juga dibandingkan. Pada dataset yng relative kecil *running time* tidak ada perbedaan yang signifikan, untuk dataset yamg lain yang besar ada perbedaa yang terlihat. Hasil *running time* yang dihasilkan adalah ketika menggunakan PCA dengan 60 komponent dihasilkan waktu 37.27 detik, dengan 43 komponent dihasilkan waktu 20.08 detik, dan dengan menggunakan LDA 0.56 detik. Dengan menggunakan PCA keecpatan meningkat 4 kali lipat dan dengan menggunakan LDA kecepatan meningkat 175 kali.

Penelitian tentang klasifikasi pewarna dan konsentrasi pada serat , penelitian ini mencapai sparsity menggunakan *logistic regresi* dengan penyusutan absolut paling kecil dan operator seleksi (Lasso) (Rich et al., 2020). Metode yang digunakan menggunakan PCA dengan 4 komponen yang menghasilkan variabelitas sebesar 97.6 % dengan perincian komponen 1 sebesar 57.5%, komponen 2 sebesar 29.9%, komponen 3 sebesar 6.4% dan komponen 4 sebesar 33.8%. Dengan menggunakan LDA menghasilkan akurasi prediksi sebesar 89.3% dihitung dari data validasi dengan 28 kesalahan klasifikasi. Dengan menggunakan lasso nilai akurasinya 96.68 %. Kesimpulan dari penelitian ini adalah metode lasso lebih baik dibandingkan dengan PCA + LDA dalam mengklasifikasikan 29 klas serat akrilik biru.

Dari beberapa penelitian yang pernah dilakukan berkaitan dengan dataset berdimensi tinggi pada algoritma regresi logistik yang pernah diteliti sebelumnya, penulis akan meneliti dengan metode yang berbeda. Masalah yang terjadi adanya sejumlah fitur input pada dataset berdimensi tinggi yang menyebabkan buruknya

kinerja algoritma *regresi logistic* sehingga dibutuhkan teknik pengurangan dimensi.

Hasil *review* beberapa penelitian yang sudah ada mengenai metode LDA dan PCA, kedua metode ini memiliki persamaan yaitu kedua metode merupakan metode yang mereduksi dimensi data menjadi dimensi yang lebih kecil dengan cara membentuk sebuah persamaan yang terdiri atas kombinasi linear dari berbagai variabel (K.Raju et al., 2015), maka dalam penelitian ini dilakukan pengujian dengan algoritma *regresi logistic* menggunakan metode LDA dan PCA untuk mengetahui kinerja mana yang lebih baik dalam mengatasi masalah dataset yang memiliki dimensi tinggi dan untuk mengetahui tingkat akurasinya. Dalam penelitian ini dilakukan 2 percobaan yaitu dengan menggunakan jumlah dataset yang berbeda. Kontribusi pada penelitian ini adalah :

- Pada menggunakan metode reduksi dimensi menggunakan algoritma LDA dan PCA untuk mengekplorasi teknik pengurangan dimensi sehingga dapat menangani masalah data pada algoritma regresi logistic . Metode reduksi dimensi tersebut dengan menggunakan teknik menghapus fitur yang berkorelasi dari fitur yang memiliki korelasi tinggi untuk mengatasi masalah overfitting dan kesalahan korelasi yang disebabkan ketidakseimbangan data.
- Hasil dari penelitian ini menambahkan hasil *time running* selain hasil akurasi, presisi, recall, f1-score dan grafik roc. Dari hasil *time running* tersebut dapat dibandingkan hasil komputasi mana yang lebih baik dari reduksi dimensi yang digunakan.

1.2. Rumusan Masalah

Algoritma regresi logistic merupakan metode yang sederhana dalam menyelesaikan masalah pada analisa klasifikasi. Algoritma ini memiliki kelebihan yaitu hasil klasifikasi berupa probabilitas jadi untuk nilai prediksi menjadi lebih mudah. Kelemahan algoritma ini adalah ketika menggunakan data yang berdimensi tinggi. Untuk menangani kelemahan algorima ini maka perlu dilakukan dengan metode mereduksi dimensi. Metode mereduksi dimensi dapat dilakukan dengan menggunakan LDA dan PCA. Maka pada penelitian ingin diketahui beberapa masalah, yaitu :

a. Bagaimana cara kerja metode PCA dan metode LDA dapat menangani masalah data berdimensi tinggi ?
b. Berapa tingkat akurasi yang dihasilkan pada penerapan metode PCA dengan algoritma Regresi Logistic dibandingkan dengan menggunakan metode LDA pada algoritma regresi logistic ?
c. Faktor apa yang mempengaruhi akurasi algoritma klasifikasi harus menggunakan metode PCA dan LDA?

1.3. Batasan Masalah

Berdasarkan perumusan masalah terdapat batasan masalah, yaitu:

a. Data yang dipakai pada penelitian ini hanya berupa teks yaitu data yang digunakan terdiri dari satu set data *Water Quality* dari *Kaggle Repository*, dan satu set data *Breast Cancer Dataset* berasal dari *Kaggle Repository*.

b. Penelitian ini menggunakan metode *Principal Component Anaysis* (PCA) dan *Linear Diskriminant Analist* (LDA) sebagai metode untuk menyederhanakan variabel pada algoritma regresi logistic dengan dataset berdimensi tinggi menjadi dataset yang lebih sedehana tanpa menghilangkan fungsi .
c. Performa metode pada penelitian ini diukur berdasarkan nilai akurasi yang dihasilkan.
d. Pada penelitian ini, untuk membandingkan keberhasilan metode LDA dan PCA dalam menangani dataset yang berdimensi tinggi.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

a. Menggunakan metode LDA dan PCA pada regresi logistic dalam menangani masalah dataset yang berdimensi tinggi.
b. Untuk mendapatkan hasil yang terbaik dalam membandingkan metode LDA dan PCA pada algoritma regresi logistic pada dataset berdimensi tinggi.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah :

a. Sebagai referensi untuk metode LDA dan PCA khususnya pada algoritma regresi logistic
b. Mengetahui akurasi perhiungan hasil perbandingan kemampuan menangani dataset berdimensi tinggi pada metode LDA dengan PCA pada algoritma regresi logistik

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian yang membahas tentang *regresi logistic* , *principal component analist* dan *linear discriminant analis*t telah dilakukan , Şahin dkk. Melakukan penelitian ini bertujuan untuk mendeteksi malware dengan analisis statis teknik berbasis pembelajaran mesin. Karena sumber daya sistem perangkat seluler yang terbatas, principal analisis komponen dan analisis diskriminan linier, yang sering digunakan dalam masalah pembelajaran mesin dengan jumlah atribut yang tinggi, diterapkan untuk deteksi malware Android. Ketika hasil diperiksa, diamati bahwa teknik pengurangan dimensi memiliki efek positif pada kinerja klasifikasi di umum Pada penelitian ini menghasilkan PCA, teknik pengurangan dimensi, digunakan dalam deteksi malware Android berbasis pembelajaran mesin. Dengan cara ini pengurangan vektor akan menghasilkan pembelajaran mesin algoritma ditampilkan. Meskipun kinerja PCA menurut dataset berbeda, diamati bahwa itu mempengaruhi klasifikasi kinerja positif secara umum. Performa klasifikasi menurun ketika algoritma NB digunakan dengan PCA. Selain PCA, LDA, teknik pengurangan dimensi lain, digunakan dalam penelitian ini. Hasil yang diperoleh dengan LDA lebih baik daripada PCA menurut empat perbedaan dataset. Kedua teknik reduksi, PCA dan LDA berkurang secara signifikan waktu berjalannya sistem pendeteksi malware. Secara khusus, kinerja algoritma klasifikasi sangat meningkat dengan LDA, karena ukuran dataset tumbuh. Ini

diamati di kedua Malgenome-215 dan kumpulan data VirusShare (Şahin et al., 2021).

Penelitian tentang Penyakit arteri koroner adalah salah satu patologi kronis yang paling umum di dunia modern, menyebabkan kematian ribuan orang, baik di Amerika Serikat maupun di Eropa. Artikel ini melaporkan penggunaan teknik data mining untuk menganalisis populasi 10.265 orang yang dievaluasi oleh Departemen Ilmu Biomedis Lanjutan untuk iskemia miokard. Secara keseluruhan, 22 fitur diekstraksi, dan analisis diskriminan linier diimplementasikan dua kali melalui platform analisis Knime dan statistik R bahasa pemrograman untuk mengklasifikasikan pasien sebagai normal atau patologis. Yang pertama dari analisis ini hanya mencakup klasifikasi, sedangkan metode yang terakhir mencakup analisis komponen utama sebelum klasifikasi untuk membuat fitur baru. Akurasi klasifikasi yang diperoleh untuk metode ini adalah 84,5 dan 86,0 persen, masing-masing, dengan spesifisitas lebih dari 97 persen dan sensitivitas antara 62 dan 66 persen. Artikel ini menyajikan implementasi praktis dari teknik penambangan data tradisional yang dapat digunakan untuk membantu dokter dalam pengambilan keputusan; Selain itu, analisis komponen utama digunakan sebagai algoritma untuk pengurangan fitur (Ricciardi et al., 2020).

Penelitian dengan metode ensemble yang melibatkan analisis diskriminan linier dan regresi logistik dalam pengaturan online, tanpa perlu untuk menyimpan semua data yang diperoleh sebelumnya dilakukan oleh (Lalloué et al., 2022). Bootstrap Poisson dan stokastik proses aproksimasi digunakan dengan data standar online untuk menghindari ledakan numerik, konvergensi yang telah ditetapkan

secara teoritis. Konvergensi empiris dari ansambel online ini menghasilkan referensi Skor "batch" dipelajari pada lima kumpulan data berbeda dari mana aliran data disimulasikan, membandingkan enam proses yang berbeda untuk membangun skor online. Untuk setiap skor, 50 ulangan menggunakan total 10N pengamatan (N menjadi ukuran dataset) dilakukan untuk menilai konvergensi dan stabilitas metode, menghitung mean dan standar deviasi dari konvergensi kriteria. Sebuah studi pelengkap menggunakan pengamatan 100N juga dilakukan. Semua proses yang diuji pada semua set data konvergen setelah N iterasi, kecuali untuk satu proses pada satu dataset. Proses terbaik adalah proses rata-rata menggunakan data standar online dan ukuran langkah konstan sepotong demi sepotong.

Penelitian yang menganalisa untuk mereduksi variabel yang mempengaruhi fungsi ginjal pada tikus. Metode yang akan digunakan dalam penelitian ini adalah metode principal component analysis (PCA) atau analisa komponen utama. PCA merupakan salah satu metode dalam analisis multivariat yang secara khusus dikembangkan untuk mereduksi dimensi data yang ukurannya besar menjadi lebih sederhana tanpa harus kehilangan informasi data asli. Pada penelitian ini, metode PCA digunakan untuk mereduksi jumlah variabel, sehingga dari 8 variabel yang ada hanya akan diketahui 3 variabel yang benar-benar mempengaruhi perbaikan fungsi ginjal tikus, dimana 3 variabel yang dihasilkan tersebut dapat mewakili 8 variabel yang ada pada dataset. tikus, dimana 3 variabel yang dihasilkan tersebut dapat mewakili 8 variabel yang ada pada dataset. Variabel baru hasil reduksi akan dijadikan sebagai variabel input untuk membuat model persamaan regresinya untuk

melihat sebagaimana pengaruh variabel tersebut terhadap perbaikan fungsi ginjal tikus (Fitrianingsih & Sugiyarto, 2019).

Penelitian tentang prestasi belajar siswa dengan menggunakan metode analisis Principal Component Analysis (PCA). Penelitian dilakukan dengan mengumpulkan data melalui kuesioner kepada responden atau sampel penelitian yaitu siswa/i kelas X dan XI SMK Raksana 2 Medan, Sehingga diperoleh 3 faktor yaitu: faktor utama (PC1) memiliki nilai eigenvalue sebesar 3.11 dengan jumlah varians sebesar 31%. Faktor pendukung (PC2) memiliki nilai eigenvalue sebesar 1.50 dengan jumlah varians sebesar 15%. Faktor tambahan (PC3) memiliki nilai eigenvalue sebesar 1.16 dengan jumlah varians sebesar 12%. Keseluruhan faktor memberikan proporsi keragaman kumulatif sebesar 57.70%, artinya ketiga faktor tersebut menurut persepsi siswa/i yang menjadi responden dalam penelitian ini dapat mempengaruhi prestasi belajar siswa/i di SMK Raksana 2 Medan sebesar 57.70% dilakukan oleh (Nasution, 2019).

Pengolahan data dalam jumlah besar secara manual berpeluang menghasilkan banyak kesalahan. Untuk itu diperlukan pendekatan teknologi untuk dapat meminimalisir kesalahan yang dapat terjadi. Data mining merupakan suatu proses pengekstrakan informasi dari kumpulan data yang besar. Proses ini bertujuan untuk mendapatkan intisari dari kumpulan data tersebut. Proses data mining dapat menghasilkan menghasilkan informasi penting berupa klasifikasi (*classification)*, pengelompokan (*clustering*), bahkan prediksi (*prediction*). *Clustering* merupakan suatu proses analisis data untuk membentuk sekelompok objek berdasarkan sifat dan cirinya sehingga terbentuk suatu kelompok yang bersifat homogen antar

anggota pada kelompok yang sama. Namun, beberapa algoritma clustering menemui masalah ketika dihadapkan pada data dengan dimensi tinggi, termasuk juga KMeans. Reduksi dimensi dapat dijadikan sebagai salah satu langkap optimasi algoritma clustering. Proses reduksi dimensi yang umumnya diterapkan pada tahap pre-processing data bertujuan untuk mengurangi jumlah fitur (dimensi) tanpa menghilangkan informasi penting dari suatudata. Metode PCA akan membentuk sekumpulan dimensi baru yang kemudian di ranking berdasarkan varian datanya, schingga tercipta kumpulan data dengan fitur yang lebih sederhana. Penelitian ini akan menguji kinerja PCA sebagai salah satu metode optimasi algoritma clustering K-Means yang diterapkan pada data pertanian Kab. Bojonegoro pada tahun 2017 hingga 2020. Dataset hasil clustering yang didapatkan dari situs BPS akan dibandingkan dengan dataset dari sumber yang sama namun telah mengalami proses reduksi dimensi menjadi 1 PC, 2 PC, dan 3 PC. Evaluasi data hasil clustering menggunakan nilai DB Index menunjukkan nilai paling optimal pada dataset yang direduksi menjadi 1 PC dan dibentuk menjadi 3 klaster, yaitu 0.4072. sedangkan dengan jumlah klaster yang sama, dataset dengan 2PC menghasilkan nilai DB Index 0.6168, dataset dengan 3 PC menghasilkan nilai 0.6598, dan dataset tanpa proses reduksi dimensi menghasilkan nilai DB Index 0.4598. Penelitian ini dilakukan oleh (Hediyati & Suartana, 2021).

Sumber Daya Manusia (SDM) yang baik dapat menghasilkan kualitas dan kuantitas yang maksimal untuk mewujudkan bisnis yang berkelanjutan. Dengan demikian proses penerimaan karyawan menjadi penting, yang mana untuk memenuhi kebutuhan SDM yang berkualitas, sebuah perusahaan dapat membuka

lowongan pekerjaan dan menentukan seleksi SDM sesuai dengan kualifikasi pekerjaan. Penelitian ini memberikan kontribusi pada penentuan standar spesifikasi kerja di cafe atau coffee shop berdasarkan Big Data. Analisa dilakukan dengan metode Latent Dirichlet Allocation (LDA), dan metode Analytical Hierarchy Process (AHP) dalam proses pemilihan jenis pekerjaan dan kualifikasi dari perkerjaan yang tersedia di sektor food & beverages. Kuesioner dan wawancara dilakukan untuk pengumpulan data primer dan tiga website utama platform pencari kerja online (Jobstreet.com, Gawean.id, dan Lokerindonesia.com) digunakan sebagai referensi dan pengumpulan data. Hasil penelitian ini memberikan masukkan terkait 4 (empat) jenis pekerjaan di cafe yakni sebagai manager, barista, chef dan waiter yang mana terdapat 11 spesifikasi manager, 12 spesifikasi barista, 11 spesifikasi chef, dan 15 spesifikasi waiter (Natalia et al., 2021).

Latent Dirichlet Allocation (LDA) merupakan metode untuk pemodelan topik adalah yang didasarkan kepada konsep probabilitas untuk mencari kemiripan suatu dokumen dan mengelompokkan dokumen-dokumen menjadi beberapa topik atau kelompok. Metode ini masuk dalam unsupervised learning karena tidak ada label atau target pada data yang dianalisis. Penelitian ini bertujuan untuk mengelompokkan persepsi tentang pembelajaran online ke dalam beberapa topik menggunakan metode LDA. Data penelitian ini adalah data primer yang dikumpulkan melalui formulir online. Hasil analisis menunjukkan bahwa pemodelan LDA menggunakan 6 topik memiliki coherence score paling besar. Hasil visualisasi data text menggunakan wordcloud didapatkan kata tidak memiliki frekuensi kemunculan terbesar. Penentuan jumlah topik yang optimal berdasarkan

coherence score, didapatkan pemodelan LDA dengan 6 topik adalah yang paling optimal. secara garis besar terdapat beberapa kata yang saling beririsan dengan topik yang lain. Hasil pemodelan memberikan gambaran bahwa persepsi/pandangan mahasiswa terdapat pembelajaran online terkait pemahaman materi yang diberikan dosen, sinyal atau jaringan internet, kuota, dan tugas. Pada kata-kata terkait pemahaman materi, mahasiswa memberikan pandangan bahwa mereka tidak dapat memahami dengan baik materi yang diberikan oleh dosen (Fernanda, 2021).

2.2. Keaslian Penelitian

2.3.Tabel 2.1. Matriks literatur review PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTICT REGRETION*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Analisis Komponen Utama Faktor-Faktor Pendahulu (Antecendents) Berbagi Pengetahuan Pada Usaha Mikro, Kecil, Dan Menengah (Umkm) Di Indonesia	Anita Ilmaniati, Jurnal Teknologi, 2019	Manajemen pengetahuan bertujuan untuk meningkatkan keuntungan usaha melalui komunikasi dan meningkatkan penguasaan pengetahuan melalui berbagi pengetahuan	Faktor-faktor pendahulu berbagi pengetahuan pada UMKM di Indonesia dapat dikategorikan menjadi tiga komponen yaitu komponen 'suasana bekerja', 'intensi positif karyawan' dan 'pola pikir karyawan'. Hasil tersebut mereduksi komponen dari penelitian sebelumnya yang terdiri dari lima komponen. Variabel yang digunakan untuk mengukur komponen-komponen juga berkurang dari 19 variabel/item menjadi 14 variabel/item.	Tidak dijelaskan dengan detail tentang data pada penelitian	Pada penelitian untuk tesis yang dilakukan berbeda dalam penggunaan algoritma yaitu regresi logistic dan metodenya ditambah dengan LDA. Penelitian yang dirujuk ini menggunakan hanya menggunakan PCA.

Tabel 2.1. Matriks literatur review PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTICT REGRETION* (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
2	Application of the VNS heuristic for feature selection in credit scoring Problems Journal Machine Learning with Applications	Victor Gomes Helder, Tiago Pascoal Filomena, Luciano Ferreira, Guilherme Kirch Journal Machine Learning with Applications. 2022	Tujuan utama dari penelitian ini adalah untuk meningkatkan akurasi prediksi dalam model penilaian kredit dengan mengurangi dimensi dari ruang variabel. Untuk tujuan ini, teknik pemilihan variabel diusulkan berdasarkan konsep Variable Neighborhood Search (VNS).	Berdasarkan hasil yang diperoleh, VNS telah menunjukkan kinerja yang lebih baik dari PCA, secara umum, karena untuk kebanyakan kasus perbedaan yang signifikan dalam akurasi diperoleh, di samping AUC yang selalu lebih baik. Mengenai hasil tanpa teknik pemilihan fitur,	Tidak dijelaskan dengan detail tentang data pada penelitian	Penulis menggunakan penelitian sebagai dasar terhadap perbandingan penelitian berbeda dalam penggunaan algoritma dan metodenya. Pada penelitian rujukan menggunakan metode VNS dan PCA.
3	Pemodelan Persepsi Pembelajaran Online Menggunakan Latent Dirichlet Allocation	Jerhi Wahyu Fernanda, Statistika, 2021	Penelitian ini bertujuan untuk mengelompokkan persepsi tentang pembelajaran online ke dalam beberapa topik menggunakan metode LDA	Hasil pemodelan LDA, terdapat 6 topik yang dibentuk berdasarkan coherence score. Hasil analisis terhadap 6 topik tersebut, terdapat kata-kata yang saling beririsan	Saran : Kurangnya analisa perhitungan coherence score	Penulis menggunakan tambahan algoritma regresi logistic dan penambahan metode PCA. Penelitian rujukan menggunakan metode LDA.

Tabel 2.1. Matriks literatur review PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTICT REGRETION* (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
4	Complex Chemical Data Classification and Discrimination Using Locality Preserving Partial Least Squares Discriminant Analysis	Noor Atinah, Ahmad Muhammad Aminu ACS Omega 2020	Tujuan dari penelitian ini bukan untuk mengembangkan teknik diskriminasi yang baru melainkan untuk memperkenalkan LPPLS-DA untuk komunitas chemometrics dibandingkan PLS-DA dalam kimia	algoritma LPPLS-DA adalah efektif metode untuk pengurangan dimensi dan diskriminasi data kimia. Hasil eksperimen menunjukkan bahwa LPPLS DA adalah pilihan yang baik untuk klasifikasi praktis kompleks data kimia.	Saran : Membandingkan dengan algoritma yang lain	Penulis memiliki tujuan dan pemakaian algoritma yang berbeda untuk memproses hasil. Penelitian rujukan menggunakan algoritma LPPLS-DA.
5	Variable Selection in Count Data Regression Model based on Firefly Algorithm	Zakariya Yahya Algamal,Statistics, Optimization And Information Computing Vol. 7, June 2019	Penelitian untuk membuktikan efisiensi metode yang diusulkan dan mengungguli metode populer lainnya.	Penelitian untuk membuktikan efisiensi metode yang diusulkan dan mengungguli metode populer lainnya.	Tidak dijelaskan dengan detail tentang data pada penelitian.	Penulis menambahkan metode LDA dan PCA , sedangkan pada penelitian rujukan ini menggunakan algoritma regresi logistic dan algoritma firefy.

Tabel 2.1. Matriks literatur review PERBANDINGAN METODE REDUKSI DIMENSI PADA ALGORITMA *LOGISTICT REGRETION* (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6	Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Prestasi Belajar Siswa	Muhammad Zulfahmi, Jurnal Teknologi Informasi) Vol.3, No.1. Juni 2019	Penelitian ini bertujuan untuk menyederhanakan dan mereduksi faktor – faktor tersebut sehingga memperoleh faktor dominan yang mempengaruhi prestasi belajar siswa dengan menggunakan metode (PCA).	setelah dilakukan analisis faktor dengan metode principal component analysis, diperoleh 3 faktor yang mempengaruhi	Saran untuk menambahkan dataset pada pengujian	Penulis menambahkan algoritma regresi logistic
7	Feature Extraction and Dimensionality Reduction of Cancer Data Using Folded LDA	Samson Damilola Fabiyi, Dismas Ndubuisi Ezechukwu, 3rd International Informatics and Software Engineering Conference, IISEC, 2022 .	Penelitian ini untuk mengevaluasi efektifitas LDA sebagai teknik reduksi dimensi dalam klasifikasi data kanker ketika menggunakan data pelatihan yang kecil	Hasil percobaaan dengan menggunakan dataset kanker payudara dan kanker prostat didapat bahwa F-LDA dapat mengurangi dimensi secara efektif pada hasil klasifikasi sample kecil.	Hasil penelitian belum dibandingkan dengan data traing yang besar.	Pada penelitian yang penulis lakuakan menggunakan perbandingan algoritma LDA dan PCA pada algoritma regresi logistic.

2.3. Landasan Teori

2.3.1. *Regresi logistic*

Baik masalah regresi dan klasifikasi adalah contoh pembelajaran yang cdiawasi metode, di mana (Abdulhafedh, 2022) :

Masalah regresi digunakan untuk memprediksi nilai keluaran berdasarkan data sebelumnya pengamatan. Pada regresi, sering menggunakan variabel kuantitatif yang mengambil nilai numerik, seperti usia, tinggi badan, atau pendapatan seseorang, nilainya rumah, dan harga saham. Dengan kata lain, teknik Regresi memprediksi tanggapan terus menerus, misalnya, perubahan suhu atau fluktuasi permintaan daya.

• Masalah klasifikasi digunakan pada variabel output yang dapat dikategorikan, seperti ya atau tidak, lulus atau gagal. Dalam masalah klasifikasi, menggunakan variabel kualitatif yang mengambil nilai di salah satu kelas yang berbeda, atau kategori, seperti jenis kelamin seseorang (laki-laki atau perempuan), dan jenis produk dibeli (tipe A, B, atau C). Masalah klasifikasi bertindak mirip dengan clas sifier yang dapat memiliki dua atau lebih tingkat, dan tingkat mungkin atau mungkin tidak atau dinal. Dengan kata lain, teknik klasifikasi memprediksi tanggapan kategoris, misalnya, apakah email itu asli atau spam.

2.3.2. Data Berdimensi tinggi

Analisis data modern berurusan dengan sejumlah besar informasi. Berikut adalah beberapa contoh domain di mana mengumpulkan informasi dalam skala besar membuka jalur baru penelitian atau eksploitasi.(Azhari & Fitriani, 2022)

2.3.3. PCA (*Principal Component Analysis*)

PCA adalah metode statistik yang melakukan transformasi untuk menghasilkan kumpulan data yang tidak berkorelasi disebut sebagai komponen utama dari data terkait yang besar ditetapkan (Rana et al., 2020) Metode ini membantu mengurangi kompleksitas masalah dengan mengurangi dimensi dari kumpulan data yang besar. Kumpulan data yang tidak berkorelasi disebut set fitur membuat sistem bekerja lebih cepat dan akurat. PCA tidak lain adalah vektor Eigen yang terkait dengan nilai Eigen tertinggi matriks kovarians yang berasal dari kumpulan data standar.

Prosedur pengerjaan *Principal Component Analysis* bertujuan untuk menyederhanakan dan menghilangkan faktor atau indikator skrining yang kurang dominan dan kurang relevan tanpa mengurangi maksud dan tujuan dari data asli dari variabel acak x (matrik berukuran n x n, dimana baris-baris yang berisi observasi sebanyak n dari variabel acak x) adalah sebagai berikut: (Nasution, 2019)

i. Menghitung matrik varians kovarian dari data observasi.

$$Var\ x = \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (\dot{z}_{ij} - \mu_j)^2 \quad (1)$$

$$Con(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (2)$$

Dengan μx dan μy merupakan rata – rata (mean) sampel dari variabel x dan y, dimana xi dan yi merupakan nilai observasi ke i dari variabel x dan y. Dari data nilai yang digunakan,

ii. Mencari eigenvalues dan eigenvector dari matrik kovarian yang telah diperoleh yaitu: Nilai eigen dan vektor eigen untuk matriks kovarians dihitung. Nilai

eigen yang dikomputasi kemudian ditransformasikan (rotasi orthogonal varimax) menggunakan persamaan berikut :

$$Det\ (A - \lambda I) = 0 \quad (3)$$

Dimana :

$A = matrix\ nxn$

$\lambda = nilai\ eigenvalue$

$I = matrix\ identitas\ persegi$

iii. Mmmmm Menentukan nilai proporsi Principal Component (proporsi Principal Component (%) dengan persamaan :

$$PC\ \% = \frac{Nilai\ Eigen}{Varians\ Covarian}\ x\ 100\ \% \quad (4)$$

iv. Menghitung bobot factor (factor loading) berdasarkan eigenvector dengan persamaan :

$$Ax = \lambda x \quad (5)$$

Sehingga diperoleh kombinasi linier :

$a. \lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n, eigenvalue\ matriks\ A$

$b. x_1, x_2, x_3, \dots, x_n, eigenvector\ sesuai\ eigenvalue - nya\ \lambda n$

Persamaan eigenvalue & eigenvector merupakan Eigen Value Decomposition (EVD), dengan persamaan sebagai berikut:

$$X = XD \quad (6)$$

$$A = X\,D\,X^{-1} \quad (7)$$

Dimana

$A = matrix\ nxn\ yang\ memiliki\ n\ eigenvalue\ \lambda n$

$D = Eigenvalue\ dari\ eigenvector - nya$

$X = Eigenvector\ dari\ matrix\ A$

$X^{-1} = Invers\ dari\ eigenvector\ X$

2.3.4. LDA (*Linear Discriminant Analist*)

Linear Discriminant Analist r juga digunakan untuk pengurangan dimensi. LDA digunakan untuk menentukan kombinasi linier dari kumpulan fitur yang mencirikan dua atau lebih kelas. Ini adalah dimensi yang diawasi teknik reduksi yang akan digunakan dengan variabel bebas kontinu dan avariabel terikat kategoris. LDA bertujuan untuk memproyeksikan kumpulan data yang lebih besar ke ruang dimensi rendah melalui fitur diskriminan. Ronald A Fisher punya merumuskan metode diskriminan linier dan dia telah menunjukkan penggunaannya sebagai pengklasifikasi.

Metode LDA mengidentifikasi vektor proyeksi untuk memaksimalkan matriks pencar antar kelas sambil meminimalkan matriks pencar kelas dalam ruang fitur, tujuannya adalah untuk menemukan fungsi linier (Ricciardi et al., 2020)

$$y = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \cdots + a_q x_{iq2} \quad (8)$$

Dimana :

$$a^t = [\{ a_1, a_2, a_3, \dots, a_q \}] \quad (9)$$

Adalah vector koefisien yang harus ditentukan, sedangkan

$$x_i = [x_{i1}, x_{i2}, x_{i3}, \dots, x_{iq}] \quad (10)$$

Rotasi orthogonal varimax

$$Det\ (A - \lambda I) = 0 \quad (11)$$

Dimana :

$A = matrix\ nxn$

$\lambda = nilai\ eigenvalue$

$I = matrix\ identitas\ persegi$

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

3.1.1. Jenis Penelitian

Jenis penelitian yang digunakan oleh peneliti merupakan penelitian kuantitatif, dimana peneliti melakukan perhitungan matematis untuk menemukan hasil yang di inginkan.

3.1.2. Sifat Penelitian

Sifat dari penelitian yang akan dilakukan adalah ekspremental dimana peneliti melakukan sebuah eksperiment guna mendeteksi performa penerapan algoritma Regresi Logistic dalam perbandingan akurasi dan efisiensi antara metoda PCA dan metode LDA pada dataset berdimensi tinngi.

3.1.3. Pendekatan Penelitian

Pendekatan kuantitatif di gunakan oleh peneliti dimana penelitian akan melakukan penelitian sesuai alur yang telah peneliti buat.

3.2. Metode Pengumpulan Data

Dalam penelitian ini metode pengumpulan data untuk mendapatkan sumber data yang digunakan adalah metode pengumpulan data sekunder. Data penelitian ini diambil dari data publik Kaggle.

Pada penelitian ini, untuk mengetahui kinerja dari metode yang digunakan maka akan digunakan 2 set data. Adapun data yang digunakan terdiri dari satu set

data *Water Quality* dari *Kaggle Repository*, dan satu set data *Breast Cancer Dataset* berasal dari *Kaggle Repository*.

3.3. Metode Analisis Data

Pada penelitian ini, langkah pertama menentukan variabel independen dari dataset lalu menetapkannya ke 'X' dan menetapkan variabel dependen kemudian menerapkannya ke 'Y', selanjutnya menempatkan kolom berdasarkan indeksnya. Langkah selanjutnya membagi variabel dependen (y_train dan y_test) karena ingin memiliki nilai variabel dependen yang terdistribusi dengan baik dalam set pelatihan dan pengujian. Selanjutnya menerapkan penskalaan fitur pada dataset Teknik Penskalaan Fitur seperti normalisasi dan standarisasi digunakan untuk menormalkan rentang variabel independen atau fitur dari dataset untuk menghindari hasil yang miring (bias) saat menangani beberapa fitur yang mencakup berbagai tingkat besaran, rentang, dan satuan. Umumnya dapat memilih untuk menormalkan (normalisasi) ketika data terdistribusi normal, dan menskalakan (standardisasi) ketika data tidak berdistribusi normal. Metode penskalaan diuji dan kinerjanya dibandingkan untuk mendapatkan hasil terbaik karena beberapa algoritme pembelajaran mesin sangat sensitif terhadap fitur ini.

3.4. Alur Penelitian

Metode yang digunakan pada penelitian ini menggunakan metode *Principal Componant Analisist* (PCA) dan *Linear Diskriminat Analist* (LDA) sebelum diklasifikasikan dengan menggunakan algoritma regresi logistic.

Gambar 3.1. Diagram alur penelitian

Berdasarkan gambar 1, dijelaskan bahwa untuk membandingkan kinerja metode PCA dan LDA pada algoritma regresi logistic. Langkah – langkahnya sebagai berikut :

1. Pendekatan Penelitian (Validasi Penelitian, Studi Literatur, Identifikasi Masalah).

   Tahap awal penelitian yaitu melakukan validasi penelitian tentang permasalahan serta studi literatur untuk mengetahui masalah yang akan di analisa serta penggunaan algoritma yang tepat berdasarkan studi literatur sehingga dalam pemilihan algoritma tersebut didasarkan pada hasil ilmu yang dianalisa oleh penelitian sebelumnya.

2. Pengumpulan Data (Pemilihan Data Uji, Data Validasi)

   Pemilihan Dataset, data yang digunakan terdiri dari satu set data *Water Quality* Status dan satu set data *Breast Cancer* Dataset dari Kaggle Repository, kemudian data dipisahkan menjadi data uji dan data validasi.

3. Analisa Data

   Pada tahap ini menentukan variabel independen dari dataset lalu menetapkannya ke 'X' dan menetapkan variabel dependen kemudian menerapkannya ke 'Y', selanjutnya menempatkan kolom berdasarkan indeksnya. Langkah selanjutnya membagi variabel dependen (y_train dan y_test) karena ingin memiliki nilai variabel dependen yang terdistribusi dengan baik dalam set pelatihan dan pengujian. Selanjutnya menerapkan penskalaan fitur pada dataset Teknik Penskalaan Fitur seperti normalisasi dan standarisasi digunakan untuk menormalkan rentang variabel independen atau fitur dari

dataset untuk menghindari hasil yang miring (bias) saat menangani beberapa fitur yang mencakup berbagai tingkat besaran, rentang, dan satuan. Umumnya dapat memilih untuk menormalkan (normalisasi) ketika data terdistribusi normal, dan menskalakan (standardisasi) ketika data tidak berdistribusi normal. Metode penskalaan diuji dan kinerjanya dibandingkan untuk mendapatkan hasil terbaik karena beberapa algoritme pembelajaran mesin sangat sensitif terhadap fitur ini.

4. Algoritma

Menjalankan PCA (Saffari et al., 2020) adalah teknik ekstraksi ciri sehingga komponen-komponennya bukan merupakan salah satu dari variabel independen asli. Ini adalah yang baru, seperti semacam transformasi dari yang asli. Hanya dengan pemilihan fitur untuk mendapatkan variabel independen asli.. Variabel baru adalah arah yang paling banyak variansnya, yaitu arah penyebaran data paling banyak. Selanjutnya menganalisis data untuk menerapkan objek (mendapatkan nilai eigen dan vektor eigen dari matriks kovarians.) sehingga mendapatkan informasi yang diperlukan untuk menerapkan transformasi PCA. Yaitu, mengekstrak beberapa fitur teratas yang paling menjelaskan varians.

Pemilihan komponen utama menjelaskan varians terbanyak dalam dataset dapat dilihat dari jumlah varians yang dijelaskan oleh masing-masing komponen yang dipilih. Selanjutnya melatih model Regresi Logistik untuk menghitung dan mendapatkan bobot (koefisien) dari model Regresi Logistik dari dataset pelatihan tertentu yang terdiri dari X_train dan y_train sehingga

memiliki model regresi logistik yang sepenuhnya terlatih pada data pelatihan dan siap untuk memprediksi hasil baru t metode prediksi.

Menjalankan LDA (Rodrigues et al., 2020) untuk memodelkan perbedaan antara kelas dalam data. Dua variabel independen adalah variabel independen baru yang tidak ada di antara variabel independen asli yang merupakan variabel independen yang benar-benar baru yang diekstraksi melalui LDA. Selanjutnya melatih model Regresi Logistik.

5. Perbandingan

   Dalam membandingkan kinerja model PCA dan LDA penelitian ini mengunakan konfosius matrik sehingga dapat diketahui berapa persentase nilai akurasi masing-masing metode yang digunakan.

6. Dokumentasi Hasil Penelitian

   Hasil pada penelitian ini adalah mengetahui hasil performa dari metode PCA, metode LDA dan hasil dari perbandingan kedua metode yang diterapkan pada model regresi logistic.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Hasil penelitian menggunakan teknik pengurangan dimensi *linier discriminant analist* (LDA) dan *principal component analist* (PCA) untuk mengetahui kinerja dari algoritma regresi logistic (LR) pada penggunaan dataset yang berdimensi tinggi. Dataset yang digunakan adalah dataset yang memiliki atribut dan record yang banyak, dalam penelitian ini menggunakan dataset *breast cancer* dan dataset *water quality*. Hasil performa kinerja dari algoritma LR dapat dilihat dari nilai akurasi, presisi, recall, f1-score, dan waktu yang digunakan dalam menjalankan model / algoritma. Dalam penerapan teknik pengurangan dimensi yang digunakan pada penelitian ini akan dilihat dan dibandingkan hasil kinerjanya.

4.1. Dataset

Penelitian ini menggunakan platform *google colaboratory* dan bahasa pemrograman python. Dataset yang digunakan pada penelitian ini disesuaikan dengan aloritma klasifikasi yang digunakan yaitu algortima regresi logistic karna ketika LDA dan PCA digunakan pada algoritma logistik regresi, mereka berfungsi sebagai metode reduksi dimensi dan preprocessing data yang membantu dalam meningkatkan performa model regresi. Dataset yang digunakan merupakan dataset yang berdimensi tinggi sehingga memerlukan metode LDA dan PCA untuk mereduksi dimensi dataset tersebut. Dataset yang digunakan adalah dataset *breast cancer* dan *water quality* yang diambil dari *kaggle repository*. Dataset *breast cancer* terdiri dari 31 atribut dan 569 record, dataset *water quality* terdiri dari 21 atribut dan 7999 record. Untuk hasil dari dataset yang digunakan dapat dilihat pada gambar 4.1 berupa gambar sample dataset *breast cancer* dan 4.2 berupa gambar sample dataset *water quality*.

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

Gambar 4.1 Sample Dataset Breast Cancer

	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	fluoride	bacteria	...	lead	nitrates	nitrites	mercury	perchlorate
0	1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.20	...	0.054	16.08	1.13	0.007	37.75
1	2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.90	0.65	...	0.100	2.01	1.93	0.003	32.26
2	1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	...	0.078	14.16	1.11	0.006	50.28
3	1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	...	0.016	1.41	1.29	0.004	9.12
4	0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57	0.61	0.13	...	0.117	8.79	1.79	0.003	16.90

Gambar 4.2 Sampel Dataset Water Quality

4.2. Hasil Menentukan Nilai X dan Y

Tahap penelitian selanjutnya setelah data diperoleh adalah menentukan variabel X dan Y. Untuk menentukan variabel X dan Y, menggunakan pemrograman python seperti terlihat pada gambar 4.3 dan 4.4. Untuk rincian dari data atribut yang digunakan sebagai katagori X dan Y dapat dilihat pada tabel 4.1 dan 4.2.

```
X = data.drop('target', axis = 1)
y = data['target']
X.shape, y.shape
```

```
((569, 30), (569,))
```

Gambar 4.3. Nilai X dan Y Breast Cancer

```
# x and y vectors to train
X = data.drop('target', axis = 1)
y = data['target']
X.shape, y.shape
```

```
((7999, 20), (7999,))
```

Gambar 4.3. Nilai X dan Y Water Quality

Dataset *breast cancer* terdiri dari 30 atribut dan 569 record, terlihat pada gambar 4.2. Nilai Y yang digunakan adalah variabel target yang memiliki dua kelas yaitu 0 dan 1. Untuk angka 0 menunjukkan diagnosa kanker ganas dan angka 1 sebagai diagnosa kanker jinak. Variabel X adalah semua variabel selain dari variabel target. Dataset *water quality* terdapat 20 atribut dan 7999 record seperti terlihat pada gambar 4.3. Nilai Y terdiri dari 2 katagori yaitu 0 dan 1. Angka 0 untuk berbahaya (tidak aman), dan angka 1 untuk kondisi aman.

Perincian untuk variabel yang termasuk pada katagori dependen dan independen pada dataset yang digunakan dapat dilihat pada tabel 4.1 yang berisi sample rincian dataset *breast cancer* dan tabel 4.2 berisi sample rincian dataset *water quality*. Untuk menentukan faktor dependen dan independen digunakan dengan bantuan bahasa pemrograman python.

Tabel 4.1 Sample Rincian Dataset *Breast Cancer*

No	Atribut	Katagori
1	Target	Y (dependent)
2	radius_mean	X (Independent)
3	texture_mean	X (Independent)
4	perimeter_mean	X (Independent)
5	area_mean	X (Independent)
6	smoothness_mean	X (Independent)
7	compactness_mean	X (Independent)
8	concavity_mean	X (Independent)
9	concave_points_mean	X (Independent)
10	symmetry_mean	X (Independent)
11	fractal_dimension_mean	X (Independent)
12	radius_se	X (Independent)
13	texture_se	X (Independent)
14	perimeter_se	X (Independent)
15	area_se	X (Independent)

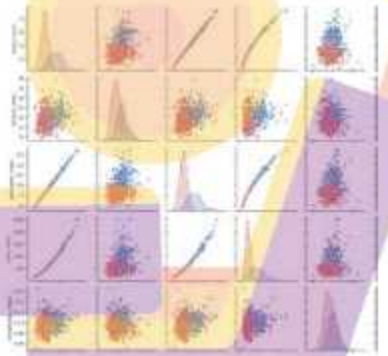
Tabel 4.2 Sampel Rincian Dataset *Water Quality*

No	Atribut	Katagori
1	Target	Y (dependent)
2	aluminium	X (Independent)
3	ammonia	X (Independent)
4	arsenic	X (Independent)
5	barium	X (Independent)

Tabel 4.2 Sampel Rincian Dataset *Water Quality* (Lanjutan)

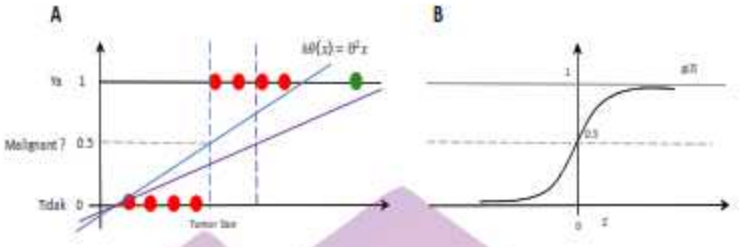
No	Atribut	Katagori
6	cadmium	X (Independent)
7	chloramine	X (Independent)
8	chromium	X (Independent)
9	copper	X (Independent)
10	flouride	X (Independent)
11	bacteria	X (Independent)
12	viruses	X (Independent)
13	lead	X (Independent)
14	nitrates	X (Independent)

Tampilan dataset *breast cancer* dapat dilihat dalam bentuk scatter plot dengan menggunakan kumpulan fungsi pasangan yang disediakan oleh seaborn pada python. Untuk menyederhanakan visualisasi, scatter plot yang dihasilkan hanya ditampilkan kombinasi dari lima pasang fitur pertama, mewakili nilai rata-rata dari parameter Radius, Texture, Perimeter, Area, dan Smoothness.



Gambar 4.5 Scatter Plot Dataset Breast Cancer

Setiap gambar mewakili plot pencar dari beberapa parameter. Visualisasi ini membuat identifikasi elemen penting untuk klasifikasi lebih mudah diakses. Beberapa dari pasangan ini, seperti Radius vs. Tekstur atau Perimeter vs. Kehalusan, memiliki tingkat pemisahan yang tepat pada target (titik biru = Jinak; titik merah = Ganas).



Gambar 4.6 Klasifikasi Regresi Logistik Breast Cancer

Plot pada Gambar 4.5A menjelaskan mengapa tidak dapat menerapkan metode linier pada klasifikasi biner. Dengan memplot sampel dengan hasil yang bisa jinak atau ganas (lingkaran merah). Dapat menerapkan model regresi linear untuk memisahkan sampel menjadi dua kelompok berbeda. Dalam upaya mengklasifikasikan nilai biner sebagai 0 dan 1, Regresi Linier mencoba untuk memprediksi nilai yang lebih besar dari 0,5 sebagai "1" dan semua yang kurang dari 0,5 sebagai "0" karena keluaran pengklasifikasi ambang batas untuk h0(x) adalah 0,5.

Secara teoritis, model Regresi Linier dapat bekerja dengan baik juga untuk klasifikasi dataset breast cancer seperti yang ditunjukkan oleh garis biru. Tapi jika kita memasukkan sampel lain dengan diagnosa Ganas (lingkaran hijau pada Gambar 4.5A). Model regresi linear mengadaptasi garis untuk memasukkan sampel baru (garis magenta).

Namun, model regresi logistik ini tidak dapat berfungsi dengan benar untuk semua sampel baru yang akan dipakai, karena klasifikasi bukan fungsi linier. Yang digunakan adalah Hipotesis baru h0(x) yang dapat menghitung probabilitas bahwa keluaran target bisa 0 atau 1: model baru ini adalah regresi Logistik.

$h\theta(x) = g(\theta^2 x)$

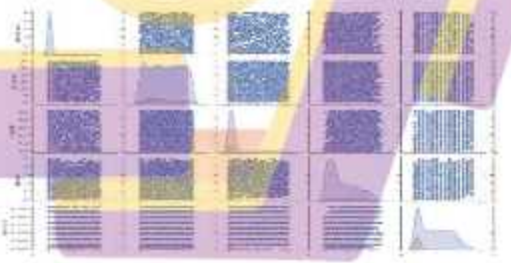
Model regresi logistik pada Persamaan 1, terlihat mirip dengan model Regresi Linier. Tetapi perbedaan sebenarnya ada pada fungsi g yang menggunakan produk dari vektor θ yang diterjemahkan dengan vektor x disebut z. Fungsi g didefinisikan seperti pada Persamaan 2:

$z = \theta^T x$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Fungsi g(z), yang merupakan fungsi sigmoid (Fungsi Logistik) adalah non-linear. Ini menghitung probabilitas bahwa keluaran Diagnosis bisa 0 atau 1 (Gambar 6B).

Tampilan dataset water quality dalam bentuk scatter plot , diwakilkan pada lima pasang fitur pertama yang terdiri dari nilai alumunium, ammonia,arsenic, barium, dan cadmium.



Gambar 4.7 Scater Plot Dataset Water Quality

Setiap gambar mewakili plot pencar dari beberapa parameter. Visualisasi ini membuat identifikasi elemen penting untuk klasifikasi lebih mudah diakses. Dari tampilan gambar scatter plot dataset pada water quality menunjukkan tingkat pemisahan yang tidak jelas pada target (titik biru = aman; titik merah = tidak aman).

4.3. Hasil Menghapus Fitur Berkorelasi

Pada dataset *breast cancer* hasil dari menghapus fitur yang berkorelasi adalah data menjadi 455 record dan 4 atribut untuk data training dan 114 record dan 4 atribut untuk data testing. Seperti terlihat pada gambar 4.8. Dataset *water quality* terdapat data sebanyak 6399 record dan 15 atribut terlihat pada gambar 4.9

```
X_train_uncorr = X_train_unique.drop(labels=corr_features, axis = 1)
X_test_uncorr = X_test_unique.drop(labels = corr_features, axis = 1)
X_train_uncorr.shape, X_test_uncorr.shape
```

```
((455, 4), (114, 4))
```

Gambar 4.8 Fitur Berkorelasi *Breast Cancer*

```
X_train_uncorr = X_train_unique.drop(labels=corr_features, axis = 1)
X_test_uncorr = X_test_unique.drop(labels = corr_features, axis = 1)
X_train_uncorr.shape, X_test_uncorr.shape
```

```
((6399, 15), (1600, 15))
```

Gambar 4.9 Fitur Berkorelasi *Water Quality*

4.4. Hasil Pengurangan Dimensi Menggunakan LDA

Data yang dihasilkan setelah menggunakan LDA pada dataset *breast cancer* adalah sebanyak 455 record dan 1 atribut untuk data training dan 114 record dan 1 atribut untuk data testing, seperti terlihat pada gambar 4.10. Sedangkan pada dataset *water quality* terdapat data 6399 record dan 1 atribut, seperti pada gambar 4.11.

```
X_train_lda.shape, X_test_lda.shape
```

```
((455, 1), (114, 1))
```

Gambar 4.10 Hasil Reduksi LDA *Breast Cancer*

```
X_train_lda.shape, X_test_lda.shape
```

```
((6399, 1), (1600, 1))
```

Gambar 4.11 Hasil Reduksi LDA *Water Quality*

4.5. Hasil Menjalankan *Time Running*

Dari hasil percobaan yang dilakukan menjalankan *time running* dataset *breast cancer* pada proses setelah LDA dan sebelum LDA tanpak pada gambar 4.12 dan gambar 4.13 Gambar 4.12 adalah hasil menggunakan LDA dan gambar 4.13 hasil tanpa menggunakan LDA.

```
%%time
run_LogisticRegression(X_train_lda, X_test_lda, y_train, y_test)

Accuracy on test set:
0.9122807017543859
CPU times: user 10.5 ms, sys: 0 ns, total: 10.5 ms
Wall time: 13 ms
```

Gambar 4.12 Waktu LDA *Breast Cancer*

Waktu yang digunakan untuk menjalankan model logistik regresi dengan menggunakan LDA sebesar 13 ms, terlihat pada gambar 4.8.

```
%%time
run_LogisticRegression(X_train, X_test, y_train, y_test)

Accuracy on test set:
0.9473684210526315
CPU times: user 54.5 ms, sys: 72.1 ms, total: 127 ms
Wall time: 95 ms
```

Gambar 4.13 Waktu Tanpa LDA *Breast Cancer*

Waktu yang digunakan untuk menjalankan model logistik regresi tanpa LDA adalah selama 95 ms seperti terlihat pada gambar 4.14. Waktu yang digunakan pada *water quality* dataset menggunakan algoritma LDA sebesar 25.3 ms dapat dilihat pada gambar 4.15 dan waktu yang digunakan tanpa menggunakan LDA sebesar 110 ms dapat dilihat pada gambar 4.15.

```
%%time
run_LogisticRegression(X_train_lda, X_test_lda, y_train, y_test)

Accuracy on test set:
0.895
CPU times: user 19.2 ms, sys: 0 ns, total: 19.2 ms
Wall time: 25.3 ms
```

Gambar 4.14 Waktu LDA *Water Quality*

```
%%time
run_LogisticRegression(X_train, X_test, y_train, y_test)

Accuracy on test set:
0.903125
CPU times: user 119 ms, sys: 78.1 ms, total: 197 ms
Wall time: 110 ms
```

Gambar 4.15 Waktu Tanpa LDA *Water Quality*

Waktu yang diperlukan untuk menjalankan model regresi logistic dengan PCA pada dataset *breast cancer* adalah 13.8 ms, dapat dilihat pada gambar 4. 16. Untuk waktu yang diperlukan tanpa menggunakan PCA adalah selama 54 ms, terlihat pada gambar 4.17. Pada dataset *water quality*, waktu yang diperlukan untuk menjalankan model regresi dengan PCA sebesar 20.6 ms dapat dilihat pada gambar 4.18. Untuk waktu yang dibutuhkan tanpa PCA sebesar 133 ms, terlihat pada gambar 4.19.

```
%%time
run_LogisticRegression(X_train_pca, X_test_pca, y_train, y_test)

Accuracy on test set:
0.8771929824561403
CPU times: user 12 ms, sys: 389 µs, total: 12.4 ms
Wall time: 13.8 ms
```

Gambar 4.16 Waktu PCA *Breast Cancer*

```
%%time
run_LogisticRegression(X_train, X_test, y_train, y_test)

Accuracy on test set:
0.9473684210526315
CPU times: user 47.4 ms, sys: 41.6 ms, total: 89 ms
Wall time: 54 ms
```

Gambar 4.17 Waktu Tanpa PCA *Breast Cancer*

```
%%time
run_LogisticRegression(X_train_pca, X_test_pca, y_train, y_test)

Accuracy on test set:
0.88625
CPU times: user 20.3 ms, sys: 9.18 ms, total: 29.5 ms
Wall time: 20.6 ms
```

Gambar 4.18 Waktu PCA *Water Quality*

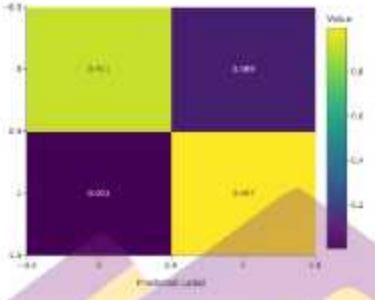
```
%%time
run_LogisticRegression(X_train, X_test, y_train, y_test)

Accuracy on test set:
0.903125
CPU times: user 138 ms, sys: 87.1 ms, total: 225 ms
Wall time: 133 ms
```

Gambar 4.19 Waktu Tanpa PCA *Water Quality*

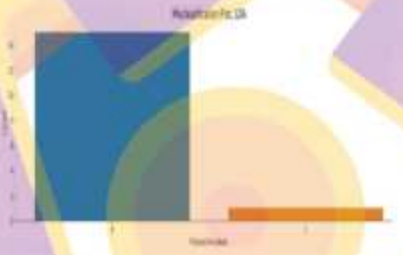
Dengan menggunakan teknik reduksi dimensi baik menggunakan LDA maupun menggunakan PCA sangat terlihat perbedaan yang besar dalam penggunaan waktu dalam menjalankan algoritma regresi logistic dibandingkan tanpa menggunakan teknik reduksi dimensi.

4.6. Hasil Prediksi dengan LDA

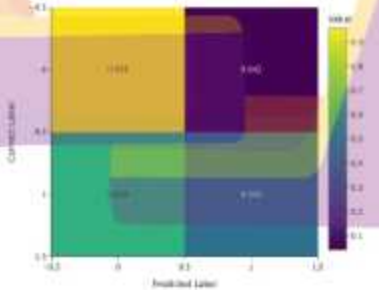
Menggunakan konfosius matrik, hasil prediksi dataset *breast cancer* tampak pada gambar 4.20. Nilai yang didapat true positive (TP) sebesar 0.911, true negative (TN) sebesar 0.003, false positive (FP) sebesar 0.089, dan false negatve (FN) sebesar 0.997. Kesalahan klasifikasi pada proses LDA terdapat 16 kasus dari 455 kasus yang ada, seperti terlihat pada gambar 4.21. Untuk dataset *water quality* diadapat nilai true positive (TP) sebesar 0.958, true negative (TN) sebesar 0.628, false positive (FP) sebesar 0.042, dan false negatve (FN) sebesar 0.372 dapat dilihat pada gambar 4.22. Kesalahan klasifikasi pada proses LDA terdapat 250 kesalahan dari 6399 kasus, seperti terlihat pada gambar 4.23.



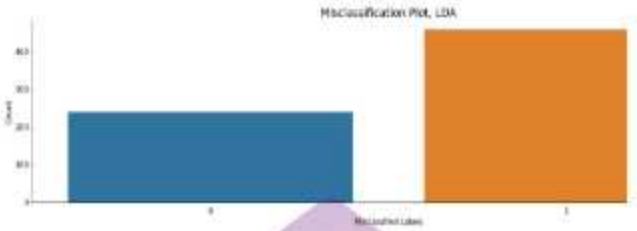
Gambar 4.20 Hasil Prediksi LDA *Breast Cancer*



Gambar 4.21 Nilai Kesalahan Klasifikasi LDA *Breast Cancer*



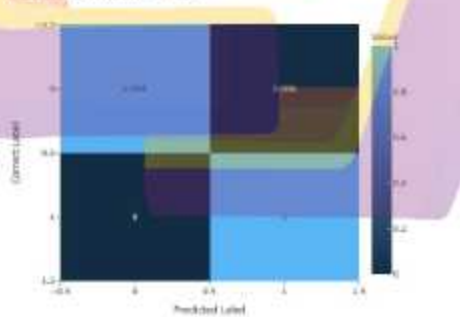
Gambar 4.22 Hasil Prediksi LDA *Water Quality*



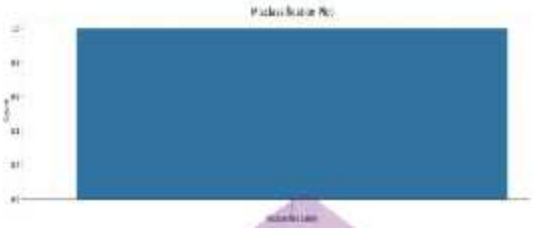
Gambar 4.23 Nilai Kesalahan Klasifikasi LDA *Water Quality*

4.7. Hasil Prediksi Menggunakan PCA

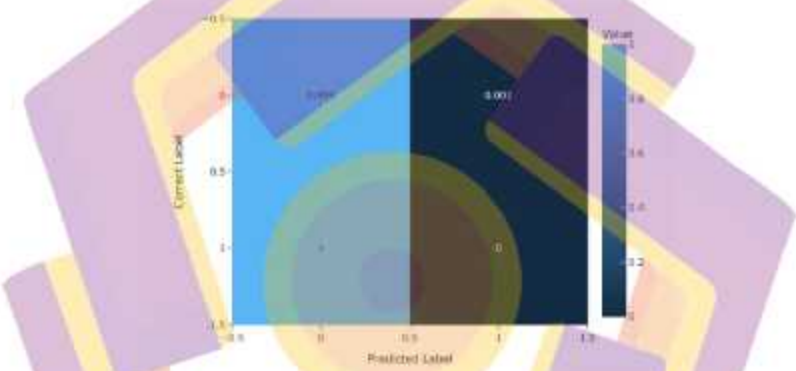
Hasil prediksi menggunakan PCA menggunakan dataset *breast cancer* dapat dilihat pada gambar 4.24, Nilai yang didapat true positive (TP) sebesar 0.994, true negative (TN) sebesar 0, false positive (FP) sebesar 0.006, dan false negatve (FN) sebesar 1. Kesalahan klasifikasi pada proses LDA terdapat 1 kasus dari 426 kasus yang ada, seperti terlihat pada gambar 4.25. Dengan menggunakan dataset *water quality* , Nilai yang didapat true positive (TP) sebesar 0.999, true negative (TN) sebesar 1, false positive (FP) sebesar 0.001, dan false negatve (FN) sebesar 0 seperti terlihat pada gambar 4.26. Kesalahan klasifikasi pada proses PCA terdapat 0 kasus dari 6399 kasus yang ada, seperti terlihat pada gambar 4.27.



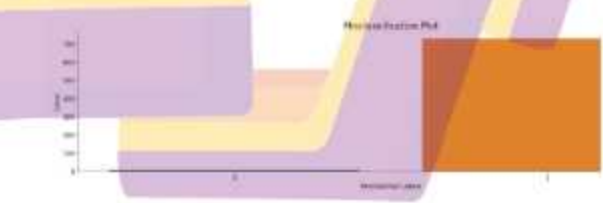
Gambar 4.24 Hasil Prediksi PCA *Breast Cancer*



Gambar 4.25 Nilai Kesalahan Klasifikasi PCA *Breast Cancer*



Gambar 4.26 Hasil Prediksi PCA *Water Quality*



Gambar 4.27 Nilai Kesalahan Klasifikasi PCA *Water Quality*

Error of rate yang dihasilkan pada dataset *breast cancer* sebesar 0.235% yang didapat dari nilai kesalahan klasifikasi 1 kasus dari total 426 total kasus. Pada dataset water quality mendapatkan error of rate sebesar 0% artnya tidak ada kasus kesalahan klasifikasi.

Tabel 4.3 Akurasi PCA *Breast Cancer*

Komponen	Nilai Akurasi (%)
1	84.2
2	87.7
3	89.4
4	91.2

Tabel 4.4 Akurasi PCA *Water Quality*

Komponen	Nilai Akurasi (%)
1	88.6
2	88.6
3	88.6
4	88.5

Akurasi yang didapat pada reduksi dimensi PCA pada dataset *breast cancer* menggunakan 5 komponen dapat dilihat pada tabel 4.3.Nilai akurasi tertinggi terdapat pada komponen 4. Pada dataset *water quality* nilai akurasi menggunakan PCA dapat dilihat pada tabel 4.4, dengan hasil akurasi sama pada setiap komponen.

4.8. Hasil Prediksi dengan Regresi Logistik

Performa dari model diukur dengan algoritma logistik dengan menggunakan konfosius matrik. Hasil yang didapat berupa nilai akurasi, presicion, recall, dan f1_score. Pada dataset *breast cancer* hasil dari evaluasi model regresi logistick setelah melalui proses reduksi dimensi dengan menggunakan LDA dapat dilihat pada gambar 4.28. Gambar 4.28 menunjukkan nilai akurasi sebesar 98%. Untuk nilai performa model dengan menggunakan PCA dapat dilihat pada gambar 4.29. Nilai akurasi yang didapat adalah 93%. Untuk hasil performa model regresi logistic tanpa reduksi dimensi dapat dilihat pada gambar 4.30, hasil akurasi yang didapat sebesar 96%.

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.99	0.95	0.97	169
1	0.97	0.99	0.98	286
accuracy			0.98	455
macro avg	0.98	0.97	0.98	455
weighted avg	0.98	0.98	0.98	455

Gambar 4.28 Performa LR dengan LDA *Breast Cancer*

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.92	0.90	0.91	170
1	0.94	0.95	0.95	285
accuracy			0.93	455
macro avg	0.93	0.93	0.93	455
weighted avg	0.93	0.93	0.93	455

Gambar 4.29 Performa LR dengan PCA *Breast Cancer*

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.95	0.94	0.94	170
1	0.96	0.97	0.97	285
accuracy			0.96	455
macro avg	0.95	0.95	0.95	455
weighted avg	0.96	0.96	0.96	455

Gambar 4.30 Performa LR Tanpa LDA dan PCA *Breast Cancer*

Hasil performa model regresi logistic dengan menggunkan dataset *water quality* dapat dilihat pada gambar 4.31, 4.32, dan 4.33. Pada gambar 4.31

menunjukkan hasil performa menggunaka LDA, hasil akurasi yang didapat sebesar 89 %, nilai akurasi yang didapat dengan menggunakan PCA sebesar 88% terlihat pada gambar 4.32. Hasil performa model regresi logistic tanpa reduksi dimensi LDA dan PCA sebesar 90%, dapat dilihat pada gambar 4.33.

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.91	0.98	0.94	5670
1	0.57	0.25	0.35	729
accuracy			0.89	6399
macro avg	0.74	0.61	0.65	6399
weighted avg	0.87	0.89	0.87	6399

Gambar 4.31 Performa LR dengan LDA *Water Quality*

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.89	1.00	0.94	5670
1	0.00	0.00	0.00	729
accuracy			0.88	6399
macro avg	0.44	0.50	0.47	6399
weighted avg	0.79	0.88	0.83	6399

Gambar 4.32 Performa LR dengan PCA *Water Quality*

```
print(classification_report(y_train, LRtrain_preds))
```

	precision	recall	f1-score	support
0	0.91	0.98	0.95	5670
1	0.66	0.28	0.39	729
accuracy			0.90	6399
macro avg	0.79	0.63	0.67	6399
weighted avg	0.88	0.90	0.88	6399

Gambar 4.33 Performa LR Tanpa LDA dan PCA *Water Quality*

4.9. Pembahasan

Hasil pengujian dari perbandingan kinerja regresi logistic menggunakan teknik reduksi dimensi LDA dan PCA berupa nilai akurasi, presisi, recall, dan f1 score dapat dilihat pada tabel 4.5 untuk dataset *breast cancer* dan 4.6 untuk dataset *water quality*.

Tabel 4.5 Hasil Perbandingan Kinerja LR *Breast Cancer*

	LDA + LR	PCA + LR	LR
Akurasi	98	96	94
Presisi	99	98	95
Recall	99	95	95
F1 score	99	97	95

Tabel 4.6 Hasil Perbandingan Kinerja LR *Water Quality*

	LDA + LR	PCA + LR	LR
Akurasi	89	91	91
Presisi	67	71	71
Recall	29	35	33
F1 score	41	47	45

Selain dalam bentuk tabel untuk hasil performa perbandingan dari metoda LDA dan PCA pada algoritma regresi logistic dibuat juga dalam bentuk diagram batang seperti pada gambar 4.34. Gambar 4.33 merupakan tampilan hasil yang didapat model logistik regresi pada dataset *breast cancer* menggunakan LDA, PCA, dan LR tanpa reduksi dimensi. Hasil yang didapat berupa nilai akurasi , precision, recall, dan f1-score. Perbandingan hasil performa dari model regresi menggunakan dataset *water quality* dengan menggunakan LDA, PCA, dan LR tanpa reduksi dimensi dapat dilihat pada gambar 4.35.

Gambar 4.34 Perbandingan Performa LR *Breast Cancer*

Gambar 4.35 Perbandingan Performa LR *Water Quality*

Gambar 4.36 Performa Waktu Tempuh LR *Breast Cancer*

Gambar 4.36, dan 4.37 menampilkan perbandingan waktu tempuh untuk menjalankan model logistic regresi. Gambar 4.36 hasil dari percobaan pada dataset *breast cancer* dan 4.37 merupakan hasil dari dataset *water quality*. Hasil yang didapat berupa waktu yang digunakan ketika menggunakan reduksi dimensi LDA, PCA, LR tanpa LDA, dan LR tanpa PCA. Hasil detail dari performa waktu pada pemodelan logistic regresi (LR) dapat juga dilihat pada tabel 4.7 yang berisi hasil performa waktu untuk dataset *breast cancer* dan tabel 4.8 berisi hasil performa waktu untuk dataset *water quality*.

Tabel 4.7 Hasil Waktu *Breast Cancer*

	LDA	LR-LDA	PCA	LR-PCA
CPU time (ms)	16.8	83.7	12.4	89
Wall time (ms)	20.8	54.6	13.8	54

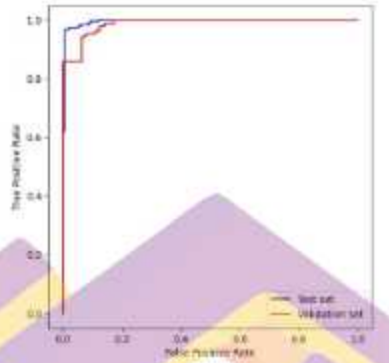
Tabel 4.8 Hasil Waktu *Water Quality*

	LDA	LR-LDA	PCA	LR-PCA
CPU time (ms)	19.2	197	29.5	225
Wall time (ms)	25.3	110	20.6	133

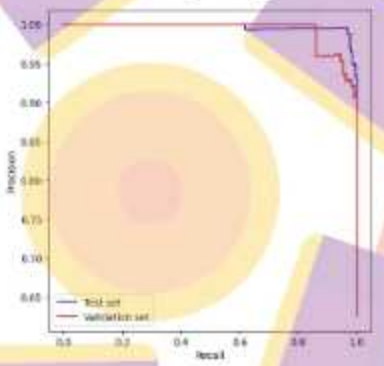
Gambar 4.37 Performa Waktu Tempuh LR *Water Quality*

Dari hasil penelitian berupa nilai akurasi, presisi, recall, f1 score dan waktu yang digunakan, nilai yang paling signifikan berbeda adalah nilai waktu yang digunakan dalam pemrosesan algoritma. Waktu yang digunakan dalam menjalankan algoritma regresi logistic dengan menggunakan metode LDA dan PCA adalah lebih sedikit dibandingkan dengan waktu tanpa menggunakan metode reduksi dimensi LDA dan PCA.

Hasil penelitian dapat dilihat dari hasil kurva ROC untuk melihat perubahan hasil batas klasifikasi. Hasil kurva ROC tidak tergantung pada keseimbangan data. Kurva presisi recall mirip dengan kurva ROC, yang memberikan gambaran yang lebih baik pada hasil klasifikasi. Pada gambar 4.38 merupakan kurva ROC pada regresi logistic dataset breast cancer yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 0.999 dan untuk validasi sebesar 0.974. Pada gambar 4.39 kurva presisi recall pada regresi logistic dengan dataset breast cancer . Hasil yang ditunjukkan oleh kurva presisi sebesar 0.999 pada training set dan 0.97 pada validasi set. Nilai ROC AUC sama dengan 1 menunjukkan hasil klasifikasi yang sempurna.

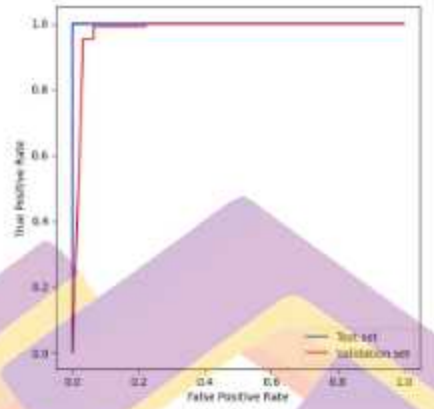


Gambar 4.38 Kurva ROC Regresi Logistik Breast Cancer

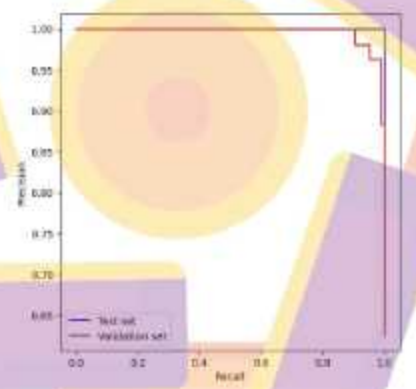


Gambar 4.39 Kurva Presisi Recall Regresi Logistik Breast Cancer

Pada gambar 4.40 merupakan tampilan kurva ROC hasil PCA pada dataset breast cancer yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 1 dan untuk validasi sebesar 0.987. Gambar 4.41 merupakan tampilan kurva Presisi recall PCA dataset breast cancer yang memiliki nilai untuk training set sebesar 1 dan pada validasi set sebesar 0.983.

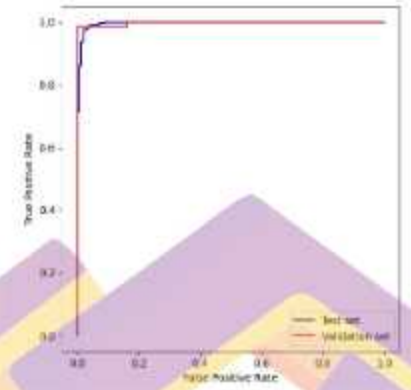


Gambar 4.40 Kurva ROC PCA Breast Cancer

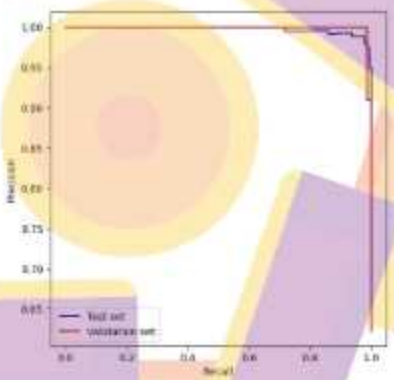


Gambar 4.41 Kurva Presisi Recall PCA Breast Cancer

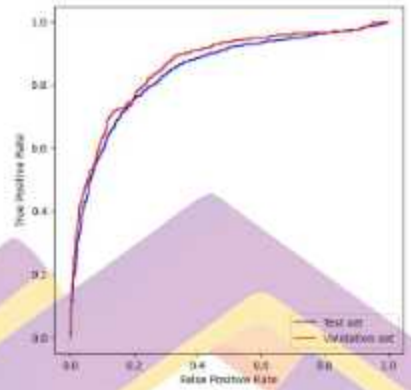
Pada gambar 4.42 merupakan tampilan kurva ROC hasil LDA pada dataset breast cancer yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 0.996 dan untuk validasi sebesar 0.998. Gambar 4.43 merupakan tampilan kurva Presisi recall LDA dataset breast cancer yang memiliki nilai untuk training set sebesar 0.998 dan pada validasi set sebesar 0.999.



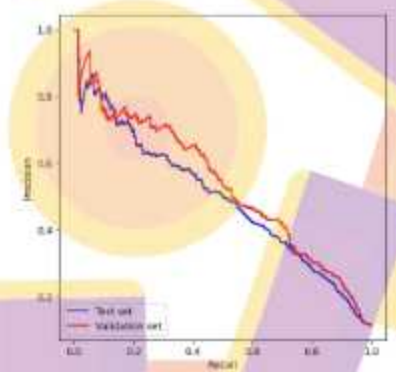
Gambar 4.42 Kurva ROC LDA Breast Cancer



Gambar 4.43 Kurva Presisi Recall LDA Breast Cancer

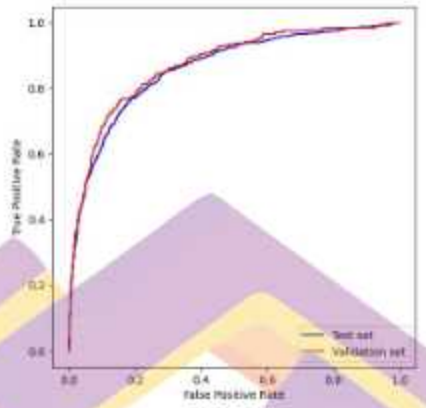


Gambar 4.44 Kurva ROC Regresi Logistik Water Quality

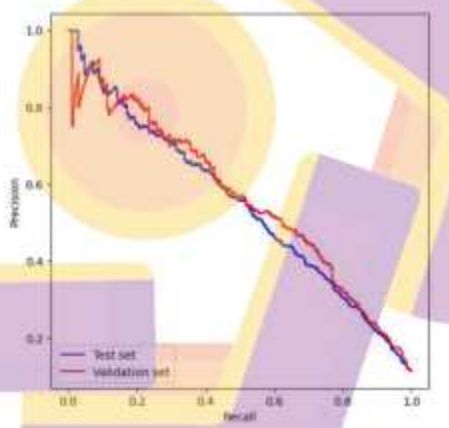


Gambar 4.45 Kurva Presisi Recall Regresi Logistik Water Quality

Pada gambar 4.44 merupakan kurva ROC pada regresi logistic dataset water quality yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 0.843 dan untuk validasi sebesar 0.862. Pada gambar 4.45 kurva presisi recall pada regresi logistic dengan dataset water quality . Hasil yang ditunjukkan oleh kurva presisi sebesar 0.495 pada training set dan 0.536 pada validasi set. Nilai ROC AUC sama dengan 1 menunjukkan hasil klasifikasi yang sempurna.

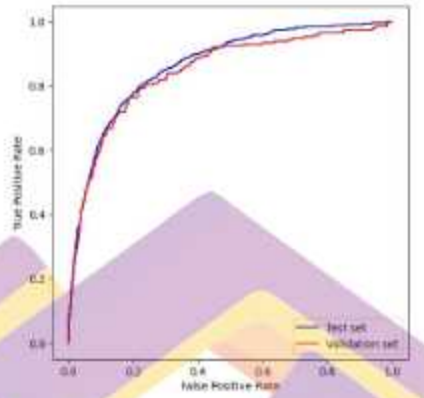


Gambar 4.46 Kurva ROC PCA Water Quality

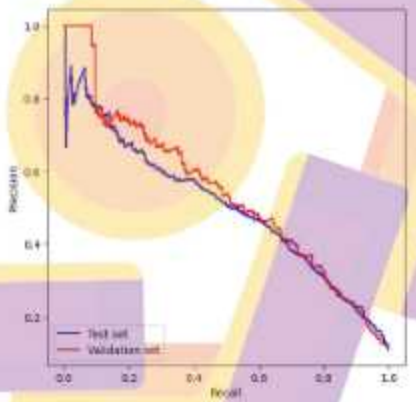


Gambar 4.47 Kurva Presisi Recall PCA Water Quality

Pada gambar 4.46 merupakan tampilan kurva ROC hasil PCA pada dataset water quality yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 0.861 dan untuk validasi sebesar 0.872. Gambar 4.47 merupakan tampilan kurva Presisi recall PCA dataset water quality yang memiliki nilai untuk training set sebesar 0.547 dan pada validasi set sebesar 0.564.



Gambar 4.48 Kurva ROC LDA Water Quality



Gambar 4.49 Kurva Presisi Recall LDA Water Quality

Pada gambar 4.48 merupakan tampilan kurva ROC hasil LDA pada dataset water quality yang memiliki nilai daerah dibawah kurva ROC untuk training set sebesar 0.864 dan untuk validasi sebesar 0.846. Gambar 4.49 merupakan tampilan kurva Presisi recall LDA dataset breast cancer yang memiliki nilai untuk training set sebesar 0.502 dan pada validasi set sebesar 0.54.

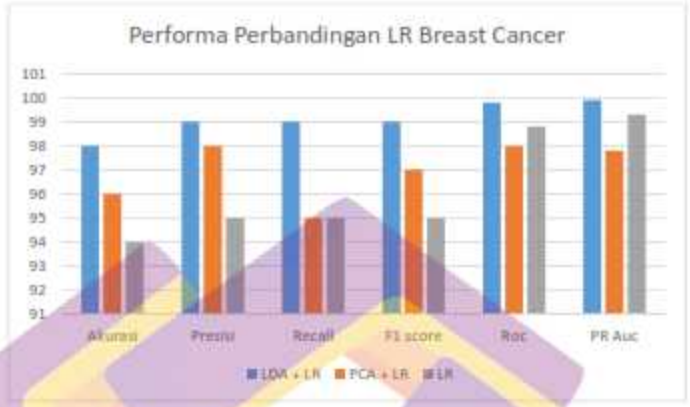
Setelah melakukan penelitian mencari nilai akurasi, presisi, recall, f1 score, roc, dan pr auc dengan menggunakan 2 jenis dataset yaitu breast cancer dan water quality maka metode yang terbaik adalah dengan menggunakan reduksi dimensi linear discriminant analist (LDA). Hasil dari penelitian dapat dilihat pada tabel 4.9 dan 4.50 yang menunjukkan hasil dari dataset breast cancer , dan hasil dari penelitian menggunakan dataset water quality dapat dilihat pada tabel 4.10 dan gambar 4.51.

Tabel 4.9 Hasil Perbandingan Kinerja Keseluruhan LR *Breast Cancer*

	LDA + LR	PCA + LR	LR
Akurasi	98	96	94
Presisi	99	98	95
Recall	99	95	95
F1 score	99	97	95
Roc	99.8	98	98.8
PR Auc	99.9	97.8	99.3

Tabel 4.10 Hasil Perbandingan Kinerja Keseluruhan LR *Water Quality*

	LDA + LR	PCA + LR	LR
Akurasi	89	91	91
Presisi	67	71	71
Recall	29	35	33
F1 score	41	47	45
Roc	84.6	87.2	86.2
PR Auc	54	56.4	53.6



Gambar 4.50 Hasil Perbandingan Keseluruhan Dataset Breast Cancer

Gambar 4.51 Hasil Perbandingan Keseluruhan Dataset Water Quality

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian pada 2 dataset berdimensi tinggi yang digunakan pada perbandingan metode reduksi dimensi LDA dan PCA pada algoritma regresi logistic maka dapat ditarik kesimpulan :

1. PCA terbukti sebagai metode reduksi dimensi yang lebih baik dengan dataset yang memiliki atribut yang lebih banyak atau berdimensi tinggi. Hal ini dapat dilihat dari hasil pengujian penelitian redukssi dengan PCA, atribut yang direduksi pada dataset *breast cancer* menjadi 2 atribut dari 30 atribut awal dan 426 record dari 569 record awal serta menghasilkan *error of rate* sebesar 0.235% . Pada dataset *water quality* atribut yang ddireduksi menjadi 2 atribut dari 21 atribut, dan 6399 record dari 7999 record awal, serta *error of rate* yang dihasilkan sebesar 0%.
2. Penerapan metode reduksi dimensi yang optimal dengan performa yang terbaik adalah dengan menggunakan LDA, hal ini terbukti dari hasil performa kinerja yang dihasilkan. Pada dataset *breast cancer*, akurasi sebesar 98%, presisi sebesar 99%, recall sebesar 95%, f1 score sebesar 97% dan waktu yang diperlukan sebanyak 20.8 ms. Pada dataset *water quality* , akurasi sebesar 89%, presisi sebesar 91%, recall sebesar 98%, f1 score sebesar 94% dan waktu yang diperlukan sebanyak 25.3 ms.

3. Akurasi yang didapatkan dari satu algoritma dengan algoritma lainnya bisa berbeda. Perbedaan ini dikarenakan jumlah data yang dipakai, tipe data yang dipakai dan keterkaitan antar data. Jumlah data mempengaruhi karena semakin banyak data akan semakin baik akurasinya, tipe data akan berpengaruh terhadap algoritma yang dipakai dan akurasi algoritmanya, karena algoritma terkadang hanya dapat menggunakan satu tipe data.

5.2. Saran

Untuk pengembangan penelitian tentang metode klasifikasi regresi logistic dalam permasalahan data yang berdimensi tinggi menggunakan teknik reduksi dimensi yang ada, maka penulis menyarankan untuk penelitian selanjutnya:

1. Pengujian untuk metode klasifikasi dengan teknik reduksi dimensi tidak hanya menggunakan dataset public, sebaiknya ditambahkan dataset yang sebenarnya ada di lapangan.
2. Penelitian selanjutnya menambahkan metode klasifikasi yang lain selain algoritma logistic regresi.
3. Penelitian berikutnya juga bisa menambahkan teknik reduksi dimensi yang lain seperti Singular Value Decomposition sebagia perbandingan dalam model klasifikasi.

DAFTAR PUSTAKA

PUSTAKA BUKU

Begg, M. D. (2009). An introduction to categorical data analysis (2nd edn). Alan Agresti, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007. No. of Pages: 400. Price: $100.95. ISBN: 978-0-471-22618-5. In *Statistics in Medicine* (Vol. 28, Issue 11). https://doi.org/10.1002/sim.3564

Hosmer andLemeshow. (2013). *Epdf.Pub_Applied-Logistic-Regression-Wiley-Series-in-Probab.Pdf*.

Kantardzic, M. (2019). *Data Mining Concepts, Models, Methods, and Algorithms* (Third Edit). John Wiley & Son Inc.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Abdulhafedh, A. (2022). Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. *OALib*, *09*(02), 1–19. https://doi.org/10.4236/oalib.1108414

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Alharthi, A. M., Lee, M. H., & Algamal, Z. Y. (2022). Improving Penalized Logistic Regression Model with Missing Values in High-Dimensional Data. *International Journal of Online and Biomedical Engineering*, *18*(2), 40–54. https://doi.org/10.3991/ijoe.v18i02.25047

Azhari, M. F., & Fitriani, F. A. (2022). Coronary Heart Disease Risk Prediction Using Binary Logistic Regression Based on Principal Component Analysis. *Enthusiastic : International Journal of Applied Statistics and Data Science*, *2*(1), 47–55. https://doi.org/10.20885/enthusiastic.vol2.iss1.art6

Begg, M. D. (2009). An introduction to categorical data analysis (2nd edn). Alan Agresti, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007. No. of Pages: 400. Price: $100.95. ISBN: 978-0-471-22618-5. In *Statistics in Medicine* (Vol. 28, Issue 11). https://doi.org/10.1002/sim.3564

Bielza, C., Robles, V., & Larrañaga, P. (2011). Regularized logistic regression

without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, *38*(5), 5110–5118. https://doi.org/10.1016/j.eswa.2010.09.140

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., Dwi, A., Putra, P., & Artikel, G. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(2), 2828–9099. https://doi.org/10.55123/jomlai.v1i2.598

Dewanta, F., & Abdillah, M. (n.d.). *Klasifikasi Beban Listrik dengan Machine Learning Menggunakan Metode K- Nearest Neighbor*. *5*(2), 163–172.

Fernanda, J. W. (2021). Pemodelan Persepsi Pembelajaran Online Menggunakan Latent Dirichlet Allocation. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *9*(2), 79. https://doi.org/10.26714/jsunimus.9.2.2021.79-85

Fitrianingsih, F., & Sugiyarto, S. (2019). Implementasi Analisa Komponen Utama untuk Mereduksi Variabel yang Mempengaruhi Perbaikan pada Fungsi Ginjal Tikus. *Jurnal Ilmiah Matematika*, *6*(2), 62. https://doi.org/10.26555/konvergensi.v6i2.19549

Fujiwara, T., Wei, X., Zhao, J., & Ma, K. L. (2022). Interactive Dimensionality Reduction for Comparative Analysis. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 758–768. https://doi.org/10.1109/TVCG.2021.3114807

Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *Journal of Information Engineering and Educational Technology*, *5*(2), 49–54. https://doi.org/10.26740/jieet.v5n2.p49-54

Hosmer andLemeshow. (2013). *Epdf.Pub_Applied-Logistic-Regression-Wiley-Series-in-Probab.Pdf*.

K.Raju, Rao, Y. S., & Yadav, M. N. (2015). Performance Analysis of PCA and LDA. *International Journal of Innovative Research in Electronics and Communications* , *2*(2), 17–22. www.arcjournals.org

Kaya, S., & Yaganoglu, M. (2020). An Example of Performance Comparison of Supervised Machine Learning Algorithms before and after PCA and LDA Application: Breast Cancer Detection. *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020*, 0–5. https://doi.org/10.1109/ASYU50717.2020.9259883

Lalloué, B., Monnez, J.-M., & Albuisson, E. (2022). Construction and Update of an

Online Ensemble Score Involving Linear Discriminant Analysis and Logistic Regression. *Applied Mathematics*, *13*(02), 228–242. https://doi.org/10.4236/am.2022.132018

Liang, S., Singh, M., Dharmaraj, S., & Gam, L. H. (2010). The PCA and LDA analysis on the differential expression of proteins in breast cancer. *Disease Markers*, *29*(5), 231–242. https://doi.org/10.3233/DMA-2010-0753

Lim, N., Ahn, H., Moon, H., & Chen, J. J. (2010). Classification of high-dimensional data with ensemble of logistic regression models. *Journal of Biopharmaceutical Statistics*, *20*(1), 160–171. https://doi.org/10.1080/10543400903280639

Lisnawati, L., & Syafril, A. S. (2021). Pengaruh Likuiditas, Profitabilitas Dan Solvabilitas Terhadap Opini Audit Going Concern (Studi Pada Perusahaan Retail Trade Yang Terdaftar Di Bursa Efek Indonesia). *Land Journal*, *2*(2), 1–14. https://doi.org/10.47491/landjournal.v2i2.1274

Nasution, M. Z. (2019). PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) DALAM PENENTUAN FAKTOR DOMINAN YANG MEMPENGARUHI PRESTASI BELAJAR SISWA (Studi Kasus : SMK Raksana 2 Medan). *Jurnal Teknologi Informasi*, *3*(1), 41. https://doi.org/10.36294/jurti.v3i1.686

Natalia, C., Suprata, F., Surbakti, F. P. S., & Clarence, S. (2021). Penentuan Standar Spesifikasi Kerja di Café Berdasarkan Big Data dengan Metode LDA dan AHP. *Jurnal Rekayasa Sistem Industri*, *10*(2), 211–226. https://doi.org/10.26593/jrsi.v10i2.5228.211-226

P. Vatcheva, K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Rana, D., Jena, S. P., & Pradhan, S. K. (2020). Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(9), 2353–2360.

Rashid, N. A., Hussain, W. S. E. C., Ahmad, A. R., & Abdullah, F. N. (2019). Performance of classification analysis: A comparative study between PLS-DA and integrating PCA+LDA. *Mathematics and Statistics*, *7*(4), 24–28. https://doi.org/10.13189/ms.2019.070704

Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., Picone, I., Santini, S., & Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Informatics Journal*, *26*(3), 2181–2192.

https://doi.org/10.1177/1460458219899210

Rich, D. C., Livingston, K. M., & Morgan, S. L. (2020). Evaluating performance of Lasso relative to PCA and LDA to classify dyes on fibers. *Forensic Chemistry*, *18*(January), 100213. https://doi.org/10.1016/j.forc.2020.100213

Rodrigues, L., Parayil, A., Shetty, T., & Mirza, I. (2020). Use of Linear Discriminant Analysis (LDA), K Nearest Neighbours (KNN), Decision Tree (CART), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector Machines (SVM) to Predict Admission for Post Graduation Courses. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3683065

Saffari, E., Yildirimoglu, M., & Hickman, M. (2020). A methodology for identifying critical links and estimating macroscopic fundamental diagram in large-scale urban networks. *Transportation Research Part C: Emerging Technologies*, *119*(March), 102743. https://doi.org/10.1016/j.trc.2020.102743

Şahin, D. Ö., Kural, O. E., Akleylek, S., & Kılıç, E. (2021). Permission-based Android malware analysis by using dimension reduction with PCA and LDA. *Journal of Information Security and Applications*, *63*(October), 102995. https://doi.org/10.1016/j.jisa.2021.102995

Sihombing, P. R., & Yuliati, I. F. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(2), 417–426. https://doi.org/10.30812/matrik.v20i2.1174

Tsoulfidis, L., & Athanasiadis, I. (2022). A new method of identifying key industries: a principal component analysis. *Journal of Economic Structures*, *11*(1). https://doi.org/10.1186/s40008-022-00261-z

Yunial, A. H. (2020). Analisa Perbandingan Algoritma Klasifikasi Support Vector Machine, Decession Tree Dan Naive Bayes. *Prosiding Seminar Nasional Informatika Dan Sistem Informasi* , *5*(2).

PUSTAKA LAPORAN PENELITIAN

Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., Picone, I., Santini, S., & Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health Informatics Journal*, *26*(3), 2181–2192. https://doi.org/10.1177/1460458219899210