



Penerapan Metode Regresi Logistik Untuk Memprediksi Peristiwa Biner Pasien Pasca Operasi Kanker Payudara

Sylvia Sujana, Ayu Ratna Juwita*, Rahmat, Sutan Faisal

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia

Email: ¹if20.sylviasujana@mhs.ubpkarawang.ac.id, ²*ayurj@ubpkarawang.ac.id, ³rahmat@ubpkarawang.ac.id,
⁴sutan.faisal@ubpkarawang.id

Email Penulis Korespondensi: ayurj@ubpkarawang.ac.id

Submitted: 08/07/2024; Accepted: 23/07/2024; Published: 26/07/2024

Abstrak—Kanker payudara merupakan penyebab kematian kedua tertinggi pada perempuan di seluruh dunia. Untuk mengatasi masalah yang terus berkembang ini, penelitian ini dirancang sebuah model yang dapat memprediksi kanker payudara dengan memanfaatkan dataset kemudian diolah menggunakan metode Prediksi Regresi Logistik. Metode ini tepat untuk memprediksi data yang digunakan karena kemampuannya dalam menangani variabel dependen yang bersifat kategorikal serta memberikan output dalam bentuk probabilitas. Penelitian ini menggunakan dataset sebanyak 306 sample dengan 4 atribut. Data yang digunakan Langkah penelitian meliputi pengumpulan data, preprocessing, pemodelan dengan regresi logistik dan evaluasi hasil menggunakan matrix seperti confusion matrix, MAE, MSE, dan R-Square. Hasil menunjukkan akurasi prediksi sebesar 86%, dengan nilai MSE 0,137 dan R-Square 0,309. Penelitian ini menunjukkan efektivitas regresi logistic dalam memprediks kelangsungan hidup pasien pasca operasi kanker payudara. Meskipun demikian dengan menerapkan algoritma yang berbeda penelitian ini dapat memilih set atribut terbaik yang signifikan guna meningkatkan nilai akurasi prediksi pada pasien pasca operasi kanker payudara.

Kata Kunci: Kanker Payudara; Regresi Logistik; Prediksi Kelangsungan Hidup; Data Mining; Evaluasi Model

Abstract—Breast cancer is the second leading cause of death in women worldwide. To overcome this growing problem, this study designed a model that can predict breast cancer by utilizing datasets and then processed using the Logistic Regression Prediction method. This method is appropriate for predicting the data used because of its ability to handle dependent variables that are categorical and provide outputs in the form of probabilities. This study uses a dataset of 306 samples with 4 attributes. Data used Research steps include data collection, preprocessing, modeling with logistic regression and evaluating results using matrices such as confusion matrix, MAE, MSE, and R-Square. The results showed a prediction accuracy of 86%, with an MSE value of 0.137 and R-Square of 0.309. This study shows the effectiveness of logistic regression in predicting the survival of patients after breast cancer surgery. However, by applying different algorithms, this study can select the best set of significant attributes to increase the prediction accuracy value in postoperative breast cancer patients.

Keywords: Breast Cancer; Logistic Regression; Survival Prediction; Data Mining; Model Evaluation

1. PENDAHULUAN

Saat ini, kanker payudara telah menjadi penyebab kematian kedua tertinggi akibat kanker pada perempuan, setelah kanker leher Rahim, dan menjadi jenis kanker yang paling umum di kalangan wanita. Menurut World Health Organization (WHO), sekitar 8-9% wanita berpotensi mengalami kanker payudara. Fakta ini menunjukkan bahwa kanker payudara merupakan jenis kanker yang paling sering dijumpai pada poplasi wanita. Setiap tahun, lebih dari 250.000 kasus baru kanker payudara diagnosis di Eropa dan sekitar 175.000 kasus di Amerika Serikat[1]. Pada tahun 2018, jumlah kasus kanker meningkat menjadi 18,1 juta dan jumlah kematian meningkat menjadi 9,6 juta. Salah satu dari 5 laki-laki dan satu dari 6 wanita diseluruh dunia akan menderita kanker selama hidup mereka. Kanker paru-paru, prostat, usus besar, perut dan hati adalah kanker yang paling umum terjadi pada pria, sedangkan kanker payudara, usus besar, paru-paru, leher rahim, dan tiroid adalah kanker yang paling umum terjadi pada wanita[2].

Penyebab kanker payudara melibatkan berbagai faktor genetik, hormonal, lingkungan dan gaya hidup. Risiko kanker payudara meningkat seiring bertambahnya usia, dengan memahami faktor risiko tersebut upaya pencegahan dan deteksi dini dapat juga dilakukan dengan menggunakan metode Prediksi Regresi Logistik yang dimanfaatkan untuk memprediksi probabilitas terjadinya suatu peristiwa biner. Peristiwa biner mengacu pada peristiwa atau situasi yang hanya memiliki dua kemungkinan atau hasil yang berlawanan. Dalam kasus ini, variabel dependen adalah suatu kategori dengan dua pilihan (contohnya, sukses atau gagal, hidup atau mati, ya atau tidak). Fungsi logistik atau sigmoid digunakan dalam regresi logistik untuk mengubah hasil prediksi linier menjadi probabilitas yang berkaitan antara 1 bertahan 5 tahun kedepan dan 2 tidak bertahan atau meninggal 5 tahun kedepan[3].

Dengan statistik yang mengkhawatirkan ini, sangat jelas bahwa upaya intensif diperlukan untuk mendeteksi dan mengobati kanker payudara sejak dini. Salah satu model yang sangat berharga dalam hal ini adalah Metode Prediksi Regresi Logistik, yang mengacu pada situasi dengan dua kemungkinan hasil yang berlawanan. Dalam konteks medis, terutama dalam kasus kanker payudara, regresi logistik dapat digunakan untuk memprediksi apakah seorang pasien akan bertahan hidup selama lima tahun ke depan atau tidak.

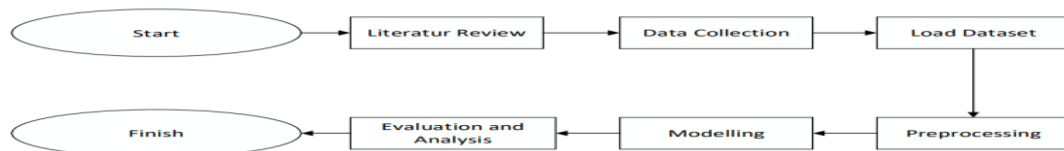
Urgensi ini tidak bisa diabaikan, mengingat setiap statistik bukan hanya angka, tetapi mencerminkan kehidupan nyata yang dipengaruhi oleh penyakit yang serius ini. Oleh karena itu, mengembangkan dan menerapkan metode prediksi yang efektif seperti regresi logistik adalah suatu upaya dalam memprediksi kanker payudara.

Beberapa penelitian sebelumnya telah membahas prediksi yang berjudul Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien oleh Irfan Solikin menggunakan teknik data mining dengan akurasi tertinggi menggunakan algoritma Support Vector Machine (SVM) mencapai 97,59% [4]. Selanjutnya penelitian yang dilakukan oleh Qatrunnada Refa Cahyani et al, dengan judul Prediksi Risiko Penyakit Diabetes Menggunakan Algoritma Regresi Logistik dengan akurasi mencapai 0,76% [5]. Berikutnya penelitian yang dilakukan oleh Nurmalitasari et al, dengan judul Prediksi Performa Mahasiswa Menggunakan Regresi Logistik dengan atingkat kesalahan prediksi 8% [6]. Selanjutnya penelitian yang dilakukan oleh Andita dani Achmad dengan judul Klasifikasi Breast Cancer Menggunakan Metode Logistic Regression dengan akurasi data latih sebesar 76,04% dan akurasi data uji sebesar 83,33% [7]. Selanjutnya penelitian yang dilakukan oleh N. K. Kutha Ardana et al dengan judul Perbandingan Metode KNN, Naïve Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara dengan hasil terbesar akurasi dari algoritma KNN 100% [8]. Penelitian terakhir yang dilakukan oleh Ayu Ratna Juwita et al dengan judul Perbandingan Algoritma C4.5 dan Regresi Linear dalam Memprediksi Keterlambatan Pembayaran Uang Kuliah hasil akurasi C4.5 nilai error yang dihasilkan lebih besar dari regresi linier, Dimana nilai error paling kecil maka akurasi tersebut semakin akurat [9].

Berdasarkan permasalahan yang telah dijelaskan sebelumnya dengan hasil penelitian yang telah dilakukan sebelumnya terakrit dengan implementasi pemilihan fitur untuk pengoptimalan algoritma machine learning dalam prediksi pasien pasca operasi kanker payudara, kami melakukan perbaikan pada penelitian ini menggunakan prediksi regresi logistik [10]. Kontribusi kami yaitu menggunakan teknik regresi logistic yang menggabungkan elemen regresi linier dengan fungsi logistic untuk memodelkan hubungan antara variable independen dan probabilitas kejadian. Sehingga model prediksi ini membantu meramalkan hasil pasien kanker payudara pasca operasi.

2. METODOLOGI PENELITIAN

Pada Gambar 1 Tahapan ini dilakukan dalam penelitian diawali dengan studi literatur. Studi literatur dilakukan dengan tujuan untuk menemukan landasan teori yang digunakan, serta mencari literatur ilmiah terkait untuk mendukung penelitian [11]. Literatur review dilakukan dengan cara pengumpulan data, pemilihan atribut penelitian, algoritma data mining yang digunakan untuk membangun model dan kriteria evaluasi yang bertujuan untuk menentukan landasan teoritis yang digunakan serta mencari literatur ilmiah yang relevan untuk mendukung penelitian [12]. Dalam keseluruhan penelitian ini, langkah-langkah atau tahapan penelitian adalah sebagai berikut:



Gambar 1. Flowchart Alur Penelitian

2.1 Data Collection dan Load Dataset

Data Collection dan Load Dataset, dimulai dengan pengumpulan dataset untuk dilakukan pemilihan data/target data. Penelitian ini digunakan dataset kanker payudara diperoleh dari Situs Kaggle dari rn is here, data tersedia di: <https://www.kaggle.com/code/raniaprillyap/prediksi-kelangsungan-hidup-pasien/notebook>. Terdiri dari 306 baris dan 4 kolom. Seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Deskripsi Dataset Kaggle Kelangsungan Hidup Pasien.

Nomor	Fitur Haberman Kaggle	Deskripsi
1	Age	Usia Pasien Pada Saat operasi
2	Patients Years	Tahun Operasi Pasien
3	N Positive ax	Jumlah Kelenjar Getah Bening Aksila Positif yang terdeteksi
4	Survival Status	Status bertahan hidup 1=pasien bertahan 5 tahun atau lebih 2=pasien meninggal dalam waktu 5 tahun

2.2 Preprocessing

Preprocessing, bertujuan untuk membersihkan, mengubah, dan menyiapkan data awal agar lebih cocok untuk analisis atau pemodelan [13], beberapa langkah yang dilakukan dalam penelitian ini yaitu :



- 1) Pembersihan data tujuan tahap ini yaitu untuk membersihkan data agar dapat digunakan dengan model. Proses ini mencakup beberapa pengecekan informasi penting dalam data, seperti nilai yang hilang (missing value), duplikasi data dan distribusi data.
- 2) Normalisasi data yaitu proses mengubah nilai-nilai dalam dataset menjadi rentang nilai yang lebih seragam atau standar.

2.3 Modelling

Modelling, dalam data mining teknik modelling mencakup berbagai pendekatan untuk mengekstraksi informasi berharga dari data besar[14]. Adapun beberapa teknik modelling yang akan dipakai seperti Prediksi, Regresi Logistik dan pengukuran kinerja dengan Confussion Matrix. Dalam data mining teknik modelling mencakup berbagai pendekatan untuk mengekstraksi informasi berharga dari data besar. Adapun beberapa teknik modelling yang akan dipakai seperti berikut:

1) Prediksi

Prediksi merupakan bidang pengetahuan yang diterapkan secara sistematis untuk memperoleh informasi berdasarkan data. Ini melibatkan proses estimasi terhadap peristiwa di masa depan dengan menggunakan berbagai informasi atau data yang memiliki signifikansi dari periode sebelumnya[15].

2) Regresi Logistik

Regresi logistik adalah salah satu algoritma machine learning yang dapat digunakan untuk melakukan tugas untuk memprediksi probabilitas pada variable target. Selanjutnya, kita menggunakan fungsi perkiraan kompleks yang dapat ditentukan sebagai fungsi sigmoid atau fungsi logistik. Persamaan regresi logistik digunakan untuk memodelkan probabilitas hasil biner (1 atau 2)[16]. Berdasarkan kumpulan data, model menggunakan fungsi logistic untuk memprediksi probabilitas. Fungsi logistic didefinisikan sebagai:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 - b_1x_1 + b_2x_2 + \dots + b_kx_k)}} \tag{1}$$

Dimana di, yaitu probabilitas kelas positif, adalah basis logaritma alami, adalah istilah intersep dan koefisien variable predictor. Parameter C dalam regresi logistic mewakili kebalikan dari kekuatan regularisasi $P(Y=1) \in [0, 1]$, $b_0, b_1, \dots, b_k, x_1, x_2, \dots, x_k$.

3) Pengukuran kinerja dengan confusion matriks

Penerapan ini menggunakan metode himpunan data dari model dengan pengukuran Akurasi, Presisi dan Recall. Perumusan kinerja sebagai berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

TP = True Positive (Diprediksikan dengan benar pasien bertahan 5 tahun ke depan)

TN = True Negative (Diprediksikan sebagai pasien tidak bertahan 5 tahun ke depan)

FP = False Positive (Diprediksikan sebagai pasien bertahan 5 tahun kedepan tetapi sebenarnya tidak)

FN = False Negatif (Diprediksikan sebagai pasien tidak bertahan 5 tahun ke depan tetapi sebenarnya iya)

Akurasi adalah suatu pola tolak alat ukur untuk memprediksi kelas data dari data yang akan datang[17]. Presisi bisa digambarkan sebagai tingkat akurasi data diberikan oleh model yang diminta dengan hasil prediksi[18]. Presisi dapat diartikan sebagai rasio antara jumlah prediksi positif yang benar dengan total jumlah prediksi positif, presisi menunjukkan berapa banyak yang benar-benar positif. Nilai presisi ini dapat dihitung menggunakan suatu persamaan. Recall digunakan untuk mengukur sejauh mana model berhasil mengidentifikasi semua data yang benar-benar positif. Dengan kata lain, recall dapat diartikan sebagai proporsi prediksi benar positif dibandingkan dengan keseluruhan data yang sebenarnya positif, nilai recall dapat dihitung menggunakan persamaan[19]. Inilah evaluasi performa individu dari setiap algoritma yang dipresentasikan melalui confusion matriks. Dalam matriks tersebut, label X mewakili data uji, sementara label Y prediksi model.

2.4 Hasil dan Evaluasi Prediksi

Evaluation and Analysis, pada tahap ini akan ditentukan apakah hasil dari tahap sebelumnya mampu memenuhi tujuan yang telah ditetapkan.

1) Pengujian MAE (Mean Absolute Error)

Secara umum, istilah MAE digunakan untuk sebagai metode pengukuran untuk menilai kesalahan prediksi dalam analisis time series, mirip dengan RMSE, tetapi dengan rumus perhitungan yang lebih sederhana[20]. Penggunaan metode MAE ini bertujuan untuk memberikan kepastian sejauh mana hasil prediksi mendekati tingkat akurasi yang diinginkan. Rumus Mae dapat dinyatakan sebagai berikut:

$$MAE = \sum \frac{|Y^1 - Y|}{n} \tag{5}$$

Rumus tersebut adalah Mean Absolute Error (MAE), yang berfungsi untuk menghitung rata-rata kesalahan absolut antara nilai prediksi (Y^1) dan nilai actual (Y). MAE diperoleh dengan menjumlahkan nilai absolut dari selisih antara nilai prediksi dan nilai actual untuk setiap periode, kemudian membagi totalnya dengan jumlah data (n). Dalam konteks ini, Y^1 merujuk pada nilai prediksi, Y pada nilai actual dan n merupakan jumlah data keseluruhan. MAE mengindikasikan seberapa besar kesalahan dalam model peramalan tanpa memperhatikan arah kesalahan, di mana nilai MAE yang lebih rendah menandakan model peramalan yang lebih akurat.

2) MSE (Mean Squared Error)

Teknik perhitungan Mean Squared Error (MSE) melibatkan penjumlahan selisih antara data peramalan dan data actual. Semakin kecil nilai MSE, semakin baik hasil peramalan. Nilai MSE dapat dihitung dengan menggunakan persamaan berikut:

$$MSE = \frac{\sum_{t=1}^n (A_t - F_t)^2}{n} \tag{6}$$

Rumus yang disebutkan adalah Mean Squared Error (MSE), yang digunakan untuk menghitung rata-rata kesalahan kuadrat antara nilai actual (A_t) dan nilai prediksi (F_t) dalam suatu periode waktu. MSE diperoleh dengan menjumlahkan kuadrat dari selisih antara nilai actual dan nilai prediksi untuk setiap periode untuk setiap periode (t) dari $t = 1$ hingga $t = n$, kemudian membagi totalnya dengan jumlah data (n). Dalam rumus ini, A_t adalah nilai actual pada periode t , F_t adalah nilai prediksi pada periode t , dan n adalah jumlah total data. MSE menunjukkan seberapa besar kesalahan dalam model peramalan, di mana nilai MSE yang lebih rendah menandakan model peramalan yang lebih akurat.

3) R-Square

R-Square merupakan ukuran statistic yang digunakan untuk mengevaluasi seberapa baik model regresi linier cocok dengan data yang diamati. Nilai R-square berkisar antara 0 hingga 1, dan semakin tinggi nilainya semakin baik modelnya sesuai dengan data. Berikut adalah rumus untuk R-square:

$$R^2 = 1 - \frac{SSE}{SST} \tag{7}$$

R^2 = Koefisien determinasi

SSE (Sum of Squared Errors) = Jumlah kuadrat dari selisih antara nilai yang diamati dan nilai yang diprediksi oleh model

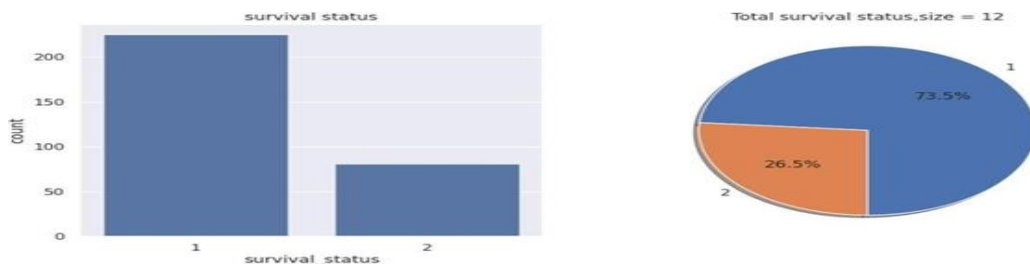
SST (Total Sum of Squares) = jumlah kuadrat dari selisih antara nilai yang diamati dan rata-rata nilai yang diamati

Rumus tersebut digunakan untuk menghitung Koefisien Determinasi (R^2), yang merupakan ukuran statistik untuk menilai seberapa baik model regresi cocok dengan data yang diamati. Nilai R^2 dihitung dengan menggunakan formula 1 dikurangi rasio antara Sum of Squared Errors (SSE) dan Total Sum of Squares (SST). SSE adalah jumlah kuadrat dari selisih antara nilai yang diamati dan nilai yang diprediksi oleh model, sedangkan SST adalah jumlah kuadrat dari selisih antara nilai yang diamati dan rata-rata nilai yang diamati. Secara sederhana, R^2 mengukur proporsi variabilitas total dalam data yang dapat dijelaskan oleh model regresi. Nilai R^2 berkisar antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan bahwa model tersebut lebih efektif dalam menjelaskan variabilitas data.

3. HASIL DAN PEMBAHASAN

3.1 Data Collection dan Load Dataset

Menampilkan distribusi data dari tahapan data collection dan load dataset sehingga diperoleh distribusi data dari survival status yang terdapat 2 kategori antara 1 dan 2 karena diperlukan untuk keputusan tahapan sebelum preprocessing. Pada distribusi dataset prediksi kelangsungan hidup pasien ini terdapat 73.5% bertahan dan 26.5% tidak ditampilkan dengan visualisasi pada Gambar 2.



Gambar 2. Kelas Distribusi EDA dalam dataset.

3.2 Preprocessing

Dalam menghasilkan data berkualitas, teknik berikut digunakan:

- 1) Pembersihan Data, himpunan data dibersihkan terlebih dahulu dan nilai pengamatan disajikan dengan tidak mengulangi awalan dari setiap fitur. Ini digunakan untuk mendefinisikan dan menghapus kolom id, data yang tidak lengkap (nilai yang hilang) dan data duplikat tanpa mengurangi konten data. Dalam data ini, tidak ada nilai yang hilang, dan data duplikat seperti yang ditunjukkan pada Gambar 3.

```
Age          0
Patients_Years  0
N_positive_ax  0
survival_status  0
dtype: int64
```

```
[ ] print(df.shape)
df.duplicated().sum()

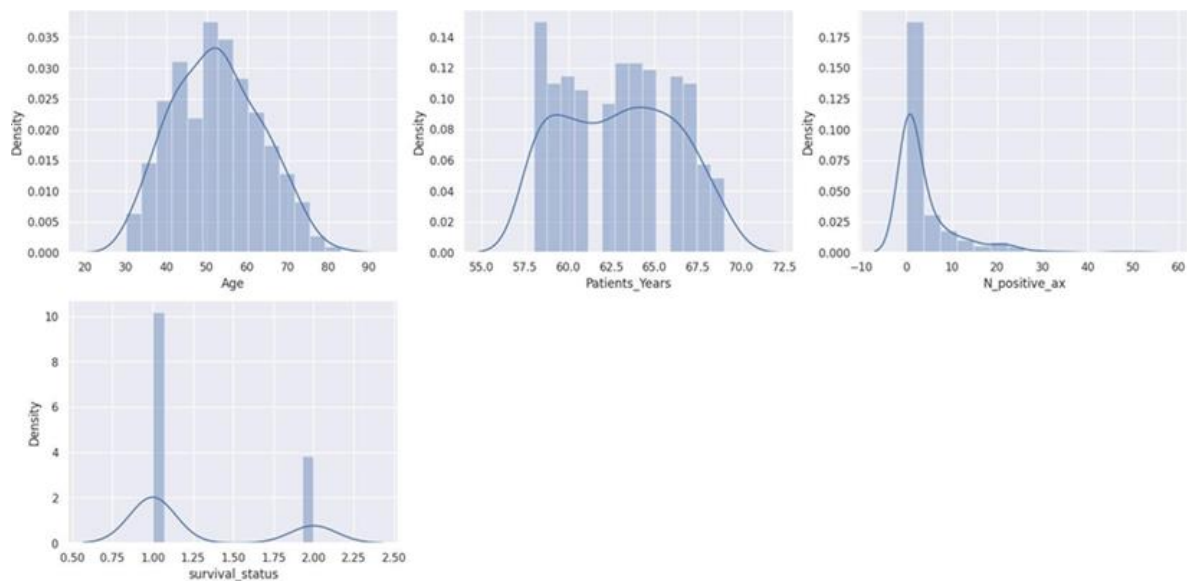
(306, 4)
17
```

Gambar 3. Mencari nilai hilang dan data duplikat

- 2) Normalisasi Data juga dilakukan dengan menggunakan MinMax agar terstandarisasi. Rumus dalam:

$$\text{Normalized Value} = \frac{\text{Original Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}} \tag{8}$$

Normalized Value adalah nilai asli yang akan dinormalisasi, Min Value adalah nilai minimum dalam rentang data dan Max Value adalah nilai maksimum dalam rentang data. Tampilkan histogram hasil data numerik sesudah normalisasi. Normalisasi Min-Max mengonversi setiap nilai dalam dataset ke dalam rentang antara 1 dan 2. Ini adalah salah satu metode normalisasi yang umum digunakan untuk memastikan bahwa semua fitur dalam dataset memiliki skala yang serupa, yang dapat bermanfaat dalam beberapa algoritma machine learning. Sebagai contoh, jika anda memiliki suatu nilai (Original Value) dalam rentang data tertentu, dan anda ingin menormalisasi nilai tersebut ke dalam rentang antara 1 dan 2, anda dapat menggunakan rumus di atas dengan mengetahui nilai minimum (Min Value) dan nilai maksimum (Max Value) dari dataset tersebut.



Gambar 4. Histogram data numerik dan sesudah normalisasi

Pada Gambar 4 Histogram di tunjukkan bahwa hasil distribusi data numerik sesudah normalisasi menggunakan MinMax bekerja dengan baik pada data tersebut. 289 data dibagi menjadi data latih dan data uji. Kami menggunakan dua data yaitu 90:10 atau dari segi presentase sebanyak 90% data uji dan 10% data latih. Karena tidak banyak outlier dalam dataset uji 10% cukup untuk mengevaluasi model akhir secara objektif. Sehingga data training sebanyak 260 dan data testing sebanyak 29.

3.3 Modelling

Hasil setelah preprocessing data, performa pengklasifikasi divisualisasikan dengan matriks performa yang berbeda. Preprocessing himpunan data mencakup penggantian nilai yang hilang dan pengambilan nilai minimum dan maksimum. Scaling data dan dimensional reduction diterapkan di semua algoritma machine learning yang digunakan pada penelitian ini. Confusion matrix dan nilai error dari MSE, MAE dan R-Square juga digunakan untuk evaluasi kinerja algoritma seperti pada Tabel 2.

Tabel 2. Performa dan Algoritma Regresi Logistik berdasarkan Confussion Matrix

Hasil	1 (bertahan 5 tahun kedepan)	2 (tidak bertahan 5tahun kedepan)
Accuracy	86%	86%

Hasil	1 (bertahan 5 tahun kedepan)	2 (tidak bertahan 5tahun kedepan)
Presisi	87%	83%
Recall	95%	62%



Gambar 5. Confussion Matrix dan Algoritma Regresi Logistik

Pada Gambar 5 Confussion matrik ini menjelaskan bahwa True Positive sebanyak 20, True Negative sebanyak 5, False Positive sebanyak 1, dan False Negative sebanyak 3. Jumlah dari keseluruhan yaitu 29 sesuai dengan data yang telah di uji.

Tabel 3. Hasil Nilai Error dari Algoritma Regresi Logistik

Hasil	Nilai
MSE	0.1373103448275862
MAE	0.1373103448275862
R-Square	0.3095238095238094

Tabel 3 menjelaskan metode Regresi Logistik memprediksi pasien pasca operasi kanker payudara, rata-rata MSE yang diperoleh adalah 0.1 dan nilai dari MAE diperoleh 0,1 dan Nilai R-square diperoleh 0.3, nilai MSE dan MAE lebih kecil mencerminkan kualitas prediksi sangat baik, ini dapat dijelaskan bahwa semakin mendekati nilai actual semakin kecil nilai yang diperoleh MSE dan MAE dibandingkan nilai R-Square. Dalam konteks ini, R2 dapat memberikan informasi tentang sejauh mana variasi dalam data dapat dijelaskan oleh model regresi. Semakin tinggi nilai R2, semakin besar proporsi variabilitas yang dijelaskan oleh model. Untuk nilai error dari Mean Square Error (MSE) yaitu 0,13, Mean Absolute Error (MAE) yaitu 0,13, dan R-Square yaitu 0,3.

4. KESIMPULAN

Berdasarkan penelitian yang dilakukan mengenai prediksi hasil pasien pasca operasi kanker payudara, Dilakukan penggunaan dengan metode regresi logistic yang diaplikasikan untuk prediksi dengan nilai 1 yaitu bertahan dan 2 itu tidak bertahan. Penelitian ini menunjukkan data sebanyak 289 dibagi menjadi data latih dan data uji 90:10 atau dari segi presentase sebanyak 90% data latih dan 10% data uji dengan menghasilkan akurasi mencapai 86%. Evaluasi melalui confussion matriks dengan menunjukkan model perhitungan TP=20, TN=5, FP=1 dan FN=3. Data ini mempunyai nilai error seperti MSE sebesar 0.13793103448275862, MAE sebesar 0.13793103448275862, dan R-Square sebenar 0.3095238095238094. Untuk penelitian selanjutnya, beberapa langkah dapat diambil guna mencapai tingkat akurasi yang optimal dan meningkatkan hasil prediksi pada pasien pasca operasi kanker payudara. Ini termasuk eksplorasi penggunaan metode yang lain seperti SVR (Support Vector Regression). Dengan menerapkan algoritma yang berbeda, penelitian ini bertujuan untuk memilih set attribute terbaik yang dapat signifikan meningkatkan nilai akurasi prediksi pada pasien pasca operasi kanker payudara.

REFERENCES

- [1] A. Milenia and D. Retnaningsih, “Penerapan Terapi Guided Imagery pada Pasien Dengan Kanker Payudara Dengan Nyeri Sedang,” *J. Manaj. Asuhan Keperawatan*, vol. 6, no. 1, pp. 35–42, 2022, doi: 10.33655/mak.v6i1.129.
- [2] Masruroh, “Perawatan Pasien Kanker,” p. 212, 2022, [Online]. Available: www.meddam.co.id
- [3] A. Avini, K. W. Patunduk, S. Sumarni, H. Harbianti, A. Pratiwi, and R. Hidayat, “Analisis Model Cox Proportional Hazard dan Regresi Logistik sebagai Upaya Pencegahan Covid-19 di Kota Palopo,” *Inferensi*, vol. 5, no. 2, p. 105, 2022,



doi: 10.12962/j27213862.v5i2.14064.

- [4] I. Solikin, “Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien,” *Fidel. J. Tek. Elektro*, vol. 3, no. 3, pp. 63–67, 2021, doi: 10.52005/fidelity.v3i3.100.
- [5] Q. R. Cahyani et al., “Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK,” *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.
- [6] N. Nurmalitasari and E. Purwanto, “Prediksi Performa Mahasiswa Menggunakan Model Regresi Logistik,” *J. Deriv. J. Mat. dan Pendidik. Mat.*, vol. 9, no. 2, pp. 145–152, 2022, doi: 10.31316/jderivat.v9i2.2639.
- [7] A. D. Achmad, “Klasifikasi Breast Cancer Menggunakan Metode Logistic Regression,” *Jtriste*, vol. 9, no. 1, pp. 143–148, 2022, [Online]. Available: <https://jurnal.kharisma.ac.id/jtriste/article/download/384/201>
- [8] N. K. K. Ardana et al., “Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara,” *Jambura J. Math.*, vol. 5, no. 2, pp. 404–418, 2023, doi: 10.34312/jjom.v5i2.21103.
- [9] T. Juwita, Ayu Ratna, Rizky Pratama, “Perbandingan algoritma c4.5 dan regresi linear dalam memprediksi keterlambatan pembayaran uang kuliah 1,” *Konf. Nas. Penelit. dan Pengabd.*, vol. 2, no. 1, pp. 55–62, 2023, [Online]. Available: <https://journal.ubpkarawang.ac.id/index.php/ProsidingKNPP/article/download/4898/3535>
- [10] Power News, “Kampus Merdeka: Studi Independen E-Commerce,” *PowerCommerce.Asia*, 2021.
- [11] H. Mubarak, “Studi Literatur Menumbuhkan Motivasi Belajar Siswa Melalui Strategi Pembelajaran Berdiferensiasi Dalam Konteks Pedagogi,” *J. Ilmu Pendidik. Nas.*, vol. 1, no. 1, pp. 01–07, 2023, doi: 10.59435/jipnas.v1i1.45.
- [12] S. Dewi and N. Suarna, “Penerapan Data Mining Dalam Algoritma Fp Growth Untuk Menentukan Pola Pembelian Pelanggan Online Shop Cirebon Laku,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 1180–1186, 2024, doi: 10.36040/jati.v8i1.8944.
- [13] S. Wahyu, “Konferensi Nasional Ilmu Komputer (KONIK) 2023 Perbandingan Model Algoritma Klasifikasi Pada Analisis Sentimen Opini Masyarakat Terhadap Layanan Kereta Cepat Jakarta Bandung (The Whoosh),” pp. 218–225, 2023.
- [14] R. Zapar, D. Pratama, K. Kaslani, C. L. Rohmat, and F. Faturrohman, “Penerapan Model Regresi Linier Untuk Prediksi Harga Saham Bank Bca Pada Bursa Efek Indonesia,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 196–202, 2024, doi: 10.36040/jati.v8i1.8215.
- [15] K. F. Irnanda, A. P. Windarto, and I. S. Damanik, “Optimasi Particle Swarm Optimization Pada Peningkatan Prediksi dengan Metode Backpropagation Menggunakan Software RapidMiner,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 1, p. 122, 2022, doi: 10.30865/jurikom.v9i1.3836.
- [16] A. Setiawan, F. Setivani, and T. Mahatma, “Performance Comparison of Decision Tree and Logistic Regression Methods for Classification of Snp Genetic Data,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 1, pp. 0403–0412, 2024, doi: 10.30598/barekengvol18iss1pp0403-0412.
- [17] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, “Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.
- [18] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes,” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.
- [19] D. Cahyanti, A. Rahmayani, and S. A. Husniar, “Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara,” *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.
- [20] I. Permana and F. N. S. Salisah, “Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation,” *Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022, doi: 10.57152/ijirse.v2i1.311.