JOMLAI: Journal of Machine Learning and Artificial Intelligence
Vol. 1, No. 2, Juni 2022, pp. 107~114
ISSN: 2828-9102 (Print) | 2828-9099 (Online)
DOI: 10.55123/jomlai.v1i2.598

Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik

Diabetes Risk Prediction using Logistic Regression Algorithm

Qatrunnada Refa Cahyani[1], Mochammad Januar Finandi[2], Jathu Rianti[3], Devi Lestari Arianti[4], Arya Dwi Pratama Putra[5]
[1,3] Universitas Diponegoro, Semarang, Indonesia
[2,5] Universitas Nasional, Jakarta, Indonesia
[4] Politeknik Elektronika Negeri Surabaya, Indonesia

Article Info

Genesis Artikel:

Diterima, 23 Juni 2022
Direvisi, 24 Juli 2022
Disetujui, 26 Juli 2022

Kata Kunci:

Diabetes
Regresi Logistik
Recall
Confusion Matrix

ABSTRAK

Banyak faktor yang mempengaruhi orang menderita diabetes, beberapa diantaranya yaitu tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, usia, jumlah kehamilan seseorang, ketebalan lipatan kulit, dan jumlah kadar insulin dalam tubuh. Regresi logistik merupakan salah satu alat statistik yang dapat digunakan dalam pemodelan klasifikasi tentang ada tidaknya yang mengalami diabetes. Tujuan dari penelitian ini adalah untuk memprediksi secara diagnostik apakah pasien menderita diabetes atau tidak. Hasil yang didapatkan adalah prediksi relatif rendah karena rentang nilai dari beberapa faktor penyebabnya sangat berjauhan sehingga dilakukan normalisasi agar rentang nilainya berdekatan. Hasilnya prediksi risiko diabetes menggunakan algoritma regresi logistik dengan normalisasi menghasilkan *recall* sebesar 55% sedangkan tanpa normalisasi sebesar 43%. Dengan demikian, normalisasi dapat meningkatkan kinerja prediksi risiko diabetes menggunakan algoritma regresi logistik. Model ini diharapkan dapat menjadi acuan untuk pengobatan penderita diabetes bagi dokter di rumah sakit dan di masyarakat untuk mengetahui cara menjaga pola hidup dan cara menghindari penyakit diabetes dilihat dari variabel yang mempengaruhi terjadinya penyakit.

Keywords:

Diabetes
Logistic Regression
Recall
Confusion Matrix

ABSTRACT

Many factors affect people suffering from diabetes, some of which are high blood pressure, excess sugar levels, weight, genetic history of diabetes, age, number of pregnancies, skin fold thickness, and the amount of insulin levels in the body. Logistic regression is a statistical tool that can be used in classification modeling about the presence or absence of diabetes. The aim of this study is to predict diagnostically whether a patient has diabetes or not. The results obtained are relatively low predictions because the ranges of values of several factors that cause it are very far apart so normalization is carried out so that the ranges of values are close together. The result is that diabetes risk prediction using a logistic regression algorithm with normalization resulted in a recall of 55% while without normalization it was 43%. Thus, normalization can improve the performance of diabetes risk prediction using a logistic regression algorithm. This model is expected to be a reference for the treatment of diabetics for doctors in hospitals and in the community to find out how to maintain a lifestyle and how to avoid diabetes in terms of the variables that affect the occurrence of the disease.

Penulis Korespondensi:

Qatrunnada Refa Cahyani,
Program Studi Matematika,
Universitas Diponegoro, Semarang, Indonesia
Email: refa220801@gmail.com

1. PENDAHULUAN

Diabetes merupakan suatu penyakit tidak menular yang cukup serius di mana pankreas tidak dapat memproduksi insulin secara maksimal [1], [2]. Diabetes dapat menyerang siapa saja tanpa mengenal usia baik lansia, orang dewasa, maupun anak-anak yang ditandai dengan meningkatnya kadar gula (glukosa) darah dalam tubuh manusia.

Diabetes dapat disebabkan oleh banyak faktor seperti tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, usia, jumlah kehamilan seseorang, ketebalan lipatan kulit, jumlah kadar insulin dalam tubuh, kurangnya aktivitas fisik dan pola hidup, serta diet tidak sehat [3], [4]. Faktor-faktor tersebut merupakan variabel yang digunakan dalam penelitian ini untuk membuat sistem cerdas yang dapat memprediksi penyakit diabetes.

Machine learning merupakan bagian dari kecerdasan buatan yang mampu mempelajari data dengan sendirinya. *Machine learning* adalah suatu model statistik untuk memprediksi data menggunakan komputer [5], [6]. Salah satu algoritma yang digunakan dalam *machine learning* adalah regresi logistik.

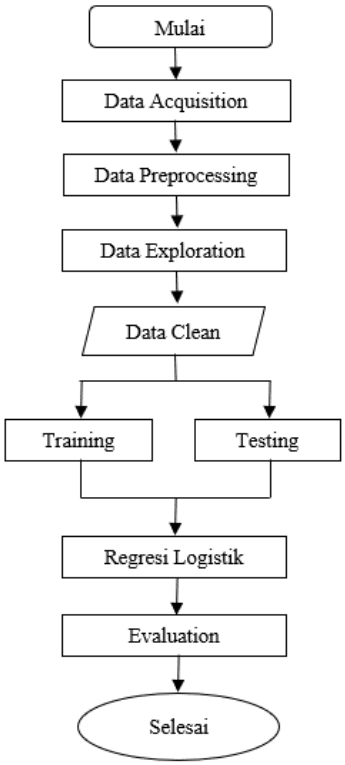
Regresi logistik adalah salah satu algoritma yang dapat digunakan dalam *machine learning* untuk melakukan tugas klasifikasi. Regresi logistik merupakan bentuk khusus analisis regresi dengan menggunakan respon biner dan prediktor yang dapat terdiri dari data kontinu, kategori, atau campuran antara keduanya. Analisis ini tidak memerlukan asumsi distribusi multivariat normal atau kesamaan matrik varian kovarian, serta dapat juga diterapkan dalam berbagai skala data [7], [8].

Penelitian menggunakan regresi logistik yang dilakukan oleh Gunawan et al dalam [9] menghasilkan prediksi akurasi penyakit diabetes melitus sebesar 72,22% pada regresi logistik tanpa *grid search*, sedangkan prediksi regresi logistik dengan *grid search* menghasilkan akurasi sebesar 83,33%. Penelitian lain yang dilakukan oleh Marna et al dalam [10] menghasilkan besarnya peluang faktor seorang mahasiswa yang memiliki ayah yang lulusan SMP dan ibu yang pendidikannya lulusan SMA (eksternal) memiliki waktu belajar kurang lebih 7 jam dan mahasiswa yang bersikap baik (internal) memperoleh IPK dibawah 3 sebesar 0,47 atau 47% sedangkan peluang memperoleh IPK lebih dari 3 adalah 0,53 atau 53%.

Berdasarkan penelitian-penelitian sebelumnya, maka penulis menggunakan algoritma regresi logistik untuk membuat sistem cerdas yang dapat melakukan prediksi diabetes atau tidak, sehingga dapat digunakan sebagai acuan untuk pengobatan penderita diabetes bagi dokter di rumah sakit dan di masyarakat untuk mengetahui cara menjaga pola hidup dan cara menghindari penyakit diabetes dilihat dari variabel yang mempengaruhi terjadinya penyakit.

2. METODE PENELITIAN

Pada penelitian ini terdapat beberapa langkah, antara lain: *data acquisition*, *data exploration*, *modelling*, dan *evaluation*. Metode penelitian dapat dilihat pada gambar 1.



Gambar 1. Langkah - Langkah Penelitian

Keterangan:

1. *Data Acquisition*
   Data acquisition merupakan suatu proses untuk mengumpulkan dan menganalisis informasi data seperti variabel dan nilai-nilai pada tiap variabel.
2. *Data Pre processing*
   Data pre processing merupakan suatu proses di mana data mentah diubah menjadi sebuah data yang mudah dimengerti.
3. *Data Exploration*
   Data exploration merupakan tahap eksplorasi data untuk memahami isi data.
4. *Data Clean*
   Data clean merupakan proses memperbaiki atau menghapus data yang salah, rusak, duplikat, atau tidak lengkap dalam suatu dataset.
5. *Training* dan *Testing*
   Training merupakan suatu proses melatih data supaya memperoleh sebuah parameter yang tepat. Sedangkan *testing* merupakan suatu proses yang dijalankan setelah validasi untuk membuktikan apakah model tersebut akurat.
6. Regresi Logistik
   Regresi logistik merupakan pemodelan dengan analisis regresi variabel yang dapat memprediksi hasil data dengan dua kemungkinan, misalnya ya dan tidak.
7. *Evaluation*
   Evaluation dilakukan dengan menghitung nilai-nilai pada metrik akurasi, presisi, *recall*, dan *f1-score*. Metriks tersebut digunakan untuk menentukan apakah model mempunyai performa yang baik atau tidak.

2.1. *Data Acquisition*

Data acquisition adalah tahap di mana dilakukan pengumpulan data apa yang diperlukan. Data yang digunakan pada penelitian ini berupa dataset yang berasal dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal dengan format .csv yang diperoleh melalui situs *kaggle* [11]. Mengenai karakteristik atribut atau variabel pada dataset dapat dilihat pada tabel 1.

Tabel 1. Karakteristik Dataset Pima India Diabetes

Variabel	Deskripsi Variabel	Jenis dan Pengukuran Variabel
Pregnancy	Jumlah kehamilan pada wanita	Numerik
Plasma Glucose	Diukur menggunakan tes toleransi glukosa oral dalam 2 jam	Numerik
Blood Pressure	Tekanan darah diastolic	Numerik (mm Hg)
Triceps	Ketebalan lipatan kulit trisep	Numerik (mm)
Serum Insulin	Serum insulin dalam 2 jam	Numerik (μ U / ml)
BMI	Body Mass Index (Indeks massa tubuh)	Numerik [berat dalam kg / (tinggi dalam m)^2]
Pedigree	Riwayat keturunan diabetes	Numerik
Age	Usia pasien	Numerik (tahun)

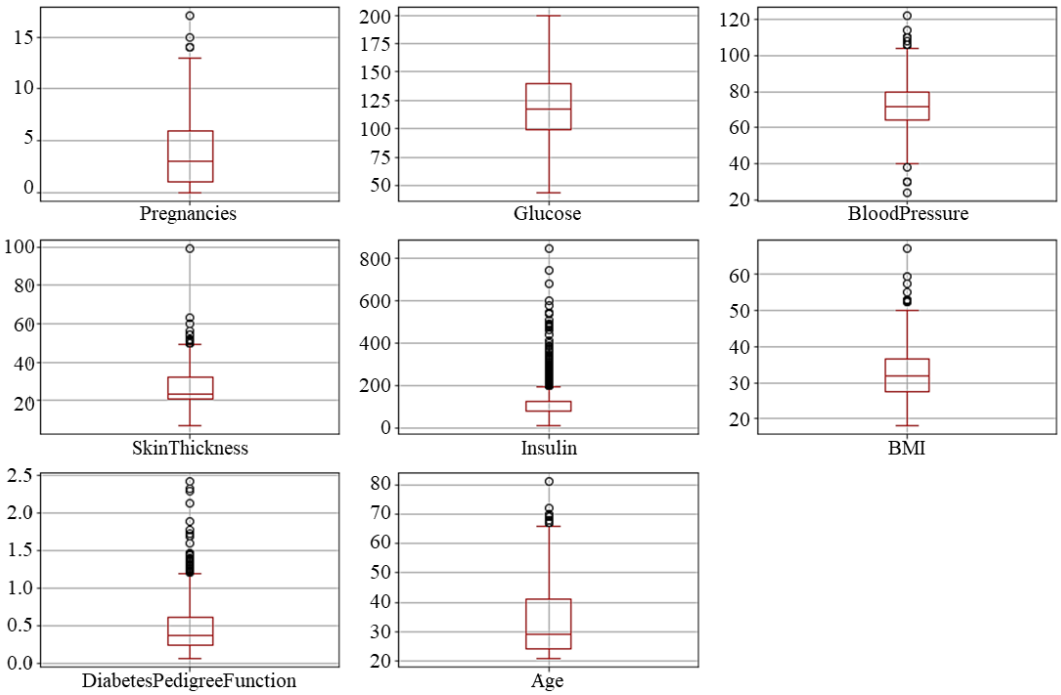
2.2. *Data Exploration*

Setelah tahap *data acquisition*, proses selanjutnya adalah *data exploration*. *Data exploration* adalah tahap yang bertujuan untuk memahami data. Pada proses eksplorasi ini kumpulan dataset yang telah didapatkan melalui situs kaggle, dilakukan *preprocessing* [12] dengan melihat data duplikat dan memeriksa *missing value*. Tabel 2 menunjukkan nilai-nilai yang hilang yaitu *NaN*.

Tabel 2. Data *Missing Value*

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Setelah mengecek data duplikat dan *missing value*, tahap *preprocessing* selanjutnya adalah melakukan pengecekan *outlier*. Gambar 1 menunjukkan *outliers* pada variabel. Dari 8 variabel Gambar 2 terlihat bahwa terdapat outliers pada setiap variabel, kecuali variabel Glucose. Outliers dihapus dengan menggunakan Z-Score. Selanjutnya dilakukan analisis korelasi antar variabelnya. Analisis korelasi variabel digunakan untuk *modelling* kemudian *evaluation*. Untuk model ini terdiri dari dua kasus yaitu adanya normalisasi (data *clean*) dan tanpa normalisasi sebelum *modelling*. Normalisasi digunakan agar nilai berada pada rentang yang berdekatan sehingga meningkatkan kinerja prediksi [13].



Gambar 2. *Outliers* pada Variabel

2.3. *Modelling* dan *Evaluation*

Modelling merupakan tahap dalam pembuatan model dari sistem klasifikasi yang dibuat. Pada penelitian ini menggunakan algoritma regresi logistik. Lib linear adalah algoritma yang baik digunakan dalam masalah optimasi regresi logistik untuk kumpulan data kecil, sedangkan sag dan saga lebih cepat untuk kumpulan data yang besar. Parameter ini mendukung regresi logistik dan mesin vektor dukungan linier [14].

Lib linear sangat efisien pada kumpulan data yang kecil, besar, dan jarang. Pemilihan algoritma ini didasarkan pada dataset yang dimiliki peneliti memiliki jumlah data yang ber kategori dan data numerik sehingga cocok menggunakan algoritma tersebut, dengan demikian dapat diketahui jumlah prediksi dan jumlah sebenarnya dari penderita diabetes. Setelah melakukan *training* dengan regresi logistik, selanjutnya melakukan hasil data *testing* dan *evaluation* model.

Evaluation dilakukan dengan memilih satu metrik diantara metrik *akurasi*, *presisi*, *recall*, atau *f1-score* yang berdasarkan perhitungan nilai *True Positive*, *True Negative*, *False Positive*, dan *False Negative* pada *confusion matrix* [15]. Nilai-nilai tersebut dapat digunakan sebagai perbandingan untuk pemilihan acuan metrik pada algoritma untuk model klasifikasi diabetes.

3. HASIL DAN PEMBAHASAN

3.1. *Data Acquisition*

Setelah melihat karakteristik variabel pada Tabel 1, lakukan analisis terhadap nilai-nilai pada setiap variabel. Tabel 3 menunjukkan lima data teratas dari dataset dengan 8 variabel independen dan 1 variabel dependen.

Tabel 3. *Head* Dataset

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

8 variabel dependen tersebut adalah *pregnancies*, *glucose*, *blood pressure*, *skin thickness*, insulin, BMI *(body mass index)*, *diabetes pedigree function*, dan age. Sedangkan 1 variabel dependen adalah *outcome*. Setiap variabel memiliki rentang nilai yang berbeda-beda. Rentang nilai tiap variabel dapat dilihat pada Tabel 4.

Tabel 4. Rentang Nilai Tiap Variabel

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.078	21.000	0.000
Max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Tabel 4 menunjukan bahwa rentang nilai setiap variabel yang berbeda. Adapun rentang nilai setiap variabel pada tabel 4 dapat dijelaskan sebagai berikut.
Variabel *pregnancies* memiliki nilai minimum 0.000 dan nilai maximum 17.00
Variabel *glucose* memiliki nilai minimum 0.000 dan maximum 199.00
Variabel *blood pressure* memiliki nilai minimum 0.000 dan maximum 122.00
Variabel *skin thickness* memiliki nilai minimum 0.000 dan maximum 99.00
Variabel insulin memiliki nilai minimum 0.000 dan maximum 846.00
Variabel BMI (*Body Mass Index*) memiliki nilai minimum 0.000 dan maximum 67.10
Variabel *diabetes Pedigree Function* memiliki nilai minimum 0.078 dan maximum 2.42
Variabel *age* memiliki nilai minimum 21.000 dan maximum 81.00
Variabel *outcome* memiliki nilai minimum 0.000 dan maximum 1.00

3.2. *Preprocessing*

Data yang telah didapatkan dari situs *kaggle* perlu dibersihkan terlebih dahulu dengan pengecekan data duplikat, *missing value*, dan *outlier*. Pada 768 data ini tidak terdapat data duplikat serta tidak terdapat *missing value*, hanya saja terdapat banyak nol (0) pada variabel *glucose*, *blood pressure*, *skin thickness*, insulin, dan BMI (*Body Mass Index*) sehingga termasuk pada nilai yang hilang. Nilai yang hilang ini kemudian diganti dengan mengisi nilai tersebut dengan nilai rata-rata seperti yang terlihat pada Tabel 5.

Tabel 5. *Missing Value* Diganti dengan Nilai Rata-rata

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148.0	72.0	35.00000	79.799479	33.6	0.627	50	1
1	1	85.0	66.0	29.00000	79.799479	26.6	0.351	31	0
2	8	183.0	64.0	20.536458	79.799479	23.3	0.672	32	1
3	1	89.0	66.0	23.00000	94.000000	28.1	0.167	21	0
4	0	137.0	40.0	35.00000	168.000000	43.1	2.288	33	1

Setelah mengatasi *missing* value, kemudian memeriksa *outlier* dan menghapusnya dengan melihat *Z-score* seperti pada Tabel 6.

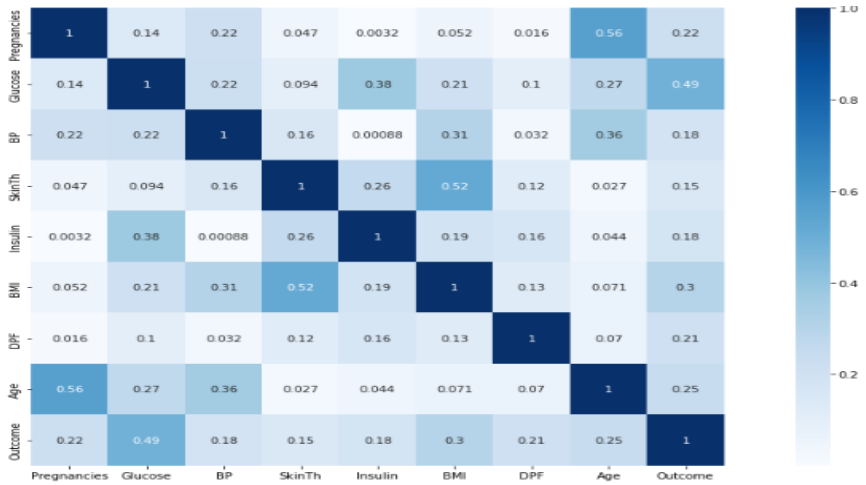
Tabel 6. *Outlier* Dihapus dengan *Z-Score*

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35.00000	79.799479	33.6	0.627	50	1
1	1	85	66	29.00000	79.799479	26.6	0.351	31	0
2	8	183	64	20.536458	79.799479	23.3	0.672	32	1
3	1	89	66	23.00000	94.000000	28.1	0.167	21	0
5	5	116	74	20.536458	79.799479	25.6	0.201	30	0

Dapat dilihat data keempat dihapus karena nilainya jauh dari *Z-score* sehingga jumlah data semula 768 menjadi 718 data. Selanjutnya melihat korelasi variabel untuk menentukan variabel apa saja yang digunakan dalam *modelling*.

3.3. *Data Exploration*

Data yang sudah bersih kemudian dilihat korelasi (hubungan) antar variabel. Hubungan antar variabel berguna untuk menentukan variabel apa saja yang digunakan untuk *modelling*. Berikut peta korelasi antar variabel yang ditunjukkan oleh Gambar 2.



Gambar 2. Peta Korelasi Antar Variabel

Jika nilai korelasi > 0 maka terdapat korelasi positif. Sementara nilai satu variabel meningkat, nilai variabel lainnya juga meningkat. Jika persamaan korelasi = 0 maka tidak ada korelasi. Jika korelasi < 0 maka ada korelasi negatif. Sementara satu variabel meningkat, variabel lainnya menurun. Ketika korelasi diperiksa, ada 2 variabel yang bertindak sebagai korelasi positif terhadap variabel dependen *outcome*, variabel tersebut adalah *glucose*. Seiring peningkatan ini, variabel dependen juga meningkat. Dengan demikian, semua variabel digunakan untuk *modelling* karena korelasinya berdekatan.

3.3. *Modelling* dan *Evaluation*

Modeling dilakukan pada data *testing*, data dipisahkan (*split*) menjadi data *training* dan *testing* dengan rasio 70:30 sehingga dari keseluruhan data berjumlah 718, jumlah data *training* sebanyak 502 dan *testing* yang digunakan untuk *modelling* sebanyak 216 data. Model ini menggunakan semua variabel independen karena hampir semua variabel memiliki korelasi yang mendekati 1.

Model dengan algoritma regresi logistik memiliki hasil yang berbeda dengan menggunakan normalisasi dan tanpa normalisasi sebelum *split* data sehingga dibagi menjadi dua kasus sebagai berikut.

1. *Modelling* dan Evaluasi Tanpa Normalisasi

   Model yang telah dirancang menggunakan algoritma regresi logistik dengan liblinear sebagai solver tanpa normalisasi mendapat skor model pada data *training* sebesar 0.787 dan pada data *testing* sebesar 0.75. Kemudian data *testing* dievaluasi dan menghasilkan jumlah prediksi 38 (*True Positive*), 127 (*True Negative*), 20 (*False Positive*), dan 31 (*False Negative*)

   Tabel 7. *Confusion Matrix* Tanpa Normalisasi

		Prediksi	
		Diabetes	Tidak Diabetes
Aktual	Diabetes	35	47
	Tidak Diabetes	7	127

   Substitusi nilai tersebut ke persamaan [16], untuk mengetahui nilai *akurasi*, *presisi*, *recall*, dan *f1-score*. Berikut perhitungannya.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} = \frac{35 + 127}{35 + 127 + 47 + 7} = \frac{162}{216} = 0.75$$

$$Precision = \frac{TP}{TP + FP} = \frac{35}{35 + 7} = \frac{35}{42} = 0.83$$

$$Recall = \frac{TP}{TP + FN} = \frac{35}{35 + 47} = \frac{35}{82} = 0.43$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{0.83 \times 0.43}{0.83 + 0.43} = 2 \times \frac{0.3569}{1.26} = \frac{0.71}{1.26} = 0.56$$

2. *Modelling* dan Evaluasi dengan Normalisasi

   Model yang telah dirancang menggunakan algoritma regresi logistik dengan liblinear sebagai solver yang sebelumnya dilakukan data *clean* yaitu normalisasi *Min-Max Scaler* agar nilai-nilai variabel berada pada rentang [0,1] mendapat skor model pada data *training* sebesar 0.793 dan pada data *testing* sebesar 0.764. Kemudian data *testing* dievaluasi dan menghasilkan jumlah prediksi 38 (*True Positive*), 127 (*True Negative*), 20 (*False Positive*), dan 31 (*False Negative*).

   Tabel 8. *Confusion Matrix* dengan Normalisasi

		Prediksi	
		Diabetes	Tidak Diabetes
Aktual	Diabetes	38	31
	Tidak Diabetes	20	127

   Substitusi nilai tersebut ke persamaan [16] untuk mengetahui nilai *akurasi*, *presisi*, *recall*, dan *f1-score*. Berikut perhitungannya.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} = \frac{38 + 127}{38 + 127 + 31 + 20} = \frac{165}{216} = \frac{55}{72} = 0.76$$

$$Precision = \frac{TP}{TP + FP} = \frac{38}{38 + 20} = \frac{38}{58} = 0.66$$

$$Recall = \frac{TP}{TP + FN} = \frac{38}{38 + 31} = \frac{38}{69} = 0.55$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{0.66 \times 0.55}{0.66 + 0.55} = 2 \times \frac{0.363}{1.21} = \frac{0.726}{1.21} = 0.6$$

Berdasarkan hasil *evaluation* maka metrik yang paling cocok digunakan dalam sistem ini adalah *recall*, metrik *recall* digunakan sebagai acuan pemilihan algoritma terbaik untuk model klasifikasi diabetes karena lebih baik terjadi banyak kesalahan prediksi positif diabetes namun sebenarnya tidak diabetes daripada kesalahan prediksi negatif namun sebenarnya positif diabetes atau lebih baik sedikit jumlah *error type II* daripada *type I* di mana semakin besar *error type* semakin membahayakan untuk kasus prediksi diabetes atau tidak [17]. Dengan membandingkan hasil dari dua kasus di atas, terlihat dengan menggunakan normalisasi dapat meningkatkan prediksi sistem yang semula nilai recall 43% (tanpa normalisasi) menjadi 55% (dengan normalisasi).

4. KESIMPULAN

Prediksi risiko diabetes menggunakan algoritma regresi logistik menggunakan liblinear dengan normalisasi menghasilkan *recall* sebesar 55% sedangkan tanpa normalisasi sebesar 43%. Dengan demikian, normalisasi dapat meningkatkan kinerja prediksi risiko diabetes menggunakan algoritma regresi logistik. Model ini diharapkan dapat menjadi acuan untuk pengobatan penderita diabetes bagi dokter di rumah sakit dan di masyarakat untuk mengetahui cara menjaga pola hidup dan cara menghindari penyakit diabetes dilihat dari variabel yang mempengaruhi terjadinya penyakit. Selain itu, disarankan untuk melakukan penelitian tentang prediksi risiko diabetes menggunakan algoritma lain agar mendapatkan kinerja model yang lebih tinggi.

REFERENSI

[1] Y. Safitri and I. K. A. Nurhayati, "Pengaruh Pemberian Sari Pati Bengkuang (Pachyrhizus Erosus) terhadap Kadar Glukosa Darah pada Penderita Diabetes Mellitus Tipe II Usia 40-50 Tahun di Kelurahan Bangkinang Wilayah Kerja Puskesmas Bangkinang Kota Tahun 2018," *J. Ners*, vol. 3, no. 1, pp. 69–81, 2019.

[2] F. Fatmawati, "Perbandingan Algoritma Klasifikasi Data Mining Model C4. 5 dan Naive Bayes untuk Prediksi Penyakit Diabetes," *Techno Nusa Mandiri J. Comput. Inf. Technol.*, vol. 13, no. 1, pp. 50–59, 2016.

[3] U. I. Lestari, A. Y. Nadhiroh, and C. Novia, "Penerapan Metode K-Nearest Neighbor untuk Sistem Pendukung Keputusan Identifikasi Penyakit Diabetes Melitus," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 4, pp. 2071–2082, 2021.

[4] F. Nasution, A. Andilala, and A. A. Siregar, "Faktor Risiko Kejadian Diabetes Mellitus," *J. Ilmu Kesehat.*, vol. 9, no. 2, pp. 94–102, 2021.

[5] R. R. Santoso, "Implementasi Metode Machine Learning menggunakan Algoritma Evolving Artificial Neural Network pada Kasus Prediksi Diagnosis Diabetes." Universitas Pendidikan Indonesia, 2020.

[6] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020.

[7] F. K. Lembang, "Analisis Faktor Resiko Penyebab Diabetes Mellitus di Kota Ambon menggunakan Model Regresi Logistik," *Stat. J. Theor. Stat. Its Appl.*, vol. 15, no. 2, pp. 65–71, 2015.

[8] M. A. Suhendra, D. Ispriyanti, and S. Sudarno, "Ketepatan Klasifikasi Pemberian Kartu Keluarga Sejahtera di Kota Semarang menggunakan Metode Regresi Logistik Biner dan Metode Chaid," *J. Gaussian*, vol. 9, no. 1, pp. 64–74, 2020.

[9] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 6, no. 3, pp. 280–284, 2020.

[10] M. Marna, M. Saftari, P. Jana, and M. Maxrizal, "Analisis Regresi Logistik Biner untuk memprediksi Faktor Internal dan Eksternal terhadap Indeks Prestasi," *Delta J. Ilm. Pendidik. Mat.*, vol. 9, no. 1, pp. 47–56, 2021.

[11] Uci Machine Learning, "Pima Indians Diabetes Database," *kaggle*, 2016. [Online], Tersedia: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database [Diakses: 22 April 2022].

[12] N. A. Kurniawati and S. P. Rahayu, "Analisis Kadar CO, Titania, dan Suhu Terhadap Kelembaban Udara Menggunakan Preprocessing Data, Distribusi Normal Multivariat, Uji Bartlett, dan T2 Hotelling".

[13] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes menggunakan Metode Normalisasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[14] scikit learn, "sklearn.linear_model.LogisticRegression — scikit-learn 1.1.1 documentation," [Online], Tersedia: *scikit learn*.https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegressi [Diakses: 08 Juni 2022].
[15] K. S. Nugroho, "Confusion Matrix untuk Evaluasi Model pada Supervised Learning," *Confusion Matrix untuk Eval. Model pada Supervised Learn.*, 2019.
[16] S. Gargate, "Evaluating your classification model," 13 Desember 2019. [Online]. Tersedia: https://medium.com/swlh/evaluating-your-classification-model-cb49338abb96 [Diakses: 12 Mei 2022].
[17] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning," 05 April 2019, [Online]. Tersedia: https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8 [Diakses: 12 Mei 2022].