Deteksi Dini Penyakit Diabetes Menggunakan *Machine Learning* dengan Algoritma *Logistic Regression*

Erlin[1*], Yulvia Nora Marlim[2], Junadhi[3], Laili Suryati[4], Nova Agustina[5]

***Intisari*—Diabetes menjadi salah satu penyakit yang mematikan di dunia, termasuk di Indonesia. Diabetes dapat menyebabkan komplikasi di banyak bagian tubuh dan secara keseluruhan dapat meningkatkan risiko kematian. Salah satu cara untuk mendeteksi penyakit diabetes adalah dengan memanfaatkan algoritma *machine learning*. *Logistic regression* merupakan model klasifikasi dalam *machine learning* yang banyak digunakan dalam analisis klinis. Pada makalah ini, dirancang model prediksi menggunakan *logistic regression* pada Python IDE untuk deteksi dini dengan memberikan prediksi seseorang terindikasi penyakit diabetes atau tidak berdasarkan data awal yang diberikan. Eksperimen dilakukan menggunakan *dataset* dari Pima Indians Diabetes Database yang terdiri atas 768 data pasien dengan delapan variabel independen dan satu variabel dependen. *Exploratory data analysis* dilakukan untuk mendapatkan wawasan maksimal dari *dataset* yang dimiliki menggunakan bantuan statistik dan mempresentasikannya melalui teknik visual. Beberapa variabel *dataset* memuat data yang tidak lengkap. Nilai data yang hilang digantikan dengan nilai median dari setiap variabel. Penanganan terhadap data yang tidak seimbang dilakukan menggunakan *synthetic minority over-sampling technique* (SMOTE) untuk meningkatkan kelas minoritas melalui sampel data sintesis. Model dievaluasi berdasarkan *confusion matrix* yang memperlihatkan kinerja yang cukup baik dengan nilai akurasi sebesar 77%, presisi 75%, *recall* 77%, dan *F*1-*score* 76%. Selain itu, pada makalah ini juga digunakan teknik *grid search* sebagai *hyperparameter tuning* yang dapat meningkatkan kinerja dari model *logistic regression*. Kinerja model dasar dengan model sesudah dilakukan penerapan teknik *grid search* diuji dan dievaluasi. Hasil percobaan memperlihatkan bahwa model berbasis *hyperparameter tuning* mampu meningkatkan kinerja algoritma *logistic regression* untuk prediksi dengan nilai akurasi sebesar 82%, presisi 81%, *recall* 79%, dan *F*1-*score* 80%.**

***Kata Kunci*—Deteksi Dini, Diabetes, *Machine Learning*, *Logistic Regression*, *Grid Search*.**

[1,2] *Program Studi Teknik Informatika, Institut Bisnis dan Teknologi Pelita Indonesia, Jl. Jend. Ahmad Yani, Pekanbaru, 28127, INDONESIA (e-mail: [1]erlin@lecturer.pelitaindonesia.ac.id; [2]yulvia.nora@lecturer.pelitaindonesia.ac.id)*

[3] *Program Studi Teknik Informatika, STMIK Amik Riau, Jl. Purwodadi Indah, Km. 10 Pekanbaru, 28294, INDONESIA (e-mail: junadhi@sar.ac.id)*

[4] *Program Studi Akuntansi, Universitas Persada Indonesia, Jl. Diponegoro No. 74 Jakarta Pusat, INDONESIA (e-mail: lailisuryati61@gmail.com)*

[5] *Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Bandung, Jl.Soekarno-Hatta No. 378, Kidul Bandung, INDONESIA (e-mail: nova@sttbandung.ac.id)*

[*] *Corresponding Author*

[Diterima: 11 Desember 2021, Revisi: 24 Februari 2022]

I. PENDAHULUAN

Diabetes merupakan penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa (gula darah) yang lebih tinggi dari kadar normal, yang disebabkan oleh gangguan sekresi insulin atau gangguan efek biologis [1]. Diabetes dapat menyebabkan komplikasi di banyak bagian tubuh dan secara keseluruhan dapat meningkatkan risiko kematian dini. Kemungkinan komplikasi yang dapat terjadi termasuk gagal ginjal, amputasi kaki, kehilangan penglihatan, dan kerusakan saraf. Orang dewasa dengan penyakit diabetes juga memiliki dua hingga tiga kali lipat peningkatan risiko serangan jantung dan strok. Pada masa kehamilan, diabetes yang tidak terkontrol dengan baik akan meningkatkan risiko kematian janin dan komplikasi lainnya [2].

Jumlah penderita penyakit diabetes meningkat dari tahun ke tahun, baik dari jumlah kasus maupun prevalensi. Pada tahun 2019, jumlah penderita diabetes di dunia sudah mencapai 463 juta orang dan diprediksi akan terus bertambah mencapai angka 700 juta orang pada tahun 2045. Penderita diabetes mayoritas tinggal di negara berpenghasilan rendah dan menengah dan 1,6 juta kematian secara langsung disebabkan oleh diabetes setiap tahunnya. Hal ini menjadikan diabetes sebagai salah satu dari sepuluh penyakit penyebab utama kematian di seluruh dunia [3].

Pada tahun 2019, Indonesia menempati posisi ke-7 di dunia, setelah China, India, Amerika Serikat, Pakistan, Brazil, dan Meksiko, sebagai negara dengan jumlah penderita diabetes tertinggi, dengan jumlah penderita sebesar 10,7 juta orang. Pada tahun 2020, angka ini meningkat menjadi 10,8 juta dengan angka prevalensi pasien pengidap diabetes mencapai 6,2% dan diperkirakan jumlah penderita diabetes di Indonesia meningkat menjadi 16,7 juta pada tahun 2045 [4]-[7].

Mengingat keterkaitan risiko berkembangnya komplikasi dari penyakit diabetes dan efek kematian yang disebabkan oleh penyakit ini, maka deteksi dini penyakit diabetes menjadi penting untuk dilakukan. Ketika terdeteksi secara dini, pasien tidak hanya dapat menunda, bahkan juga dapat mencegah perkembangan penyakit menjadi diabetes akut. Pencegahan penyakit secara signifikan lebih murah dan mudah daripada pengobatan hiperglikemia dan komplikasi diabetes. Oleh sebab itu, cara identifikasi, diagnosis, dan analisis diabetes secara cepat dan akurat merupakan topik penelitian yang sangat bermanfaat dan penting untuk dilakukan. Dalam bidang kedokteran, diagnosis penyakit diabetes dilakukan berdasarkan kadar gula darah, di antaranya kadar gula darah sewaktu, kadar gula darah puasa, dan kadar toleransi gula darah [8]-[9]. Hasil pengukuran kadar gula darah ini akan menunjukkan seseorang menderita diabetes atau tidak. Semakin dini diagnosis dan deteksi dilakukan, semakin mudah penyakit diabetes dikontrol dan diobati.

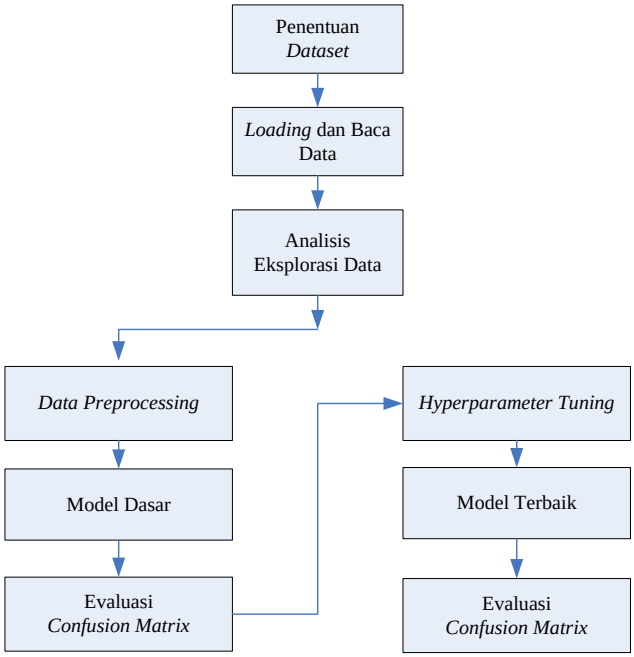
Salah satu cara untuk mendeteksi penyakit diabetes adalah dengan memanfaatkan algoritma *machine learning* [10]-[12]. Algoritma ini sudah banyak digunakan dalam berbagai bidang, termasuk dalam bidang kesehatan [13]-[14]. *Logistic regression* merupakan salah satu algoritma *machine learning* populer yang digunakan untuk masalah klasifikasi dan merupakan algoritma analisis prediktif berdasarkan konsep probabilitas. *Logistic regression* memberikan tingkat akurasi yang lebih baik dibandingkan dengan *k-nearest neighbor* (k-NN) [15], *decision tree* [16], atau model pengklasifikasi lainnya [17].

Beberapa penelitian lain juga memperkuat hasil penelitian sebelumnya mengenai keandalan *logistic regression* dalam memprediksi berbagai macam jenis penyakit. Penggunaan *logistic regression* untuk memprediksi penyakit kardiovaskular mencapai akurasi 85% [18]. *Logistic regression* juga digunakan untuk memprediksi penyakit ginjal kronis dan memperlihatkan hasil bahwa *logistic regression* cenderung memiliki *overfitting* yang lebih rendah dibandingkan dengan *random forest* dan *neural network* [19]. Identifikasi dan prediksi penyakit hati dengan membandingkan algoritma *logistic regression* dengan empat algoritma lainnya juga memperlihatkan bahwa akurasi *logistic regression* lebih baik daripada k-NN, *support vector machine* (SVM), *decision tree*, dan *random forest* [20]. Penelitian lain juga membandingkan tiga model *machine learning*, yaitu *neural network*, *naïve Bayes*, dan *logistic regression* untuk mendeteksi diabetes pada 768 data yang berasal dari data Kaggle. Hasil eksperimen menunjukkan bahwa algoritma *logistic regression* lebih baik dibandingkan dua algoritma lainnya, dengan hasil akurasi 75,78%, dibandingkan *naïve Bayes* dengan akurasi 74,87% dan *neural network* dengan akurasi 69,27% [21]. Evaluasi kinerja *logistic regression* juga dibandingkan dengan algoritma *machine learning* lainnya dan terbukti bahwa *logistic regression* sama baiknya dengan algoritma *neural network* dan SVM dalam memprediksi penyakit kronis seperti penyakit ginjal, jantung, diabetes, dan hipertensi [22].

Berdasarkan beberapa kelebihan dari *logistic regression*, makalah ini menggunakan algoritma *logistic regression* untuk deteksi dini dengan memberikan prediksi seseorang terindikasi penyakit diabetes atau tidak berdasarkan data awal yang diberikan. Penelitian menunjukkan kinerja algoritma *logistic regression* yang sangat baik dalam hal akurasi, presisi, *recall*, dan *F1-score*. Di samping itu, pada penelitian ini juga dilakukan eksperimen menggunakan teknik *grid search*, yaitu sebuah pendekatan yang terdapat dalam model *selection packet* dari scikit-learn yang dapat digunakan untuk meningkatkan kinerja model yang dihasilkan [23]-[24]. *Grid search* akan mengotomatiskan penyetelan *hyperparameter* yang jika dilakukan secara manual akan memakan banyak waktu dan sumber daya [25]. Penggunaan *grid search* pada penelitian ini terbukti mampu meningkatkan kinerja model *logistic regression* dengan nilai akurasi yang lebih baik.

II. METODOLOGI

Algoritma *logistic regression* digunakan untuk memprediksi penyakit diabetes pada sejumlah 768 data yang tersedia.



Gbr. 1 Rancangan penelitian.

TABEL I
VARIABEL YANG DIGUNAKAN DALAM PREDIKSI DIABETES

No.	Variabel	Keterangan
1	*Pregnancies*	Kehamilan: berapa kali pasien hamil
2	*Glucose*	Konsentrasi glukosa plasma selama dua jam dalam tes toleransi glukosa oral
3	*BloodPressure*	Tekanan darah: tekanan darah diastolik (mmHg)
4	*SkinThickness*	Ketebalan lipatan kulit trisep (mm)
5	*Insulin*	Insulin serum dua jam (µlU/mL)
6	BMI	Indeks massa tubuh (kg/m^2)
7	*DiabetesPedigreeFunction*/DPF	Fungsi yang menilai kemungkinan diabetes berdasarkan riwayat keluarga
8	*Age*	Usia di tahun ini
9	*Outcome/Target*	Hasil: variabel kelas (0 jika nondiabetes, 1 jika diabetes)

Prediksi menggunakan *logistic regression* memerlukan serangkaian langkah seperti diperlihatkan pada Gbr. 1, mulai dari penentuan *dataset* sampai *evaluating* dan *deployment* model. Implementasi *logistic regression* ini menggunakan *library* scikit learn dari Python yang memudahkan proses manipulasi, visualisasi, dan analisis data.

A. Penentuan Dataset

Dataset yang digunakan dalam penelitian ini berasal dari National Institute of Diabetes and Digestive and Kidney Diseases sebagai bagian dari Pima Indians Diabetes Database [26]. *Dataset* terdiri atas beberapa variabel prediktor medis (independen) dan satu variabel target (dependen), yaitu Target (hasil), seperti diperlihatkan pada Tabel I.

B. Loading dan Baca Data

Dataset dalam format *.csv* dimuat ke dalam variabel independen. Terdapat 768 data pasien yang semuanya perempuan berusia 21 tahun ke atas, yang terdiri atas sembilan variabel dengan rincian delapan variabel independen, yaitu *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin,* BMI, *DiabetesPedigreeFunction*/DPF, dan *Age*; dan satu variabel dependen, yaitu Target. Hasil pemeriksaan pada *dataset* menggunakan *data.head()* memperlihatkan adanya beberapa variabel bernilai 0 yang mengindikasikan nilai yang hilang.

C. Analisis Eksplorasi Data

Analisis eksplorasi data bertujuan menganalisis *dataset* yang digunakan untuk meringkas karakteristik utama *dataset* tersebut menggunakan bantuan statistika dan mempresentasikannya melalui teknik visual. Pada tahap ini, data diperiksa sebelum dibangunnya model, sehingga didapatkan wawasan maksimal dari *dataset* yang dimiliki.

D. Data Preprocessing

Pada tahap ini, pengecekan dilakukan terhadap nilai data yang hilang karena *dataset* bisa saja memuat data yang tidak lengkap. Nilai data yang hilang digantikan dengan nilai median dari setiap variabel, sehingga setiap data pada variabel *dataset* memiliki nilai yang lengkap. Pada tahap ini juga dilakukan pengecekan terhadap data yang tidak seimbang. Penanganan terhadap data yang tidak seimbang dilakukan menggunakan *synthetic minority over-sampling technique* (SMOTE). Teknik ini digunakan untuk meningkatkan jumlah kelas minoritas melalui sampel data sintesis dengan tetap mempertahankan jumlah kelas mayoritas.

E. Membangun Model Menggunakan Algoritma Logistic Regression

Logistic regression memodelkan hubungan antara variabel respons kategori dan *covariate*. Secara khusus, ada kombinasi linier dari variabel independen dengan log-peluang probabilitas suatu peristiwa. *Logistic regression* merupakan model linier yang lebih cocok untuk masalah klasifikasi dibandingkan penggunaannya untuk regresi. *Logistic regression* juga dikenal dalam literatur sebagai regresi *logit*, klasifikasi entropi maksimum (*MaxEnt*), atau pengklasifikasi log-linear. Dalam *logit*, probabilitas yang menggambarkan kemungkinan hasil dari percobaan tunggal dimodelkan menggunakan fungsi logistik.

Model *logistic regression* dapat berupa *logistic regression* biner, *one-vs-rest*, atau *multinomial logistic regression* dengan l_1, l_2 atau regulasi *elastic-net* [27]. *Logistic regression* biner memperkirakan kemungkinan bahwa karakteristik variabel biner tersedia, dengan mengingat nilai *covariate*. Misalnya, *Y* adalah variabel respons biner dengan $Y_i = 1$ jika karakter ada, dan $Y_i = 0$ jika karakter tidak ada dan data $[Y_1, Y_2, \dots, Y_n]$ independen. Nilai π_i dapat digunakan untuk menjadi peluang sukses sebuah *logistic regression*. Selain itu, dipertimbangkan juga nilai $x = (x_1, x_2, \dots, x_p)$ sebagai satu set variabel yang dapat berbentuk diskret, kontinu, atau kombinasi keduanya. Fungsi logistik untuk π_i diberikan oleh (1) dan (2).

TABEL II
CONFUSION MATRIX UNTUK KLASIFIKASI BINER

Kelas Aktual	Kelas Prediksi	
	Positif	***Negatif***
Positif	tp	fp
Negatif	fn	tn

$$logit\,(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad (1)$$

dengan

$$\pi_i = \frac{exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}{1 + exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}} = \frac{exp\,(x_i'\beta)}{1 + exp\,(x_i'\beta)} = \Lambda(x_i'\beta). \quad (2)$$

Di sini, π_i menunjukkan probabilitas bahwa sampel berada dalam kategori tertentu dari variabel respons dikotomis, yang biasa disebut sebagai probabilitas sukses dan sangat jelas bahwa $0 \le \pi_i \le 1$. $\Lambda(.)$ adalah *cumulative distribution function* (cdf) logistik dengan $\lambda(z) = \frac{e^z}{(1+e^z)} = 1/(1+e^z)$ serta β^s mewakili vektor parameter yang akan diestimasi. Persamaan $\frac{\pi_i}{1-\pi_i}$ disebut rasio *odd* atau risiko relatif [28].

Penelitian ini menggunakan algoritma *logistic regression* biner karena keluarannya berupa nilai 0 dan 1, yang digunakan untuk mendeteksi seseorang menderita penyakit diabetes atau tidak. Nilai keluaran 0 berarti seseorang tidak menderita diabetes dan sebaliknya, keluaran 1 berarti seseorang tersebut merupakan pasien yang mengidap penyakit diabetes.

F. Evaluasi Model Menggunakan Confusion Matrix

Evaluasi yang digunakan untuk mengukur kinerja algoritma/model adalah akurasi, presisi, *recall*, dan *F*1-*score* dalam bentuk *confusion matrix* yang sudah banyak digunakan oleh peneliti lainnya. *Confusion matrix* merupakan tabel yang bekerja dengan cara membandingkan jumlah prediksi yang benar dan yang salah yang terdapat pada masing-masing kelas, sehingga memberikan wawasan mengenai kesalahan model yang dibangun. Tabel II adalah *confusion matrix* dengan ukuran 2 × 2 yang digunakan untuk mempresentasikan kelas aktual dan kelas prediksi. Beberapa pengukuran dalam bidang *information retrieval* dan *machine learning* sudah diidentifikasi berdasarkan klasifikasi dari *confusion matrix* sebagaimana terdapat dalam (3) sampai (6).

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (3)$$

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

$$Recall/Sensitivity = \frac{tp}{tp+fn} \quad (5)$$

$$F1\ score = \frac{2\,(recall.precision)}{(recall+precision)} \quad (6)$$

Dalam (3) hingga (6), *tp* (*true positive*) adalah jumlah pasien yang diprediksi menderita diabetes dan memang benar

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Target                    768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Target
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

Gbr. 2 Informasi umum *dataset*.

penderita diabetes; *tn* (*true negative*) adalah orang yang diprediksi nondiabetes dan memang benar nondiabetes; *fp* (*false positive*) adalah orang yang diprediksi nondiabetes, tetapi orang tersebut adalah penderita diabetes; dan *fn* (*false negative*) adalah orang yang diprediksi menderita diabetes, tetapi orang tersebut bukan penderita diabetes.

G. Hyperparameter Tuning

Hyperparameter tuning digunakan untuk memilih set *hyperparameter* yang optimal untuk meningkatkan kinerja model. *Hyperparameter* berbentuk argumen model yang nilainya ditetapkan sebelum proses *learning* dimulai. Penyetelan *hyperparameter* merupakan kunci sukses dari sebuah model/algoritma. Pada makalah ini, digunakan teknik *grid search* yang mengombinasikan nilai masukan pada *hyperparameter*. Teknik *grid search* akan mencari seluruh kombinasi yang mungkin dan memilih kombinasi terbaik berdasarkan nilai *cross validation* yang paling tinggi. Dua *hyperparameter* diaplikasikan pada penelitian ini, yaitu *penalty* dan nilai C. *Penalty* menggunakan regulasi l1 dan l2 (nilai *default*-nya adalah l2), sedangkan nilai C merupakan kebalikan dari kekuatan regularisasi.

III. Hasil dan Pembahasan

A. Loading dan Baca Data

Gbr. 2 memperlihatkan hasil *loading* dan pembacaan terhadap *dataset* yang terdiri atas 768 data pasien. Tujuh data bertipe integer, yaitu *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *Age*, dan Target; serta dua data bertipe *float*, yaitu BMI dan *DiabetesPedigreeFunction*. Data terdiri atas sembilan variabel, yaitu delapan variabel independen dan satu variabel dependen.

Pada gambar juga terlihat ada nilai data yang kosong pada beberapa variabel. Dari tampilan lima data teratas ini, terlihat bahwa variabel *Pregnancies* dan *Insulin* memiliki nilai data yang kosong (0). Nilai data yang kosong akan diganti dengan nilai median dari setiap variabel untuk memudahkan proses manipulasi data.

B. Analisis Eksplorasi Data

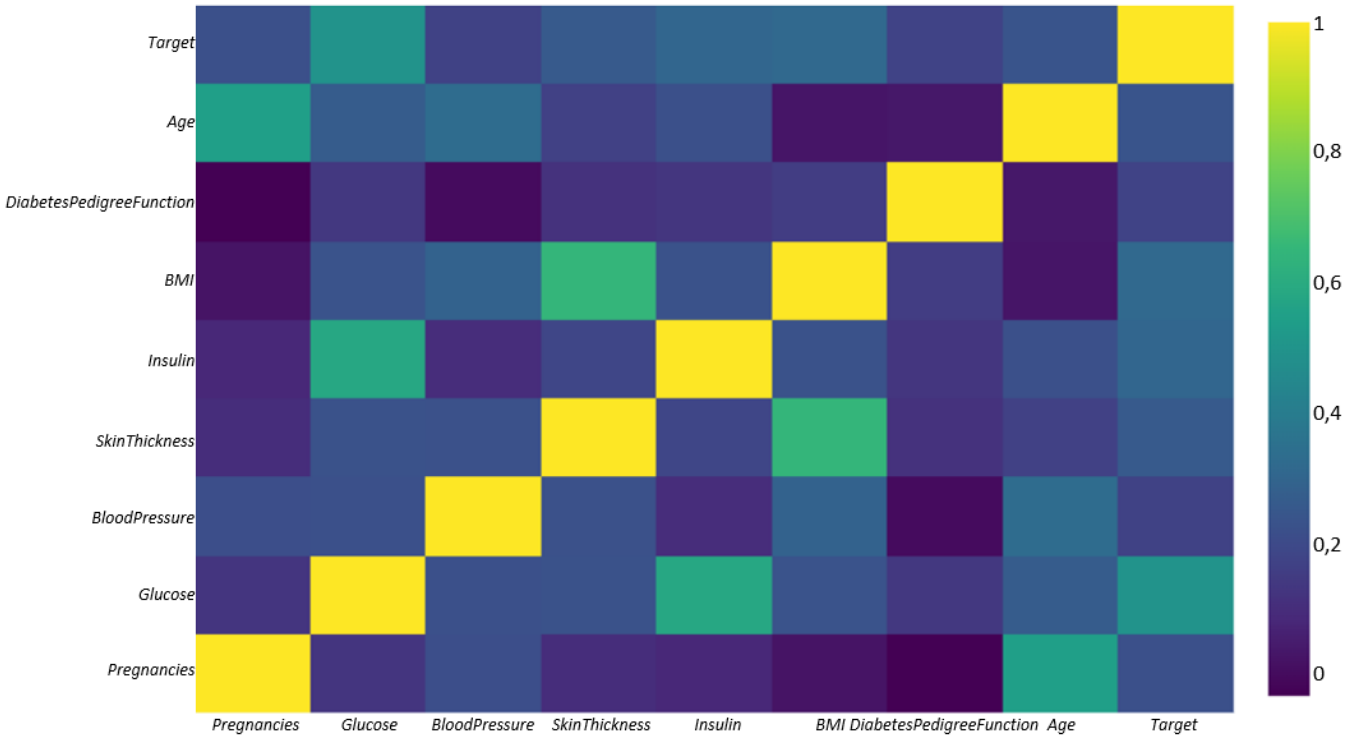
Memahami *dataset* yang ada merupakan hal penting sebelum melakukan proses selanjutnya. Gbr. 3 merupakan matriks korelasi yang menunjukkan koefisien korelasi antar himpunan variabel. Setiap variabel independen (X_i) dalam tabel berkorelasi dengan masing-masing nilai lain dalam tabel (X_j). Matriks korelasi diperlukan untuk melihat pasangan variabel yang memiliki korelasi tertinggi. Terdapat korelasi yang cukup kuat antara variabel *Insulin*, *Glucose*, BMI, *Pregnancies*, dan *Age* dengan variabel Target.

Distribusi *plot* untuk masing-masing atribut/variabel diperlihatkan pada Gbr. 4. Terlihat bahwa kolom *Age* dan *Insulin* sangat miring ke kanan, sehingga diperlukan proses normalisasi sebelum digunakan untuk proses pemodelan.

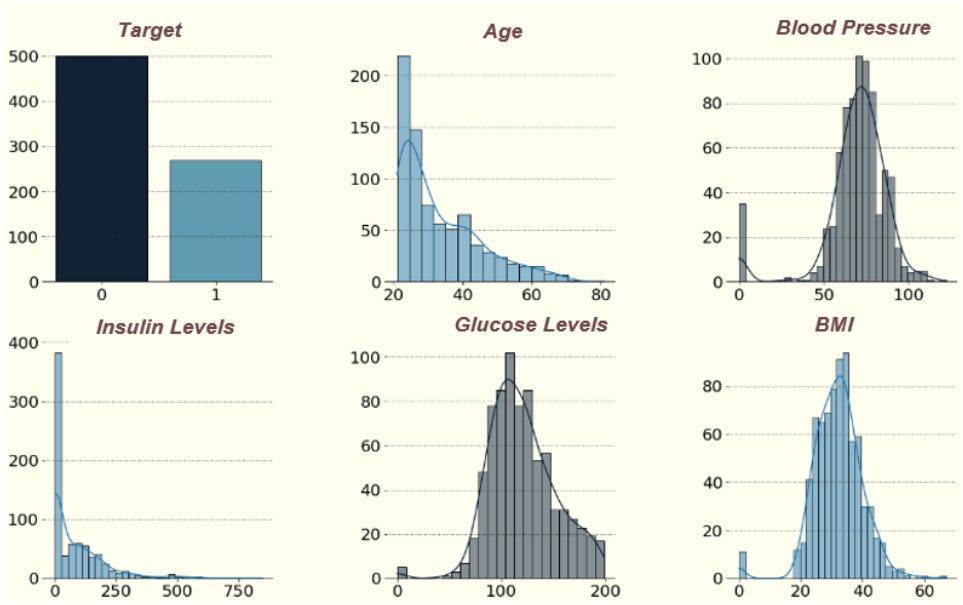
Dataset memperlihatkan bahwa sejumlah besar orang memiliki usia antara 20-40 tahun, sebagian besar orang memiliki tekanan darah antara 50-100 mmHg dan memiliki insulin 0, serta sebagian besar orang juga memiliki kadar glukosa antara 140 mg/dL sampai 199 mg/dL dan dianggap sebagai penderita pradiabetes. Nilai BMI berkisar di antara 20 sampai 50, sementara untuk orang dewasa yang sehat harus memiliki BMI antara 18,5-24,9. *Dataset* ini jelas memperlihatkan banyak orang yang kelebihan berat badan atau obesitas.

C. Data Preprocessing

1) Missing Value: Hasil pengecekan *dataset* menunjukkan sejumlah nilai data yang hilang, seperti diperlihatkan pada Gbr. 5(a). Variabel *Insulin* merupakan variabel yang memiliki nilai



Gbr. 3 Matriks korelasi.

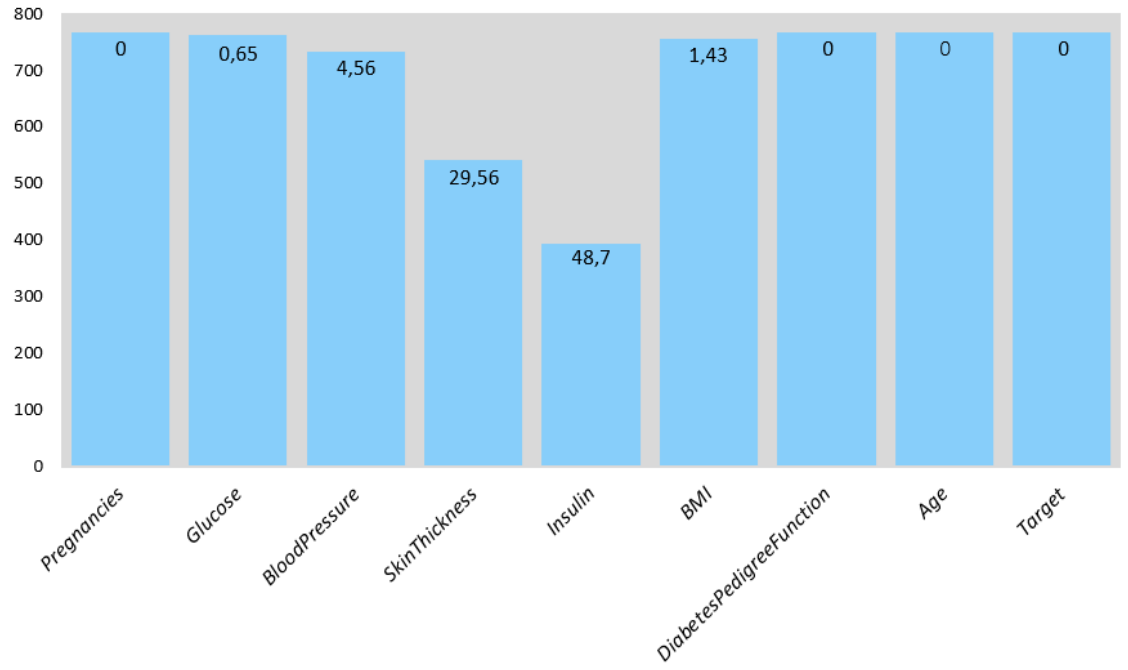


Gbr. 4 Distribusi *plot* untuk masing-masing variabel.

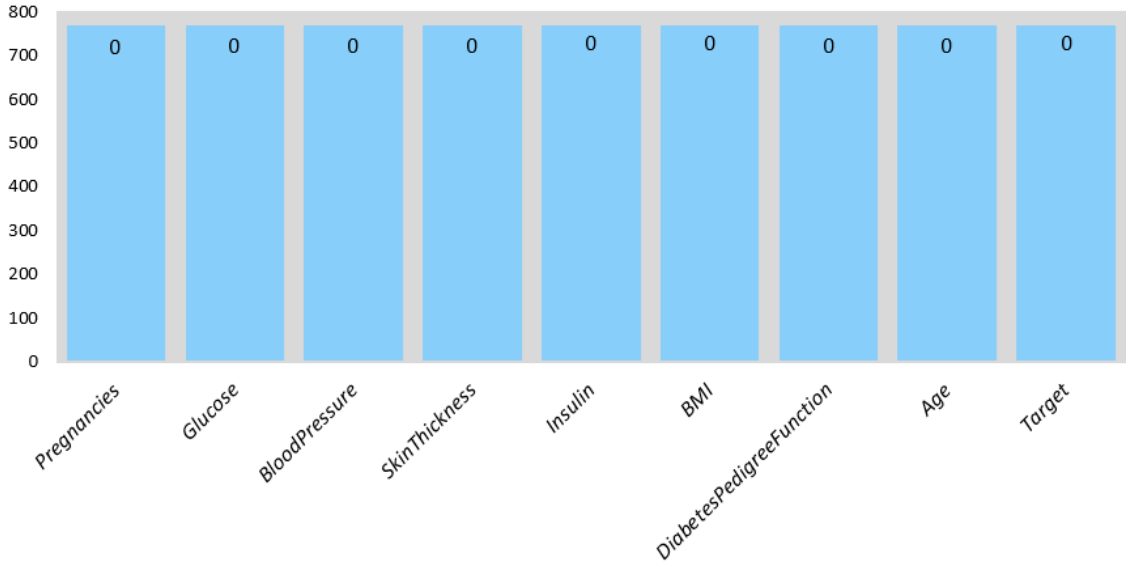
data hilang terbanyak, yaitu sebesar 374 data atau 48,7%, disusul oleh *SkinThickness* sebanyak 227 data atau 29,56%, *BloodPressure* sebanyak 35 data atau 4,56%, BMI sebanyak 11 data atau 1,43%, dan *Glucose* sebanyak 5 data atau 0,65%.

Variabel lainnya memiliki data yang lengkap. Semua nilai data yang hilang akan diganti dengan nilai median dari setiap variabel, sehingga tidak ada lagi nilai variabel yang kosong, seperti diperlihatkan pada Gbr. 5(b). Penanganan nilai data yang hilang ini digunakan untuk memudahkan proses manipulasi data.

Setelah semua variabel/atribut lengkap, dibuat atribut baru yang merupakan kombinasi dari beberapa variabel untuk melihat pengelompokan, seseorang menderita penyakit diabetes atau tidak, berdasarkan keterhubungan antara satu variabel dengan variabel lainnya. Gbr. 6 memperlihatkan salah satu atribut baru yang merupakan hubungan antara variabel *Age* dengan variabel *Glucose*, yaitu orang yang nondiabetes berusia di bawah 30 tahun dengan kadar glukosa di bawah 120 mg/dL. Orang sehat berada pada daerah yang terkonsentrasi pada usia ≤ 30 dan glukosa ≤ 120 mg/dL.



(a)



(b)

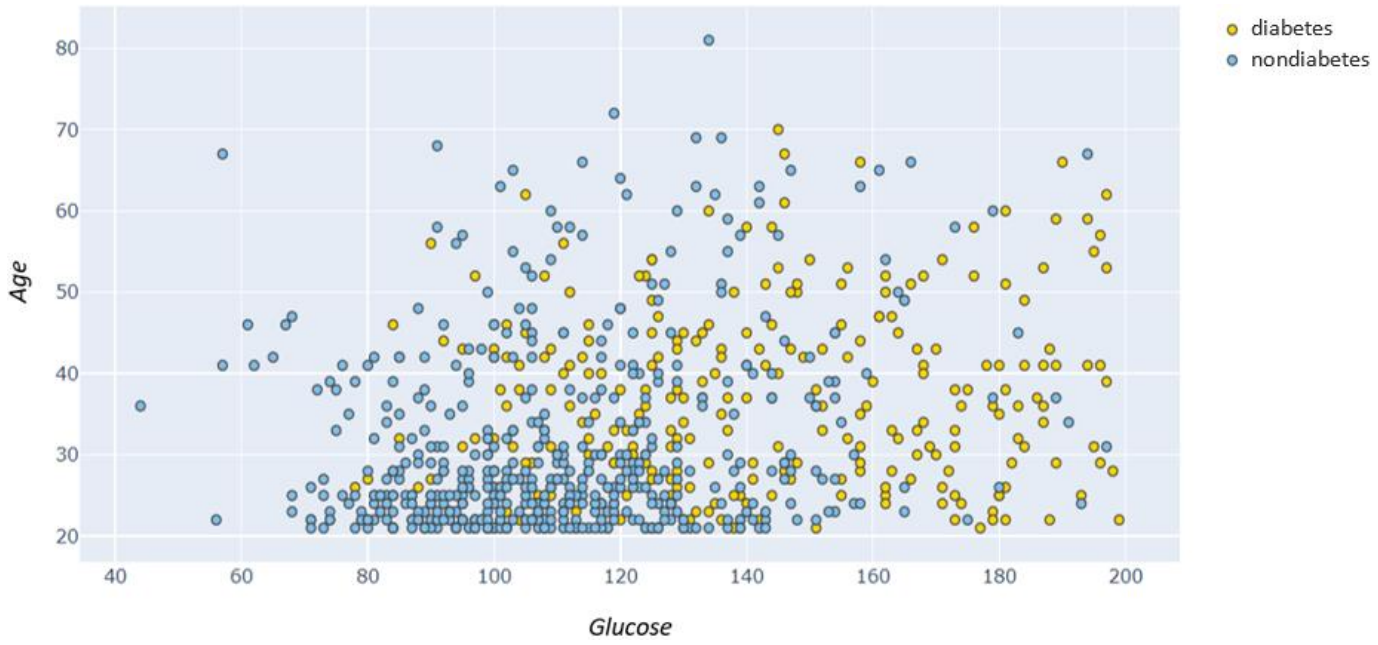
Gbr. 5 Jumlah data yang hilang pada *dataset*, (a) sebelum normalisasi, (b) sesudah normalisasi.

2) Data Tidak Seimbang (Imbalanced Dataset): Grafik hasil pengecekan terhadap data tidak seimbang diperlihatkan pada Gbr. 7. Jumlah pasien diabetes sebanyak 268 orang (34,9%), sedangkan nondiabetes sebanyak 500 orang (65,1%). Untuk mengatasi data yang tidak seimbang ini, digunakan teknik SMOTE. SMOTE adalah teknik *oversampling*, yaitu sampel sintetis dihasilkan untuk kelas minoritas untuk membantu mengatasi masalah *overfitting* yang terdapat pada *random oversampling*.

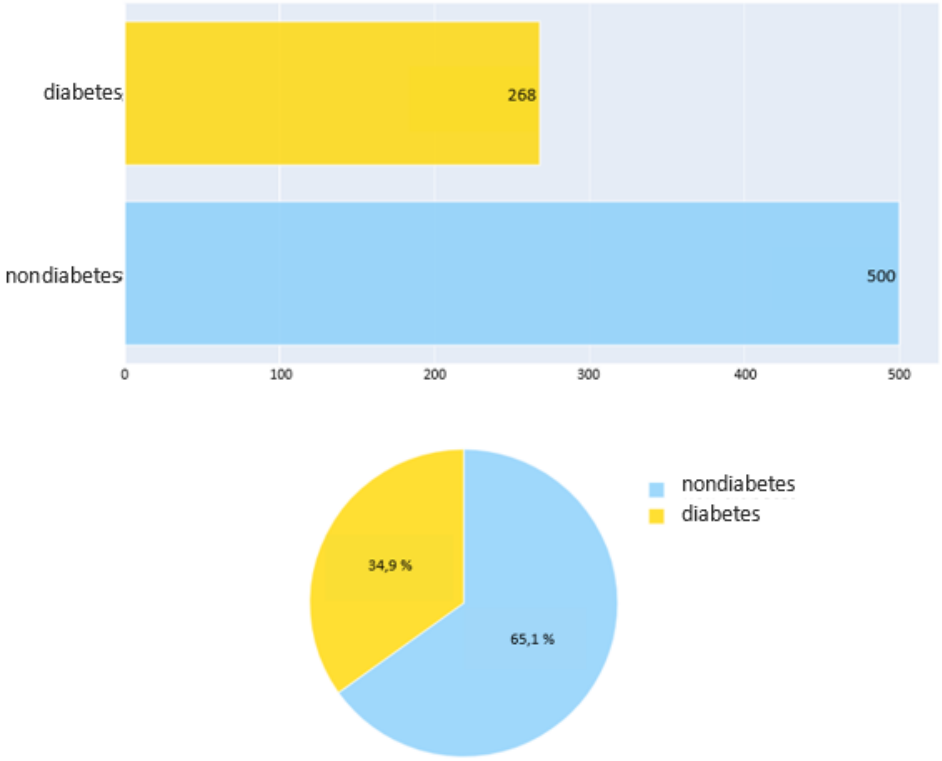
Setelah tahap *preprocessing*, langkah berikutnya adalah pemodelan menggunakan *logistic regression*. Data terlebih dahulu dibagi menjadi dua bagian, yaitu data latih dan data uji. Perbandingan data latih dan data uji adalah 70:30. Proses ini dilakukan menggunakan *library* scikit-learn dari Python. Cuplikan skrip diperlihatkan pada Gbr. 8, yang menunjukkan pembagian *dataset* untuk data latih dan data uji, lalu dilanjutkan dengan pemodelan menggunakan algoritma *logistic regression.*

D. Evaluasi Model

Evaluasi model dilakukan setelah model terbentuk. Tabel III memperlihatkan hasil evaluasi dari model dalam format *confusion matrix*. Terdapat 178 data berada dalam klasifikasi yang tepat (*true positive* dan *true negative*), yang terdiri atas



Gbr. 6 *Scatter plot* hubungan variabel umur (*Age*) dan variabel glukosa (*Glucose*).



Gbr. 7 Jumlah target *dataset* yang tidak seimbang.

116 data diprediksi diabetes dan aktualnya memang penderita diabetes serta 62 orang diprediksi nondiabetes dan aktualnya memang bukan penderita diabetes. Sebanyak 53 data lain merupakan *false positive* dan *false negative*, yaitu 35 orang nondiabetes diprediksi menderita diabetes dan delapan belas orang penderita diabetes diprediksi sebagai nondiabetes.

Tabel IV memperlihatkan pengukuran kinerja model berdasarkan nilai akurasi, presisi, *recall*, dan *F*1-*score*. Nilai presisi yang diperoleh untuk kelas nondiabetes adalah sebesar 87% dan untuk diabetes sebesar 64%, dengan rata-rata nilai presisi sebesar 75%. Nilai *recall* untuk nondiabetes adalah 77% dan untuk diabetes sebesar 78%, dengan nilai rata-rata *recall* sebesar 77%. Nilai *F*1-*score* menunjukkan angka yang tidak jauh berbeda dengan presisi dan *recall*. Nilai *F*1-*score* untuk nondiabetes sebesar 81% dan untuk diabetes sebesar 70%, dengan rata-rata 76%. Nilai akurasi adalah sebesar 77%.

```
# Membagi Dataset menjadi Dataset Training dan Dataset Testing
from sklearn.model_selection import train_test_split
X = df.drop('Target', axis=1)
y = df['Target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
```

```
logmodel = LogisticRegression(max_iter=200)
logmodel.fit(X_train, y_train)
prediction1 = logmodel.predict(X_test)
```

Gbr. 8 Cuplikan program Python untuk membangun model.

TABEL III
CONFUSION MATRIX HASIL KLASIFIKASI

Kelas Aktual	**Kelas Prediksi**	
	0 (Nondiabetes)	***1 (Diabetes)***
0 (Nondiabetes)	116	35
1 (Diabetes)	18	62

TABEL IV
PENGUKURAN KINERJA

Presisi (%)		***Recall* (%)**		***F1-Score* (%)**	
0	1	0	1	0	1
87	64	77	78	81	70
Rata-rata: 75		Rata-rata: 77		Rata-rata: 76	

Berdasarkan perhitungan nilai dari keempat variabel yang dievaluasi, model dikategorikan memiliki kinerja yang cukup baik.

E. Hyperparameter Tuning Menggunakan Teknik Grid Search

Untuk meningkatkan kinerja model, dilakukan *tuning* terhadap parameter yang digunakan. Gbr. 9 merupakan cuplikan skrip Python yang digunakan untuk memilih parameter terbaik melalui teknik *grid search*. L2 terpilih sebagai regulasi *penalty* terbaik dengan nilai C paling optimal adalah 0,6158.

F. Evaluasi Kinerja Model Sebelum dan Sesudah Hyperparameter Tuning

Perbandingan kinerja model sebelum dan sesudah penerapan teknik *grid search* dianalisis untuk mengukur dampak dari penggunaan *hyperparameter tuning*. Gbr. 10(a) menunjukkan kinerja model sebelum penerapan *hyperparameter tuning* dan Gbr. 10(b) menunjukkan kinerja model sesudah penerapan *hyperparameter tuning*. Dari kedua gambar terlihat jelas peningkatan kinerja pada semua nilai *confusion matrix*, mulai dari nilai akurasi, presisi, *recall*, hingga *F*1-*score*. Pada model dasar, nilai rata-rata akurasi adalah 77%, presisi 75%, *recall* 77%, dan *F*1-*score* 76%; sedangkan pada model yang ditingkatkan, nilai akurasi naik menjadi 82%, presisi 81%, *recall* 79%, dan *F*1-*score* 80%.

Gbr. 11 memperlihatkan bahwa model *logistic regression* menggunakan *grid search* lebih baik dibandingkan dengan model dasar untuk seluruh nilai *confusion matrix* (akurasi, *precison*, *recall*, dan *F*1-*score*). Eksperimen ini membuktikan bahwa algoritma *linier regression* dapat digunakan untuk prediksi dalam bidang klinis atau kesehatan dengan tingkat akurasi yang baik. Kinerja algoritma *logistic regression* lebih meningkat apabila disisipi dengan teknik *grid search* yang mampu menaikkan semua nilai *confusion matrix*, sehingga memberikan dampak meningkatnya kinerja algoritma *logistic regression* secara keseluruhan.

```
#Nilai hyperparameter terbaik
print('Best Penalty:', best_model.best_estimator_.get_params()['penalty'])
print('Best C:', best_model.best_estimator_.get_params()['C'])
```

```
Best Penalty: l2
Best C: 0.615848211066026
```

Gbr. 9 Pemilihan parameter terbaik.

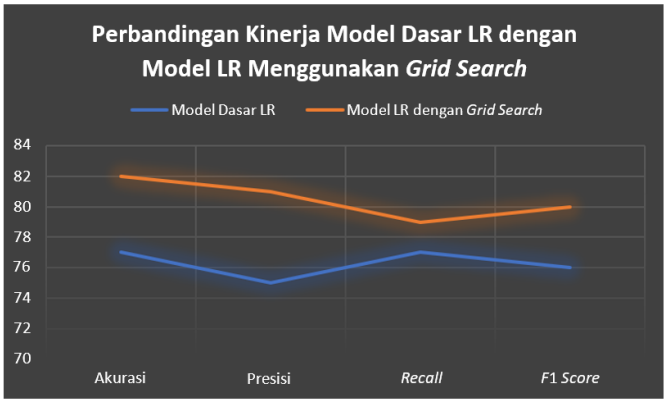
	precision	recall	f1-score	support
0	0.87	0.77	0.81	151
1	0.64	0.78	0.70	80
accuracy			0.77	231
macro avg	0.75	0.77	0.76	231
weighted avg	0.79	0.77	0.77	231

(a)

	precision	recall	f1-score	support
0	0.84	0.90	0.87	152
1	0.78	0.67	0.72	79
accuracy			0.82	231
macro avg	0.81	0.79	0.80	231
weighted avg	0.82	0.82	0.82	231

(b)

Gbr. 10 Kinerja model, (a) sebelum *hyperparameter tuning*, (b) sesudah *hyperparameter tuning*.



Gbr. 11 Perbandingan kinerja model dasar dengan model *hyperparameter tuning*.

IV. KESIMPULAN

Penelitian ini sukses mengimplementasikan algoritma *logistic regression* dalam memprediksi penyakit diabetes dengan tingkat akurasi yang baik. Pemahaman terhadap data dilakukan melalui proses eksplorasi data dan penelitian mampu menganalisis pasangan variabel yang memiliki korelasi yang cukup kuat terhadap penentuan nilai target melalui teknik visualisasi dalam bentuk distribusi dan *scatter plot*. Kinerja model dasar dari algoritma *logistic regression* ditingkatkan

menggunakan teknik *grid search*. Evaluasi model menggunakan *confusion matrix* memperlihatkan terjadinya peningkatan kinerja model sesudah penerapan *hyperparameter tuning*. Oleh sebab itu, hasil eksperimen penelitian ini membuktikan bahwa algoritma *logistic regression* dengan teknik *grid search* menjadi salah satu algoritma yang efisien dalam membangun model prediksi. Penelitian selanjutnya akan menggunakan algoritma *deep learning* pada *dataset* yang lebih besar, termasuk mencoba mengombinasikan algoritma *logistic regression* dengan algoritma klasifikasi lainnya, seperti *random forest, support vector machine, k-nearest neighbor* dengan teknik *ensemble.*

KONFLIK KEPENTINGAN

Tim penulis yang namanya tertera dalam artikel yang berjudul "Deteksi Dini Penyakit Diabetes Menggunakan *Machine Learning* dengan Algoritma *Logistic Regression*" ini menyatakan bahwa tidak terdapat konflik kepentingan dalam penelitian ini.

KONTRIBUSI PENULIS

Konseptualisasi, Erlin dan Yulvia Nora; metodologi penelitian, Erlin, Junadhi, Laili Suryati; perangkat lunak, Erlin dan Nova Agustina; validasi, Erlin, Yulvia Nora, Junadhi, Laili Suryati, Nova Agustina; analisis formal, Erlin; sumber, Erlin; tulisan—persiapan draf asli, Erlin; penulisan—review dan penyuntingan, Erlin.

REFERENSI

[1] American Diabetes Association (2020) "Diabetes Overview The path to understanding diabetes starts here." [Online], https://www.diabetes.org/diabetes, tanggal akses: 19-Nov-2021.

[2] World Health Organization (2020) "Diabetes," [Online], https://www.who.int/health-topics/diabetes#tab=tab_1, tanggal akses: 19-Nov-2021.

[3] International Diabetes Federation (2020 "Diabetes facts & figures," [Online], https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html, tanggal akses: 19-Nov-2021.

[4] J. Elflein (2019) "Number of people with diabetes, by country 2019," [Online], https://www.statista.com/statistics/281082/countries-with-highest-number-of-diabetics/, tanggal akses: 6-Des-2021.

[5] H. Nurhayati-Wolff (2020) "Projected number of people with diabetes Indonesia 2017-2024," [Online], https://www.statista.com/statistics/1052625/indonesia-diabetes-projection/, tanggal akses: 6-Des-2021.

[6] B. Hardhana, F. Sibuea, dan W. Widiantini, Eds., *Profil Kesehatan Indonesia Tahun 2019*, Jakarta, Indonesia: Kementerian Kesehatan Republik Indonesia, 2020..

[7] Badan Litbangkes Kemenkes RI (2018), "Hasil Utama Riskesdas 2018," [Online], https://drive.google.com/file/d/1MRXC4lMDera5949ezbbHj7UCUj5_EQmY/view, tanggal akses: 6-Des-2021.

[8] Diabetes UK (2018) "Diabetes the Basics," [Online], https://www.diabetes.org.uk/diabetes-the-basics, tanggal akses: 8-Des-2021.

[9] M.C. Riddle, Ed., "Standards of Medical Care in Diabetes—2022," *Diabetes Care*, Vol. 45, Supp. 1, hal. 125-143, Jan. 2022.

[10] D.J. Reddy, dkk., "Predictive Machine Learning Model for Early Detection and Analysis of Diabetes," *Mater. Today: Proc.*, akan diterbitkan.

[11] L.V.R. Kumari, dkk., "Machine Learning based Diabetes Detection," *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES 2021)*, 2021, hal. 1-5.

[12] N. Abdulhadi dan A. Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods," *Proc. 2021 Int. Conf. Inf. Technol. ICIT 2021*, 2021, hal. 350–354.

[13] R. Krishnamoorthi, dkk., "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," *J. Healthc. Eng.*, Vol. 2022, hal. 1–10, 2022.

[14] U.M. Butt, dkk., "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J. Healthc. Eng.*, Vol. 2021, hal. 1–17, 2021.

[15] P. Arsi dan O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *J. Inform. J. Pengemb. IT*, Vol. 3, No. 3, hal. 290–294, 2018.

[16] A.B. Wibisono dan A. Fahrurozi, "Perbandingan Algoritma Klasifikasi dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. dan Rekayasa*, Vol. 24, No. 3, hal. 161–170, 2019.

[17] J.J. Khanam dan S.Y. Foo, "A Comparison of Machine Learning Algorithms for Diabetes Prediction," *ICT Express*, Vol. 7, No. 4, hal. 432–439, 2021.

[18] T. Ciu dan R.S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (Int. J. New Media Technol.)*, Vol. 7, No. 1, hal. 33–38, 2020.

[19] R. Thammasudjarit, dkk., "Comparison of Machine Learning with Logistic Regression for Prediction of Chronic Kidney Disease in the Thai Adult Population," *Ramathibodi Med. J.*, Vol. 44, No. 4, hal. 1–12, 2021.

[20] N. Varshney dan A. Sharma, "Identification and Prediction of Liver Disease Using Logistic Regression," *Eur. J. Mol. Clin. Med.*, Vol. 7, No. 4, hal. 106–110, 2020.

[21] D.Y. Utami, E. Nurlelah, dan F.N. Hasan, "Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to Find the Highest Accuracy in Diabetes," *J. Inform. Telecommun. Eng.*, Vol. 5, No. 1, hal. 152–159, 2021.

[22] S. Nusinovici, dkk., "Logistic Regression was As Good As Machine Learning for Predicting Major Chronic Diseases," *J. Clin. Epidemiol.*, Vol. 122, hal. 56–69, 2020.

[23] S. Mezzatesta, dkk., "A Machine Learning-based Approach for Predicting the Outbreak of Cardiovascular Diseases in Patients on Dialysis," *Comput. Methods, Programs Biomed.*, Vol. 177, hal. 9–15, 2019.

[24] S. Ambesange, dkk., "Multiple Heart Diseases Prediction Using Logistic Regression with Ensemble and Hyper Parameter Tuning Techniques," *Proc. World Conf. Smart Trends Syst. Secur. Sustain. WS4 2020*, 2020, hal. 827–832.

[25] L. Lama, dkk., "Machine Learning for Prediction of Diabetes Risk in Middle-aged Swedish People," *Heliyon*, Vol. 7, No. 7, hal. 1–6, 2021.

[26] (2016) "Pima Indians Diabetes Database," [Online], https://www.kaggle.com/uciml/pima-indians-diabetes-database, tanggal akses: 23-Okt-2021.

[27] F. Pedregosa, dkk., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, Vol. 12, No. 85, hal. 2825–2830, 2011.

[28] R.D. Joshi dan C.K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *Int. J. Environ. Res. Public Health*, Vol. 18, No. 14, hal. 1-17, 2021.