Analisis Implementasi Seleksi Fitur Pada Klasifikasi Diabetes dengan Metode Corellation Matrix dan Algoritma Logistic Regression

Fitri Kurniawati [1], Dede Brahma Arianto[2]
Sistem Informasi Universitas Pembangunan Nasional Veteran Jakarta [1]
Magister Informatika Universitas Islam Indonesia[2]
2010512067@mahasiswa.upnvj.ac.id [1], dede.brahma2@gmail.com [2]

Abstrak. Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula darah. Diabetes merupakan salah satu penyakit penyumbang kenaikan angka peluang kematian dari tahun ketahun terhitung sejak 2000 - 2019. Penting untuk dilakukan deteksi dini dan pola hidup sehat sebagai langkah pencegahan diabetes. Tujuan penelitian ini untuk membandingkan performance dari algoritma Logistic Regression untuk prediksi diabetes dengan seleksi fitur dan tanpa seleksi fitur untuk mengetahui apakah seleksi fitur dapat meningkatkan performance model untuk prediksi diabetes. Metode yang digunakan Logistic Regression yang diuji dengan 3 skenario, 1 skenario tanpa seleksi fitur dan 2 skenario lainnya menggunakan seleksi fitur dengan *tools corellation matrix* dengan visualisasi *heatmap*. Dari penelitian ini didapatkan skenario 1 yang menggunakan algoritma Logistic Regression tanpa seleksi fitur menghasilkan performance terbaik dengan presisi 77%, akurasi 79,1%, *recall* 74% dan *f1-score* 75%. Sehingga dapat disimpulkan bahwa prediksi menggunakan model Logistic Regression tanpa seleksi fitur memiliki *performance* yang lebih unggul untuk prediksi diabetes.

Kata Kunci: Logistic Regression, Seleksi Fitur, Correlation Matrix, Heatmap

1 Pendahuluan

Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula darah. Tigginya kadar gula darah disebabkan oleh pankreas yang kurang dapat bekerja maksimal untuk menghasilkan insulin ataupun ketika tubuh tidak dapat secara maksimal menggunakan insulin yang dihasilkan. Insulin adalah hormon yang mengatur kadar glukosa pada darah. Tingginya gula darah yang tidak wajar dapat merusak sistem saraf, pembuluh darah, bahkan dapat menyebabkan buta mata permanen. Diabetes merupakan salah satu penyakit yang menyumbang angka kematian cukup tinggi. Menurut WHO pada tahun 2019, diabetes merupakan penyebab utama dari 48% kematian pada umur dibawah 70 tahun. Sebanyak 460.000 kasus kematian karena penyakit ginjal dan 20% kematian karena kardiovaskular disebabkan oleh diabetes. Melihat dari kasus-kasus tersebut, penyakit diabetes perlu dengan serius untuk diperhatikan dan dilakukan pencegahan. Pencegahan dapat dilakukan dengan mengubah pola hidup yang sehat dan tentunya adalah diagnosis dini[1].

Diabetes dikategorikan menjadi 3 tipe yaitu diabetes tipe 1, diabetes tipe 2, dan diabetes gestational. Diabetes tipe 1 disebabkan karena defisiensi insulin, sementara diabetes tipe 2 disebabkan karena tubuh penderita tidak dapat memproses insulin menjadi energi dengan baik yang mengakibatkan naiknya kadar gula darah, sedangkan diabetes gestational merupakan diabetes yang terjadi pada wanita hamil. Diabetes tipe ini disebabkan karena perubahan hormon wanita pada saat hamil. Pada penelitian ini, peneliti melakukan penelitian untuk mengimplementasikan model klasifikasi pada dataset diabetes gestational untuk wanita dengan usia diatas 21 tahun[2].

Penelitian terkait klasifikasi penyakit diabetes sebelumnya pernah dilakukan untuk predisksi penyakit diabetes menggunakan algoritma decision tree C4.5 dengan judul "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes"dan dihasilkan hasil akurasi sebesar 70.32%. Selain itu penelitian terkait implementasi seleksi fitur juga sudah pernah dilakukan menggunakan algoritma Decision Tree C4,5 dengan judul "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes Menggunakan Algoritma C4.5" dan menghasilkan akurasi 76%[3].

Penelitian lainnya adalah penelitian yang berjudul "*Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes*" yang ditulis oleh Daghistani, T., & Alshammari, R. (2020) yang membandingkan kinerja algoritma Logistic Regression dan Random Forest dalam memprediksi diabetes. Penelitian tersebut menghasilkan nilai ROC AUC 0,708 untuk Logistic Regression dan

0,944 untuk Random Forest yang kemudian disimpulkan algoritma Random Forest memiliki kinerja yang lebih unggul dalam memprediksi diabetes[4].

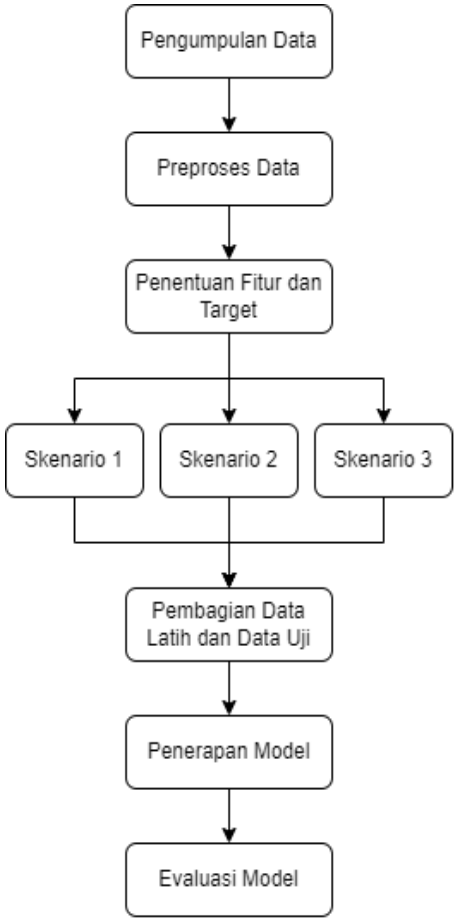
Penelitian terkait lainnya adalah penelitian tentang pengimplementasian *corellation matrix* untuk klasifikasi wine yang berjudul "Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine". Dalam penelitian tersebut, penggunakan seleksi fitur dengan menggunakan *corellation matrix* dapat meningkatkan nilai akurasi pada normalisasi Z-score dari 73,75% menjadi 75,62% dan padanormalisasi Min-max mampu meningkatkan akurasi dari 68,12% menjadi 71,25%[5].

Melihat dari penelitian-penelitian sebelumnya, penelitian yang dilakukan oleh peneliti kali ini bertujuan untuk membandingkan performance dari algoritma Logistic Regression untuk prediksi penyakit diabetes dengan seleksi fitur dan tanpa seleksi fitur untuk mengetahui apakah seleksi fitur dapat meningkatkan kinerja model untuk dalam memprediksi diabetes

2 Metode Penelitian

2.1 Alur Penelitian

Alur penelitian yang dilakukan oleh peneliti dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Tahapan yang pertama adalah mengumpulkan data. Dataset yang digunakan adalah dataset pasien diabetes perempuan dari Pima Indians Heritage yang didapatkan dari situs kaggle.com. Dataset yang digunakan memiliki data sebanyak 768 baris dengan 9 atribut yaitu *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age,* dan *Outcome.*

2.2 Preproses Data

Preproses data bertujuan untuk mengolah data yang sudah dikumpulkan menjadi data yang bersih dan siap untuk dilakukan analisis dan diterapkan model. Beberapa kegiatan yang dilakukan pada preproses data diantaranya adalah :

a. Pengecekan data kosong atau data yang hilang pada masing-masing atribut. Jika terdapat data yang kosong maka dapat dihapus jika persentasenya dibawah 5%. Jika persentasenya diatas 5% dilakukan pengisian data dengan nilai mean dari atribut yang relevan. jika persentasenya diatas 90% maka akan dilakukan penghapusan kolom
b. Pengecekan tipe data pada masing-masing atribut. Pengecekan tipe data dilakukan untuk memastikan tipe data dari masing-masing atribut sesuai dengan isi dan informasi dari masing-masing data. Jika terdapat atribut dengan tipe data yang janggal atau tidak sesuai, maka akan dilakukan konversi dengan tipe data yang relevan.
c. Konversi data kategorial ke data numerik. Tahapan ini bertujuan untuk mengubah semua atribut yang bertipe data objek menjadi numerik. Tahapan ini dilakukan sebagai syarat supaya dapat dilakukan seleksi fitur dengan *correllation matrix*.

2.3 Seleksi Fitur

Tahapan seleksi fitur berguna untuk menyeleksi fitur-fitur yang relevan dan berkorelasi kuat dengan variabel target. Seleksi fitur dilakukan dengan menggunakan *correllation matrix* yang divisualisasikan dengan visualisasi *heatmap* untuk memudahkan analisis[5]. Fitur yang terseleksi kemudian akan dijadikan sebagai prediktor yang akan digunakan oleh model untuk prediksi hasil klasifikasi dari variabel target. Matriks korelasi adalah sebuah matriks atau tabel yang berisikan nilai koefisien dari korelasi antar variabel. Nilai koefisian korelasi antar variabel didapatkan dengan rumus berikut[6]:

$$r = \frac{1}{n-1} \Sigma \left(\frac{x_i - \underline{x}}{S_x}\right)\left(\frac{y_i - \underline{y}}{S_y}\right)$$

dimana :

r = nilai korelasi antara x dengan y
n = banyaknya sample
x_i = nilai x ke-i
$\underline{x}$ = rata-rata dari x
S_x = standar deviasi dari x
y_i = nilai y ke-i
$\underline{y}$ = rata-rata dari y
S_x = standar deviasi dari y

2.4 Klasifikasi dengan Algoritma Logistic Regression

Logistic Regression adalah model statistik linear yang menghasilkan sebuah nilai biner[6]. Pada model logistic regression terdapat satu variabel dependen yang merupakan variabel biner yang diberi label 0 atau 1. Di lain sisi, model ini juga memiliki satu atau lebih variabel independen yang dapat berupa nilai kategorial biner atau nilai kontinu. Logistic regression merupakan salah satu model statistik yang umum digunakan untuk klasifikasi biner (hanya memiliki dua label atau kemungkinan, yaitu 0 atau 1). Pada penelitian ini, variabel target hanya memiliki dua nilai yaitu 0 = tidak diabetes dan 1 = diabetes.

2.5 Evaluasi

Tahapan evaluasi bertujuan untuk menilai hasil kerja dari model yang diterapkan pada dataset. Model akan dievaluasi dengan Confusion Matrix yang selanjutnya akan dihitung nilai presisi, *f1-score*, *recall*, dan akurasinya[6].

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

a. Confusion Matrix
Confusion Matrix adalah metode pengujian untuk klasifikasi yang bekerja dengan membandingkan hasil prediksi dengan hasil aktual dari dataset. Pengujian dengan Confusion Matrix pada dataset diabetes dapat dilihat di tabel 1.

Tabel 1. Confusion Matrix dari Dataset Diabetes

	Prediksi	
Aktual	Diabetes	Tidak Diabetes
Diabetes	*True Positive* (TP)	*False Negative* (FN)
Tidak Diabetes	*False Positive* (FP)	*True Negative* (TN)

b. Presisi
Presisi merupakan perhitungan untuk mengetahui jumlah data yang benar positif dari semua hasil prediksi yang positif. Adapun nilai presisi didapatkan dari persamaan :

$$presisi = \frac{TP}{TP + FP}$$

c. Recall
Recall merupakan perhitungan untuk mengetahui jumlah data yang prediksi benar positif dari hasil seluruh benar positif. Adapun nilai *recall* didapatkan dari persamaan :

$$recall = \frac{TP}{TP + FN}$$

d. F1-score
F1-score merupakan perhitungan untuk mengetahui rata-rata dari perbandingan presisi maupun *recall*. Adapun nilai *f1-score* didapatkan dari persamaan :

$$f1\ score = 2 \times \frac{recall \times presisi}{recall + presisi}$$

e. Akurasi
Akurasi merupakan perhitungan untuk mengetahui keakuratan model dalam klasifikasi yang benar. Adapun nilai akurasi didapatkan dari persamaan :

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

3 Hasil dan Pembahasan

3.1 Dataset

Dataset yang digunakan pada penelitian ini adalah dataset pasien diabetes perempuan yang terdiri dari total 768 data dengan 9 atribut yang terdiri dari 8 variabel fitur dan 1 variabel target. Variabel target dari dataset adalah data pada kolom 'Outcome'. Variabel fitur dari dataset adalah 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', dan 'Age'. Kemudian pada dataset dilakukan seleksi fitur dengan bantuan *correllation matrix*, yang kemudian diambil 4 variabel dengan korelasi tertinggi dengan variabel target. Selanjutnya dari data tersebut, dilakukan klasifikasi dengan dua skrenario yaitu skenario 1 yang menggunakan semua variabel independen sebagai variabel fitur dan skenario 2 yang hanya menggunakan 4 variabel hasil seleksi fitur sebagai variabel fitur. Tabel 2 berisi cuplikan sebagian data dari dataset diabetes yang digunakan pada penelitian ini.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

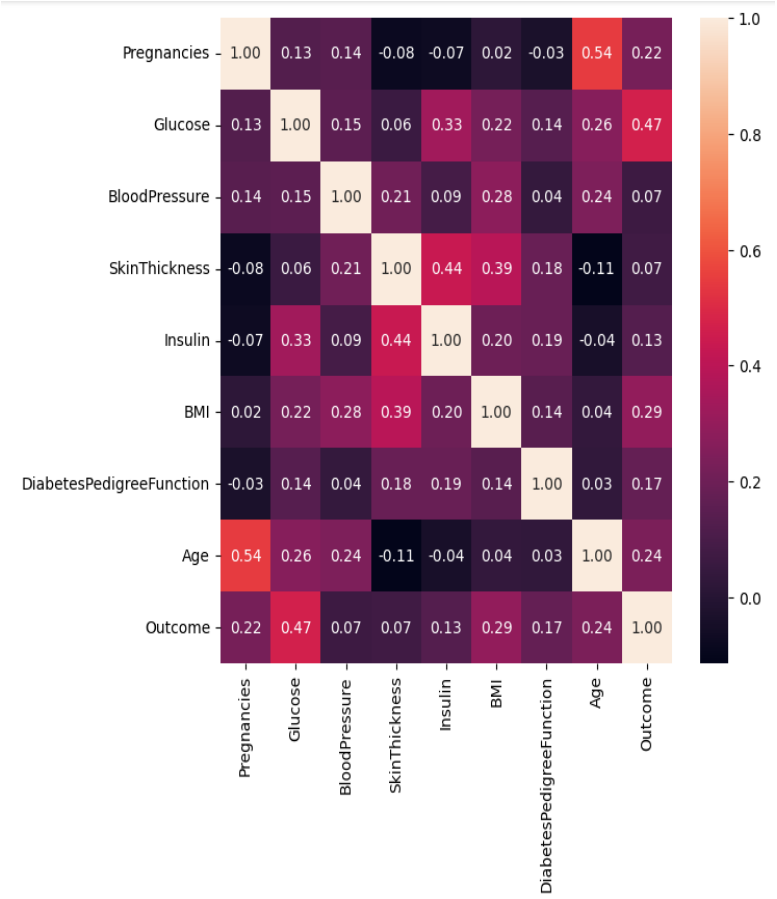
Tabel 2. Tabel Dataset Diabetes

No.	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1.	6	148	72	35	0	33,6	0,627	50	1
2.	1	85	66	39	0	26,6	0,351	31	0
3.	8	183	64	0	0	23,3	0,672	32	1
4.	1	89	66	23	94	28,1	0,167	21	0
5.	0	137	40	35	168	43,1	2,288	33	1

Masing-masing variabel pada dataset tersebut merepresentasikan sebuah nilai yang berkaitan dengan diabetes. Variabel 'Pregnancies' menunjukan berapa banyaknya kehamilan yang pernah dialami oleh, 'Glucose' menunjukan kadar plasma glukosa dalam waktu 2 jam pada test toleransi glukosa oral, 'BloodPressure' menunjukkan tekanan darah pasien dalam satuan mmHg, 'SkinThickness' menunjukkan ketebalan kulit pada sekitar otot trisep dalam satuan mm, 'Insulin' menunjukkan kadar insulin darah selama 2 jam pada insulin serum, 'BMI' menunjukkan Body Mass Indeks, 'DiabetesPedigreeFunction' menunjukkan angka kemungkinan terkena diabetes karena faktor keturunan, 'Age' menunjukkan usia dalam satuan tahun, dan 'Outcome' menunjukkan variabel target bernilai 0 jika tidak diabetes dan bernilai 1 jika diabetes.

3.2 Seleksi Fitur

Pada tahapan seleksi fitur dilakukan dengan menggunakan *correllation matrix* yang divisualisasikan dengan visualisasi *heatmap* untuk memudahkan analisis. Hasil *corellation matrix* dapat dilihat pada gambar 2.



Gambar 2. Correlation Matrix antar Variabel

Dari visualisasi di atas, korelasi dari masing-masing variabel fitur dengan variabel target (Outcome) dapat dilihat pada Tabel 3 di bawah ini yang diurutkan dari yang paling besar hingga yang paling kecil.

Tabel 3. Tabel Nilai Korelasi dengan Variabel Target

No	Variabel Fitur	Nilai Korelasi dengan Variabel Target (Outcome)
1.	Glucose	0,47
2.	BMI	0,29
3.	Age	0,24
4.	Pregnancies	0,22
5.	DiabetesPredigreeFunction	0,17
6.	Insulin	0,13
7.	BloodPressure	0.07
8.	SkinThickness	0,07

3.3 Skenario Pengujian

Dari penelitian yang dilakukan, dilakukan pengujian dengan dua skenario untuk model Logistic Regression dengan menggunakan *tools* yaitu Google Colab menggunakan bahasa pemrograman Python. Hasil kerja dari model Logistic Regression akan di evaluasi dengan berdasarkan akurasi, presisi, *recall*, dan *f1-score* yang didapatkan dari *confusion matrix*. Skenario pengujian memiliki perbedaan pada atribut yang digunakan sebagai fitur. Skenario 1 menggunakan semua atribut dari dataset, skenario 2 menggunakan 4 atribut yaitu *Pregnancies, Glucose, BMI,* dan *age* dan skenario 3 menggunakan 6 atribut yaitu *Pregnancies, Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age*. Fitur pada skenario 2 diambil dari 4 variabel dengan nilai korelasi tertinggi dengan variabel target, sedangkan fitur pada skenario 3 diambil dari 6 variabel dengan nilai korelasi tertinggi dengan variabel target. Dari ketiga skenario tersebut, peneliti kemudian membagi dataset untuk data latih dan data uji dengan perbandingan 3 : 1, 75% dari dataset dialokasikan untuk data latih dan 25% dari dataset dialokasikan untuk data uji. Adapun skenario yang dibuat dapat dilihat lebih lengkap pada Tabel 4.

Tabel 4. Skenario Pengujian

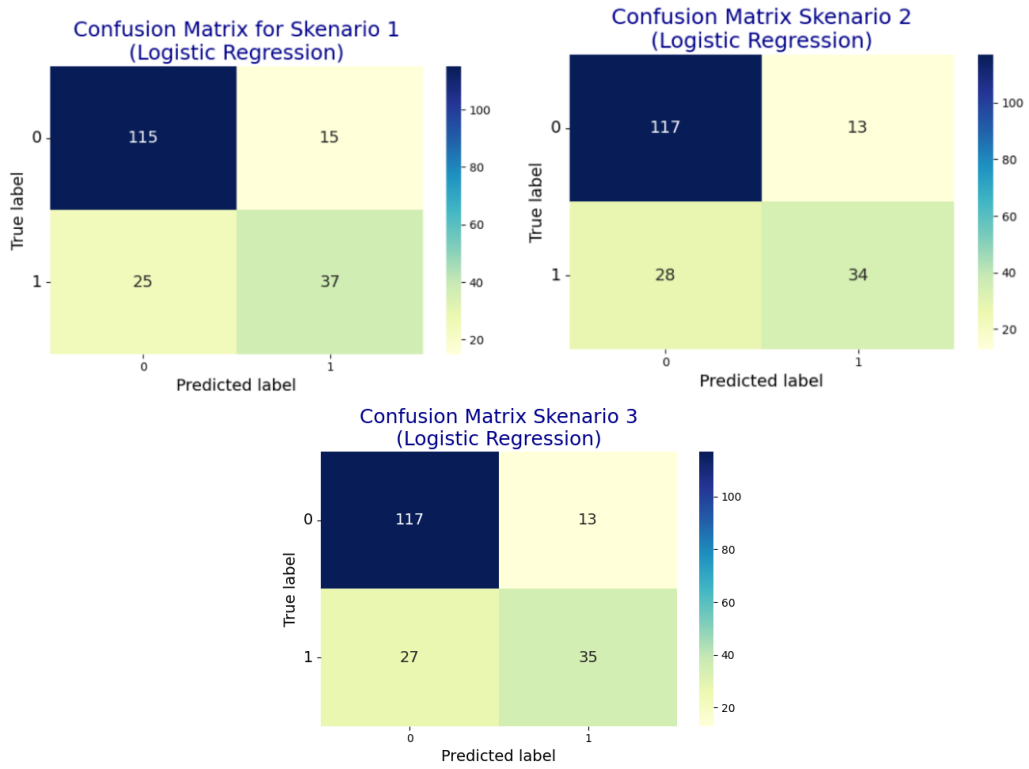
Skenario	Atribut Fitur
Skenario 1	*Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age*
Skenario 2	*Glucose, BMI, Age, Pregnancies*
Skenario 3	*Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, Insulin*

3.3 Hasil Prediksi

Hasil prediksi dari klasifikasi pada dataset dengan algoritma Logistic Regression kemudian direpresentasikan dengan confusion matrix. Gambar 3 secara berurutan adalah hasil visualisasi *confusion matrix* dari prediksi pada skenario 1, skenario 2, dan skenario 3.

Dari Gambar 3 dapat dilihat skenario 1 memiliki total prediksi yang tepat sebanyak 152 dan prediksi yang tidak tepat sebanyak 40. Skenario 2 mendapatkan total prediksi tepat sebanyak 151 dan prediksi yang tidak tepat sebanyak 41. Skenario 3 mendaparkan total prediksi tepat sebanyak 152 dan prediksi yang tidak tepat sebanyak 40.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)



Gambar 3. Confusion Matrix untuk Skenario 1, Skenario 2, dan Skenario 3

3.4 Evaluasi

Berdasarkan hasil prediksi pada dataset diabetes yang digunakan oleh peneliti, didapatkan nilai akurasi, presisi, recall, dan F1 Score seperti pada tabel 4.

Tabel 4. Hasil Evaluasi Model

Skenario	Akurasi	Presisi	Recall	F1 Score
Skenario 1	79,1%	77%	74%	75%
Skenario 2	78,6%	77%	72%	74%
Skenario 3	79,1%	77%	73%	75%

Dari tabel evaluasi model di atas, hasil evaluasi kurang menunjukkan hasil yang signifikan. Dapat dilihat dengan nilai akurasi dari masing-masing skenario yang tidak terlalu berbeda jauh, dengan skenario 1 dan 3 lebih tinggi 0,05% jika dibandingkan dengan skenario 2. Ketiga skenario juga memiliki nilai presisi yang sama. Begitu pula untuk *recall* dan *f1-score*, secara keseluruhan skenario 1 memiliki persentase tertinggi, disusul skenario 3, dan skenario 2. Namun, karena kasus yang diambil oleh peneliti adalah prediksi diabetes, sehingga akan berbahaya ketika terjadi pasien yang diabetes salah diprediksi sebagai tidak diabetes, maka dari itu peneliti lebih mempertimbangkan berdasarkan nilai presisi dan akurasi yang terbaik dari ketiga skenario di atas.

Dari hasil pengujian dari dataset Pima Indians Heritage Diabetes, skenario 1 yaitu pengujian dengan model Logistic Regression menggunakan semua variable dependent, mendapatkan nilai akurasi 79,1%, presisi 77%, *recall* 74%, dan *f1-score* 75%, kemudian disusul dengan skenario 3 dengan akurasi 79,1%, presisi 77%, *recall* 73%, dan *f1-score* 75%. Adapun skenario 2 mendapatkan nilai akurasi 78,6%, presisi 77%, *recall* 72%, dan *f1-score* 74%.

4 Kesimpulan

Dari nilai tersebut, peneliti menyimpulkan bahwa prediksi dengan skenario 1 mendapatkan performance terbaik dengan nilai presisi 77%, akurasi 79,1%, *recall* 74% dan *f1-score*75%. Sehingga dari serangkaian proses penelitian, dapat disimpulkan bahwa prediksi diabetes menggunakan model Logistic Regression tanpa seleksi fitur memiliki performance yang lebih unggul dibandingkan dengan prediksi menggunakan model Logistic Regression dengan seleksi fitur.

Referensi

[1] World Health Organization: WHO. (2023). Diabetes. www.who.int. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] R.M.M. Khan, Z.J.Y. Chua, J.C. Tan, Y. Yang, Z. Liao, Y. Zhao, From pre-diabetes to diabetes: diagnosis, treatments and translational research, Medicina (B Aires) 55 (9) (2019) 546.

[3] Noviandi, N. (2018). Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes. Jurnal INOHIM, Volume 6 Nomor 1, Juni 2018, 6(01), 1–5. https://doi.org/10.47007/inohim.v6i1.142

[4] Daghistani, T., & Alshammari, R. (2020). Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes. Journal of Advances in Information Technology, 78–83. https://doi.org/10.12720/jait.11.2.78-83

[5] Khakim, E. N. R., Hermawan, A., & Avianto, D. (2023). Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine. JIKO (Jurnal Informatika Dan Komputer), 7(1), 158. https://doi.org/10.26798/jiko.v7i1.771

[6] Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.

[7] Ardiansyah, M., Sunyoto, A., & Luthfi, E. T. (2021). Analisis Perbandingan Akurasi Algoritma Naïve Bayes Dan C4.5 untuk Klasifikasi Diabetes. Edumatic : Jurnal Pendidikan Informatika, 5(2), 147–156. https://doi.org/10.29408/edumatic.v5i2.3424

[8] Diagnosis and Classification of Diabetes Mellitus. (2013). Diabetes Care, 37(Supplement_1), S81–S90. https://doi.org/10.2337/dc14-s081

[9] Argina, A. W. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. Indonesian Journal of Data and Science, 1(2), 29–33. https://doi.org/10.33096/ijodas.v1i2.11

[10] Fadhillah, R. P., Rahma, R., Sepharni, A., Mufidah, R., Sari, B. N., & Pangestu, A. (2022). Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5. JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika), 7(4), 1265–1270. https://doi.org/10.29100/jipi.v7i4.3248

[11] Handayani, F. (2021). Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung. JEPIN (Jurnal Edukasi Dan Penelitian Informatika), Vol. 7, No. 3, Desember 2021, 7(3), 329. https://doi.org/10.26418/jp.v7i3.48053