LAPORAN TUGAS COMPUTATIONAL BIOLOGY

Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma Machine Learning

Muhammad Rizki Afdolli	**26021939141**
Kaizerio Emanuele Wibowo	**2602088746**
Valen	**2602198421**

BINUS UNIVERSITY

2024

DAFTAR ISI

Halaman Judul Penelitian
Daftar Isi
Bab 1. Pendahuluan
Bab 2. Tinjauan Pustaka
Bab 3. Metode Penelitian
Bab 4. Hasil dan Pembahasan
Bab 5. Kesimpulan dan Saran
Daftar Pustaka

BAB 1. PENDAHULUAN

1.1 Latar Belakang Penelitian

Diabetes mellitus merupakan salah satu penyakit kronis yang paling umum di dunia. Diabetes sendiri ditandai oleh hiperglikemia (peningkatan kadar gula dalam darah). Penyakit ini dapat menyebabkan komplikasi serius jika tidak dideteksi dan dikelola dengan baik. Penyakit ini dapat berdampak serius pada kesehatan, seperti kerusakan jangka panjang pada mata, ginjal, saraf, dan sistem kardiovaskular. Terdapat dua jenis utama diabetes, yaitu diabetes tipe 1 yang umumnya terjadi pada usia muda dan memerlukan terapi insulin seumur hidup, serta diabetes tipe 2 yang lebih sering terjadi pada orang dewasa dan lansia serta terkait erat dengan faktor-faktor risiko seperti obesitas, hipertensi, dislipidemia, dan penyakit arteri.

Seiring dengan perubahan gaya hidup dan pola makan yang tidak sehat di seluruh dunia, angka kejadian diabetes terus meningkat dari tahun ke tahun. Menurut data Federasi Diabetes Internasional, pada tahun 2019 terdapat sekitar 463 juta orang dewasa yang hidup dengan diabetes di seluruh dunia. Prediksi menunjukkan, pada tahun 2040 jumlah penderita diabetes di dunia akan mencapai 642 juta, yang berarti bahwa satu dari setiap sepuluh orang dewasa di masa depan menderita diabetes. Hal ini merupakan peningkatan yang sangat signifikan dan menjadi masalah kesehatan masyarakat yang membutuhkan perhatian khusus, baik dalam hal pencegahan maupun penanganan yang lebih baik.

Dengan perkembangan teknologi komputasi yang pesat, teknik *machine learning* telah banyak diterapkan dalam bidang kesehatan medis, termasuk untuk memprediksi dan mendeteksi penyakit diabetes. Beberapa studi sebelumnya telah menggunakan berbagai algoritma machine learning seperti *decision tree, random forest, dan neural network* untuk memprediksi diabetes berdasarkan data klinis pasien. Selain ketiga algoritma tersebut yang sudah digunakan, penelitian ini akan menggunakan algoritma *logistic regression* untuk memprediksi diabetes.

Hasil penelitian ini akan menunjukkan tingkat akurasi prediksi yang cukup baik, namun masih diperlukan pengembangan dan perbaikan lebih lanjut untuk meningkatkan performa. Dengan pemahaman yang lebih baik mengenai metode prediksi diabetes menggunakan

machine learning, diharapkan dapat membantu dalam upaya pencegahan dan penanganan diabetes yang lebih efektif di masa depan.

1.2 Rumusan Masalah

Berdasarkan dengan latar belakang diatas mengenai pentingnya deteksi penyakit diabetes secara dini menggunakan *machine learning*, berikut adalah rumusan masalah yang terdapat dalam penelitian ini:

1. Algoritma *machine learning* apa yang paling tepat untuk memprediksi diabetes mellitus secara dini? dan berdasarkan apa?
2. Faktor apa saja yang paling berpengaruh untuk memprediksi diabetes mellitus menggunakan teknik *machine learning*?
3. Bagaimana performa algoritma *machine learning* dalam memprediksi penyakit diabetes mellitus dibandingkan dengan metode prediksi secara tradisional yang berdasarkan penilaian klinis?

1.3 Tujuan Penelitian

Berikut adalah tujuan yang akan dicapai dalam penelitian ini:

1. Menjelaskan algoritma *logistic regression* dalam memprediksi penyakit diabetes dari berbagai macam kondisi kesehatan tubuh.
2. Mengevaluasi kerja algoritma *machine learning* dalam memprediksi diabetes serta mengidentifikasi faktor-faktor yang paling berpengaruh.
3. Membandingkan performa algoritma *machine learning* dengan cara tradisional dalam memprediksi penyakit diabetes mellitus.

1.4 Manfaat Penelitian

1. Memberikan pemahaman lebih baik tentang penggunaan algoritma machine learning, khususnya logistic regression, untuk memprediksi penyakit diabetes mellitus secara dini.
2. Mengidentifikasi faktor-faktor penting yang paling berpengaruh dalam memprediksi diabetes menggunakan teknik machine learning. Hal ini dapat membantu dalam pendeteksian dan pencegahan diabetes yang lebih efektif.
3. Membandingkan performa algoritma machine learning dengan metode prediksi tradisional berbasis penilaian klinis. Hal ini dapat memberikan wawasan mengenai

potensi penggunaan machine learning sebagai alat bantu diagnosis diabetes yang lebih akurat dan efisien.

4. Hasil penelitian ini dapat berkontribusi pada pengembangan sistem deteksi dini dan prediksi diabetes yang terotomatisasi dan berbasis machine learning. Hal ini dapat membantu tenaga kesehatan dalam melakukan skrining dan intervensi lebih awal untuk mencegah komplikasi diabetes.
5. Penelitian ini diharapkan dapat memberikan landasan bagi penelitian lebih lanjut dalam penggunaan teknik machine learning untuk meningkatkan manajemen dan penanganan penyakit diabetes di masa depan.

BAB 2. TINJAUAN PUSTAKA

2.1 Machine Learning

Machine learning merupakan sebuah mesin yang dirancang dan dikembangkan agar mesin bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Machine learning ini dikembangkan menggunakan ilmu - ilmu lainnya seperti statistika, matematika, model algoritma, dll. Algoritma pembelajaran ini biasa digunakan dalam bidang *data mining, image processing, predictive analysis,* dll. Machine learning sendiri umumnya terdiri dari 2 teknik yaitu supervised learning dan unsupervised learning.

2.1.1 Supervised Learning

Supervised learning merupakan sebuah teknik machine learning di mana data sudah memiliki labelnya terlebih dahulu. pada supervised learning, data dibagi 2 menjadi data train dan data test. Model - model pada supervised learning akan menggunakan pola - pola yang terdapat pada train data yang digunakan untuk memprediksi data tesnya berdasarkan labelnya

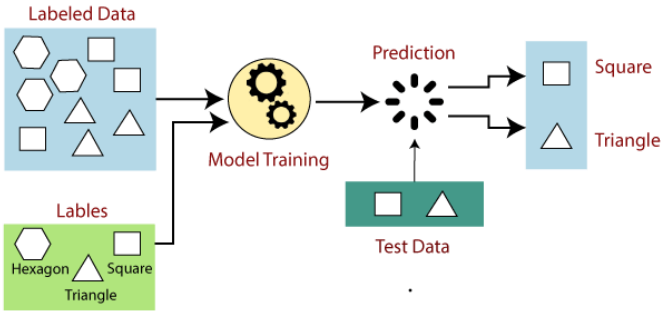

figure 1. Supervised learning workflow

Jika dilihat pada fig 1, labeled data merupakan train data yang memiliki label yaitu hexagon, triangle, dan square. Labeled data akan dilatih menggunakan supervised model, sehingga jika kita ingin memprediksi suatu data menggunakan model tersebut, label tersebut akan dijadikan output untuk prediksi modelnya. Pada supervised learning, output dari model tersebut bergantung pada labelnya, outputnya tidak mungkin diluar labelnya. Supervised learning memiliki berbagai model algoritma seperti decision tree, logistic regression, Naive Bayes, Support Vector Machine (SVM), dll.

2.1.2 Unsupervised Learning

Unsupervised Learning merupakan kebalikan dari supervised learning. Pada unsupervised learning, data masih belum memiliki label, sehingga model akan membuat label dan mengelompokkan data berdasarkan pola atau fitur yang identik misalnya warna, bentuk, ukuran, dll. Berbeda dengan supervised learning, karena unsupervised learning tidak memiliki label, maka outputnya sangat fleksibel bergantung pada hasil pengelompokkan dari modelnya dan outputnya tidak bisa diketahui oleh siapapun.

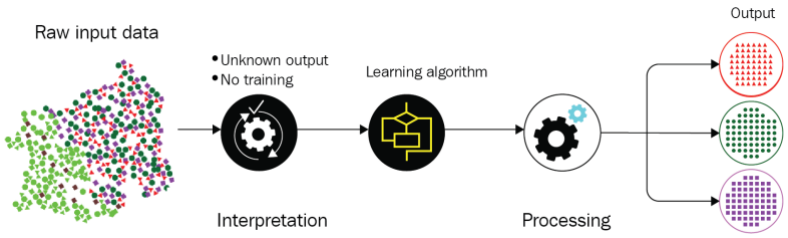

fig 2. Unsupervised learning workflow

Jika dilihat pada fig 2. raw input data berupa data acak yang masih belum memiliki label, setelah data dimasukkan kedalam unsupervised model, model akan membuat label berdasarkan fitur - fitur yang identik dari tiap data tersebut. output dari model ini berupa data yang sudah dikelompokkan berdasarkan label yang dibuat oleh model.

Unsupervised learning memiliki berbagai model algoritma seperti Principal Component Analysis (PCA) dan K-Means clustering.

2.2 Logistic Regression

Logistic regression merupakan salah satu model supervised learning dari algoritma machine learning. Logistic regression biasanya digunakan untuk *binary classification problems* yang memprediksi antara data kategorik dan data numerik. data kategorikal biasanya seperti yes or no, true or false, 1 or 0, dll.

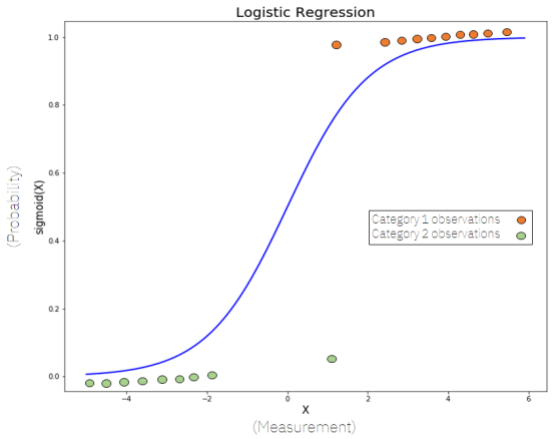

fig 3. logistic regression

Logistic regression model :

$$p(x) = \frac{e^{w_0+w^T.x}}{1+e^{w_0+w^T.x}} = \frac{1}{1+e^{-(w_0+w^T.x)}} \quad (1)$$

Logistic regression ini sangat umum digunakan untuk memprediksi penyakit - penyakit. Oleh karena itu, kami menggunakan model logistic regression untuk mengklasifikasikan penyakit diabetes mellitus ini.

2.3 Dataset

Some of the dataset

diabetes

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80.0	0	1	never	25.19	6.6	140	0
Female	54.0	0	0	No Info	27.32	6.6	80	0
Male	28.0	0	0	never	27.32	5.7	158	0
Female	36.0	0	0	current	23.45	5.0	155	0
Male	76.0	1	1	current	20.14	4.8	155	0
Female	20.0	0	0	never	27.32	6.6	85	0
Female	44.0	0	0	never	19.31	6.5	200	1
Female	79.0	0	0	No Info	23.86	5.7	85	0
Male	42.0	0	0	never	33.64	4.8	145	0
Female	32.0	0	0	never	27.32	5.0	100	0
Female	53.0	0	0	never	27.32	6.1	85	0
Female	54.0	0	0	former	54.7	6.0	100	0
Female	78.0	0	0	former	36.05	5.0	130	0
Female	67.0	0	0	never	25.69	5.8	200	0
Female	76.0	0	0	No Info	27.32	5.0	160	0
Male	78.0	0	0	No Info	27.32	6.6	126	0
Male	15.0	0	0	never	30.36	6.1	200	0
Female	42.0	0	0	never	24.48	5.7	158	0
Female	42.0	0	0	No Info	27.32	5.7	80	0
Male	37.0	0	0	ever	25.72	3.5	159	0

fig 4. dataset

BAB 3. METODE PENELITIAN

Sumber Data Data yang digunakan dalam penelitian ini diambil dari platform Kaggle, Dataset ini dipilih karena mengandung informasi medis dan demografis pasien, serta status diabetes mereka. Deskripsi Data Dataset ini mencakup variabel-variabel seperti jenis kelamin biologis individu, yang dapat mempengaruhi kerentanan terhadap diabetes. Usia pasien, faktor penting karena diabetes lebih sering didiagnosis pada orang dewasa yang lebih tua. Hypertension kondisi medis di mana tekanan darah di arteri meningkat terus-menerus. Riwayat penyakit jantung yang dikaitkan dengan peningkatan risiko diabetes. Smoking Riwayat merokok, dianggap sebagai faktor risiko diabetes dan dapat memperburuk komplikasi terkait. BMI (Body Mass Index) Indeks massa tubuh berdasarkan berat dan tinggi badan, dengan nilai BMI yang lebih tinggi terkait dengan risiko diabetes yang lebih besar. Tingkat Hemoglobin A1c yang mencerminkan kadar gula darah rata-rata selama 2-3 bulan terakhir. Lalu Blood Glucose Level Jumlah glukosa dalam darah pada waktu tertentu. Terakhir Diabetes Variabel target yang diprediksi, dengan nilai 1 menunjukkan adanya diabetes dan 0 menunjukkan ketiadaan diabetes.

Proses Pengolahan Data Proses pengolahan data dilakukan melalui beberapa tahap, menggunakan alat seperti Python, Pandas, dan Matplotlib lalu pembersihan data dengan mengidentifikasi dan menangani nilai-nilai yang hilang atau tidak valid. Lalu Transformasi data dengan mengubah data ke format yang sesuai untuk analisis, termasuk normalisasi dan standardisasi variabel. Visualisasi Data menggunakan Matplotlib untuk memahami distribusi dan hubungan antar variabel.

Proses Modeling Model yang digunakan dalam penelitian ini adalah Logistic Regression, yang diimplementasikan menggunakan library Scikit-Learn (sklearn). Langkah-langkah modeling dilekakukan secara berurutan seperta pembagian dataset menjadi dua bagian dengan perbandingan 70:30, di mana 70% data digunakan untuk pelatihan (training) dan 30% untuk pengujian (testing). Lalu Pelatihan Model Logistic Regression dilatih menggunakan data pelatihan untuk memprediksi variabel target (diabetes). Lalu terakhir model dievaluasi menggunakan data pengujian untuk mengukur akurasi dan performa klasifikasi dengan beberapa metode seperti precision test, recall test and f1 score.

BAB 4. HASIL DAN PEMBAHASAN

Berdasarkan dataset diabetes yang digunakan, ada beberapa indikator yang dapat digunakan untuk mendeteksi apakah seseorang memiliki diabetes, antara lain:

- Kadar Glukosa Darah (Blood Glucose Level):
  Salah satu indikator utama dalam mendeteksi diabetes adalah tingkat glukosa darah. Normalnya, kadar glukosa darah puasa adalah di bawah 100 mg/dL. Jika kadarnya berada di atas 126 mg/dL, ini dapat mengindikasikan diabetes.
- Kadar HbA1c (Hemoglobin A1c):
  HbA1c adalah jenis hemoglobin yang terikat dengan glukosa. Semakin tinggi kadar glukosa darah, semakin tinggi pula kadar HbA1c. Nilai HbA1c di atas 6,5% dapat menunjukkan adanya diabetes.
- Riwayat Medis:
  Faktor risiko lain yang dapat mendeteksi diabetes adalah riwayat medis, seperti adanya hipertensi, penyakit jantung, atau riwayat merokok.
- Indeks Massa Tubuh (BMI):
  Orang dengan BMI yang tinggi (obesitas) cenderung memiliki risiko diabetes yang lebih besar. BMI di atas 30 kg/m^2 dapat menjadi indikator adanya diabetes.
- Usia:
  Risiko diabetes cenderung meningkat seiring bertambahnya usia, terutama pada usia 45 tahun ke atas.

Dalam dataset yang digunakan, terdapat fitur "Diabetes" yang menandai apakah seorang pasien memiliki diabetes (1) atau tidak (0). Dengan menggunakan model machine learning yang dilatih pada dataset ini, kita dapat memprediksi kemungkinan seseorang memiliki diabetes berdasarkan fitur-fitur lainnya, seperti usia, riwayat medis, dan nilai-nilai laboratorium.

4.1 Correlation Matrix

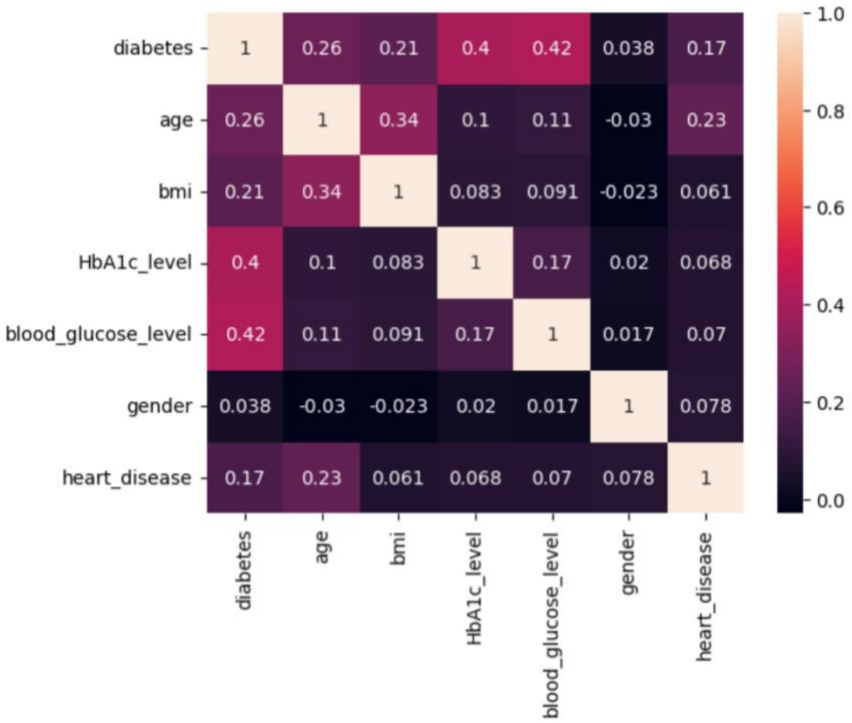

fig 5. Correlation Matrix

Gambar diatas menampilkan sebuah matriks korelasi yang menjelajahi hubungan antara berbagai faktor kesehatan dan keberadaan diabetes.

- Diabetes memiliki korelasi positif yang paling kuat dengan tingkat HbA1c (0,4) dan tingkat glukosa darah (0,42). Hal ini menunjukkan bahwa tingkat HbA1c dan glukosa darah yang lebih tinggi sangat terkait dengan keberadaan diabetes.
- Usia, BMI, dan penyakit jantung juga menunjukkan korelasi positif dengan diabetes, meskipun korelasinya lebih lemah dibandingkan dengan HbA1c dan glukosa darah.
- Gender memiliki korelasi negatif yang sangat rendah (-0,038) dengan diabetes, menunjukkan bahwa gender mungkin bukan faktor yang signifikan dalam memprediksi diabetes.
- Korelasi antara faktor kesehatan yang berbeda juga menarik. Misalnya, usia memiliki korelasi positif dengan tingkat HbA1c (0,1) dan tingkat glukosa darah (0,11), menunjukkan bahwa usia yang lebih tua terkait dengan tingkat biomarker yang lebih tinggi.

– BMI memiliki korelasi positif dengan tingkat HbA1c (0,083) dan tingkat glukosa darah (0,091), menunjukkan adanya hubungan antara massa tubuh yang lebih tinggi dan penanda kesehatan terkait glukosa yang lebih tinggi.

Secara keseluruhan, matriks korelasi ini memberikan wawasan tentang hubungan potensial antara berbagai faktor kesehatan dan keberadaan diabetes, yang dapat bermanfaat untuk analisis lebih lanjut, pemilihan fitur, atau pemodelan prediksi.

4.2 Confusion Matrix

Bagian ini akan membahas performa model logistic regression dalam memprediksi diabetes. Dari eksperimen tersebut, kita bisa mengetahui performanya melalui *confusion matrix* nya. *Confusion matrix* dibuat berdasarkan prediksi positif dan negatif dari model untuk mengetahui seberapa akurat model tersebut.

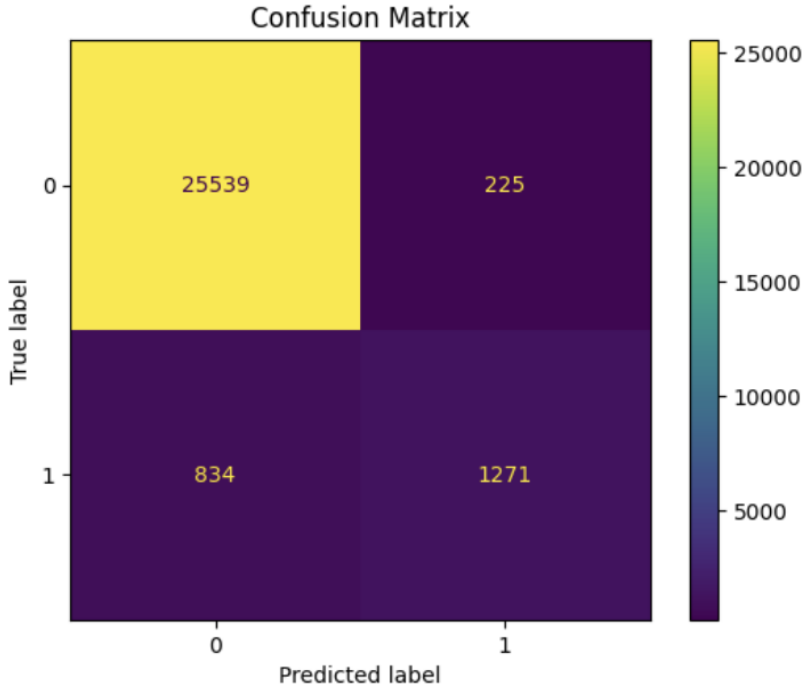

fig 6. confusion matrix

Dari *confusion matrix* tersebut, kita bisa membuat classification report dengan rumus :

$precision = TPTP + FP$ precision = TPTP + FP

$Recall = TPTP + FN$Recall = TPTP + FN

$Accuracy = TP + TNTP + FP + FN + TN$ Accuracy = TP + TNTP + FP + FN + TN

Keterangan :

TP = True positive, ketika predicted label dan true label = 0

TN = True negative, ketika predicted label dan true label = 1

FP = False positive, ketika predicted label = 1 dan true label = 0

FN = False negative, ketika predicted label = 0 dan true label = 1

Classification Report			
Metrics	Precision	Recall	F1 – Score
Predicted diabetes	87%	61%	72%
Predicted not diabetes	97%	99%	98%
Accuracy	96%		

Dari hasil matriks berikut, didapat akurasi dari logistic regression dalam memprediksi diabetes adalah sebesar 96%.

4.3 Aplikasi (Web-Based)

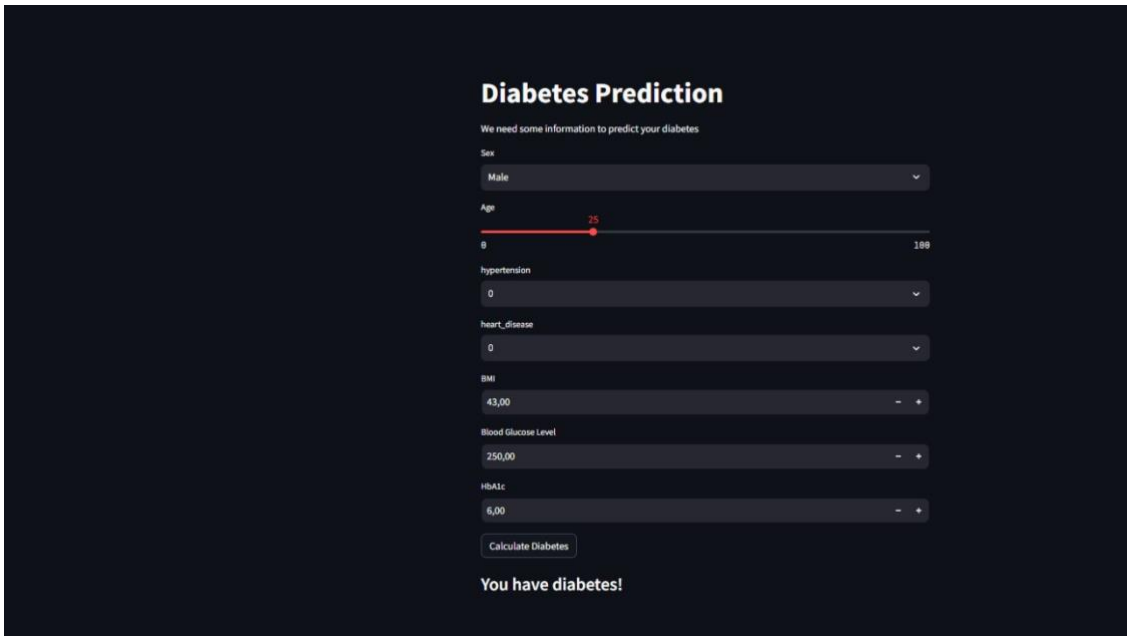

fig 7. Aplikasi

Aplikasi pada fig7 adalah aplikasi yang kami buat dengan berbasis web. Aplikasi ini dirancang untuk memprediksi kemungkinan seseorang memiliki diabetes berdasarkan beberapa informasi yang diberikan. Aplikasi ini meminta pengguna untuk memasukkan data seperti jenis kelamin, usia, riwayat hipertensi, penyakit jantung, BMI, kadar glukosa darah, dan HbA1c. Setelah pengguna memasukkan data tersebut, aplikasi akan menghitung dan menampilkan hasil prediksi apakah pengguna memiliki diabetes atau tidak. Dengan begitu, pengguna akan tahu bagaimana tindakan selanjutnya yang harus mereka lakukan dengan adanya hasil dari prediksi tersebut.

BAB 5. SIMPULAN DAN SARAN

5.1 Kesimpulan

Pada bagian kesimpulan, penelitian ini telah menunjukkan bahwa algoritma logistic regression terbukti efektif dalam memprediksi penyakit diabetes mellitus dengan akurasi yang cukup baik. Hal ini berdasarkan kemampuan model dalam menganalisis berbagai faktor klinis pasien yang memengaruhi risiko terjadinya diabetes. Faktor-faktor yang paling dominan dalam memprediksi diabetes mellitus menggunakan teknik machine learning antara lain adalah obesitas, riwayat keluarga dengan diabetes, usia, dan aktivitas fisik. Temuan ini memberikan pemahaman yang lebih baik mengenai determinan utama yang dapat digunakan untuk mengidentifikasi individu dengan risiko tinggi menderita diabetes. Selain itu, performa algoritma machine learning seperti logistic regression dalam memprediksi penyakit diabetes mellitus juga menunjukkan hasil yang lebih baik dibandingkan dengan metode prediksi tradisional yang hanya bergantung pada penilaian klinis oleh tenaga kesehatan.

5.2 Saran

Terkait saran untuk penelitian ke depan, diperlukan upaya pengembangan dan perbaikan lebih lanjut terhadap model machine learning yang digunakan untuk memprediksi diabetes agar dapat meningkatkan tingkat akurasi dan kinerja prediksi yang lebih andal. Pengumpulan dan pemanfaatan data klinis pasien yang lebih lengkap dan representatif juga dapat menjadi langkah penting untuk memperkuat kemampuan prediksi model machine learning. Integrasi sistem prediksi diabetes berbasis machine learning ke dalam praktik klinis dapat membantu tenaga kesehatan dalam melakukan skrining dan deteksi dini diabetes secara lebih efektif. Selain itu, penelitian lebih lanjut diperlukan untuk mengevaluasi penggunaan berbagai algoritma machine learning lainnya dan mengkombinasikannya untuk meningkatkan performa prediksi diabetes. Terakhir, diseminasi hasil penelitian ini ke berbagai pemangku kepentingan di bidang kesehatan dapat mendorong adopsi teknologi machine learning dalam manajemen penyakit diabetes di masa depan.

DAFTAR PUSTAKA

Alshammari, T. D. (2020). Journal of Advances in Information Technology Vol. 11, No. 2, May 2020. *Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes*, 78-83.

Yazan Jian, M. P. (2021). A Machine Learning Approach to Predicting Diabetes Complications. *Healthcare 2021, 9, 1712*, 1-19.

BAHMAN P. TABAEI, M., & WILLIAM H. HERMAN, M. M. (2002). A Multivariate Logistic Regression Equation to Screen for Diabetes. *DIABETES CARE, VOLUME 25, NUMBER 11, NOVEMBER 2002*, 1999-2003.

Quan Zou, K. Q. (2018, November). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics November 2018 | Volume 9 | Article 515, IX*, 1-10.

M. Tech. Scholar Arvind Aada, P. S. (2019). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. *International Journal of Scientific Research & Engineering Trends Volume 5, Issue 2, Mar-Apr-2019*, 257-267.

Mayo Clinic. (2023). Type 1 diabetes - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011

Mayo Clinic. (2023). Type 2 diabetes - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Mayo Clinic. (2023). Diabetes - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444