Deteksi Penyakit Diabetes Menggunakan Gaussian Naive Bayes, Regresi Logistik, dan Random Forest

Kenny Belle Lesmana[a1], I Ketut Gede Suhartana[a2],

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]kennybelle015@unud.ac.id
[2]ikg.suhartana@unud.ac.id

Abstract

Diabetes is a very common health problem in the world. The number of people with diabetes is increasing year to year. Therefore, it is necessary to realize the symptoms of diabetes as early as possible. Diabetes is a chronic disease characterized by high sugar levels in the blood. In this study, a system was made about a diabetes detection system based on numerical data using three methods. That three methods are Gaussian Naive Bayes method, Logistic Regression, and Random Forest by taking a dataset in the form of numerical data. The accuracy value on the data tested in this study using Gaussian Naive Bayes, Logistic Regression, Random Forest is 0.74; 0;78; 078.

Keywords: *Gaussian Naive Bayes, Regresi Logistik, Random Forest*

1. Pendahuluan

Diabetes merupakan salah satu penyakit kronis yang ditandai dengan kadar glukosa (gula) yang tinggi dalam darah. Gejala umum pada diabetes yaitu mulai dari haus yang berlebihan, kelelahan yang berlebihan, sampai penurunan berat badan tanpa alasan yang jelas. Maka dari itu, sangat penting untuk mengecek dan mengelola diabetes dengan baik karena akan memiliki dampak yang sangat serius jika tidak segera diobati. Sistem pendeteksi penyakit diabetes diperlukan untuk penanganan yang lebih cepat. Pada proses pembuatannya dapat menggunakan convolutional neural network (CNN), Decision Tree, pengolahan citra (image processing). Namun, pada penelitian ini, akan dibuat sistem pendeteksi penyakit diabetes pada manusia berdasarkan data numerik menggunakan tiga metode, yaitu metode Gaussian Naive Bayes, Regresi Logistik, dan Random Forest dengan mengambil dataset berupa data numerik.

2. Metode Penelitian

Proses pada penelitian ini dibagi menjadi beberapa bagian yaitu studi pengumpulan data, metode yang digunakan, dan pembuatan sistem.

2.1 Pengumpulan Data

Pada penelitian ini, dataset berupa data sekunder yang diambil dari website kaggle https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset. Data tersebut merupakan data numerik yang akan dimasukan sebagai input dari sistem yang akan dibuat pada penelitian ini.

2.2 Metode

Metode yang digunakan pada pembuatan sistem ini yaitu Gaussian Naive Bayes, Regresi Logistik, dan Random Forest.

a. Gaussian Naive Bayes

Naive Bayes merupakan salah satu metode pengklasifikasian data menggunakan perhitungan probabilitas yang akan terjadi di masa depan berdasarkan peristiwa yang pernah terjadi sebelumnya. Klasifikasi bayes merupakan salah satu metode klasifikasi yang memiliki akurasi paling tinggi diantara metode klasifikasi lainnya [1]. Pada penelitian ini, data yang dimasukkan berupa data numerik dan menghasilkan data yang numerik. Maka dari itu dapat memakai fungsi Probability Density Function (PDF) [2]. PDF merupakan sebuah konsep yang digunakan teori peluang dalam menggambarkan seberapa besar probabilitas terdistribusi dalam rentang nilai-nilai tertentu. Berikut adalah rumus dari PDF dalam persamaan 1 dan persamaan 2 merupakan rumus dasar standar deviasi atau yang disebut sebagai formula Gaussian Naive Bayes Classifier [3].

$$P(X_1 = x_1 \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\underline{x})^2}{2\sigma^2}} \quad (1)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_1-\underline{x})^2}{n-1}} \quad (2)$$

Keterangan:

P : Probabilitas
X_i : Atribut
x_i : nilai atribut
Y : Kelas yang berhubungan
y_j : sub kelas yang berhubungan
$\underline{x}$: rata-rata
σ : standar deviasi
n : banyaknya data

b. Regresi Logistik

Regresi Logistik merupakan salah satu metode analisis statistika dengan tujuan mengetahui hubungan antara variabel terkait yang memiliki dua kategori atau lebih dengan satu atau lebih peubah bebas berskala kategori atau kontinu [4]. Regresi Logistik biasanya digunakan pada analisis prediktif dan dalam pemodelan data yang mana variabel target hanya memiliki dua nilai yaitu 0 dan 1 (ya atau tidak). Maka dari itu persamaan rumus mengikuti distribusi Bernoulli yaitu sebagai berikut [5].

$$f(y_i) = \pi_i^{y_i} (1 - \pi_I)^{1-y_i} \quad (3)$$

dimana: π_i = peluang kejadian ke-i
y_i= peubah acak ke-i yang terdiri dari 0 dan 1

Bentuk model regresi logistik dengan satu variabel prediktor adalah [6]:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \quad (4)$$

Untuk mempermudah menaksir parameter regresi, maka $\pi(x)$ pada persamaan diatas ditransformasikan sehingga menghasilkan bentuk logit regresi logistik, sebagai berikut:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \quad (5)$$

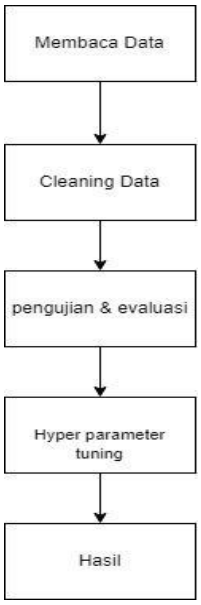
c. Random Forest

Random Forest merupakan salah satu pengembangan dari Decision Tree yang menggunakan beberapa Decision Tree, yang mana setiap decision tree telah dilakukan

pelatihan dari sampel individu dan setiap atribut dipecah pada pohon yang dipilih antara atribut subset yang bersifat acak [6]. Random Forest merupakan salah satu metode yang memiliki tingkat akurasi cukup tinggi dan dapat menangani nilai yang hilang atau fitur yang tidak berkaitan. Metode ini juga dapat bekerja baik pada data yang besar dengan parameter yang kompleks.

2.3 Pembuatan Sistem

Pada pembuatan sistem ini, dilakukan melalui beberapa tahap yaitu membaca dataset, cleaning Dataset, pengujian dan evaluasi, Hyperparameter Tuning, dan hasil. Alur pembuatan sistem dapat dilihat pada gambar 1.



Gambar 1. Alur Pembuatan Sistem

a. Membaca Data

Tahap ini merupakan tahapan import dataset pada sistem, kemudian sistem akan membaca informasi yang ada pada dataset tersebut.

b. Cleaning Data

Pada tahapan ini, dilakukan proses pembersihan sekaligus mempersiapkan dataset sebelum dilakukan analisis ataupun pemodelan. Tujuan dari tahap ini yaitu untuk menghilangkan data yang terduplikat, memperbaiki data yang hilang, tidak lengkap, dan juga tidak relevan [7].

c. Pengujian dan Evaluasi Dataset

Tahapan ini dilakukan proses pengukuran kinerja model terhadap data yang sudah dicari. Tujuan pada tahapan ini yaitu untuk mencari nilai akurasi pada sistem yang dibuat.

d. Hyperparameter Tuning

Hyperparameter Tuning merupakan tahapan dimana nilai-nilai hiperparameter dioptimalisasi untuk mendapatkan kinerja model yang baik. Hiperparameter ditentukan sebelum training pada model [8].

e. Hasil

Pada tahap ini merupakan output dari data yang telah diproses dari sistem.

3. Hasil dan Pembahasan

Pada penelitian sistem pendeteksi penyakit diabetes pada manusia ini dibuat menggunakan bahasa python. Pada penelitian ini mengambil data berupa data numerik. Pengkodean sistem diambil dari

3.1 Hasil menggunakan Metode Gaussian Naive Bayes

Nilai precision, recall, dan f-1 score menggunakan metode ini adalah 0.83; 0.79; 0.81 pada angka 0 (tidak diabetes) dan 0.58; 0.65; 0.61 pada angka 1 (diabetes). Hasil precision, recall, dan f-1 score dari penggunaan Metode Gaussian Naive Bayes menunjukan data diuji negatif terkena diabetes.

	precision	recall	f1-score	support
0	0.83	0.79	0.81	150
1	0.58	0.65	0.61	68
accuracy			0.74	218
macro avg	0.70	0.72	0.71	218
weighted avg	0.75	0.74	0.75	218

Gambar 2. Hasil uji menggunakan metode Naive Bayes

3.2 Hasil menggunakan Metode Regresi Logistik

Nilai precision, recall, dan f-1 score menggunakan metode ini adalah 0.81; 0.88; 0.85 pada angka 0 (tidak diabetes) dan 0.68; 0.56; 0.61 pada angka 1 (diabetes). Hasil precision, recall, dan f-1 score dari penggunaan Metode Gaussian Naive Bayes menunjukan data diuji negatif terkena diabetes.

	precision	recall	f1-score	support
0	0.81	0.88	0.85	150
1	0.68	0.56	0.61	68
accuracy			0.78	218
macro avg	0.75	0.72	0.73	218
weighted avg	0.77	0.78	0.77	218

Gambar 2. Hasil uji menggunakan metode Regresi Logistik

3.3 Hasil menggunakan Metode Random Forest

Nilai precision, recall, dan f-1 score menggunakan metode ini adalah 0.82; 0.87; 0.84 pada angka 0 (tidak diabetes) dan 0.66; 0.57; 0.61 pada angka 1 (diabetes). Hasil precision, recall, dan f-1 score dari penggunaan Metode Gaussian Naive Bayes menunjukan data diuji negatif terkena diabetes.

	precision	recall	f1-score	support
0	0.82	0.87	0.84	150
1	0.66	0.57	0.61	68
accuracy			0.78	218
macro avg	0.74	0.72	0.73	218
weighted avg	0.77	0.78	0.77	218

Gambar 3. Hasil uji menggunakan metode Random Forest

4. Kesimpulan

Pada penelitian ini dapat disimpulkan bahwa tingkat akurasi pada metode Gaussian Naive Bayes, Regresi Logistik, Random Forest pada data yang diuji yaitu sebesar 0.74; 0;78 ; 078. Berdasarkan nilai akurasi pada data yang diuji metode Gaussian Naive Bayes memiliki akurasi paling rendah, sementara Regresi Logistik dan Random Forest memiliki nilai akurasi yang sama.

Daftar Pustaka

[1] C. J. Hinde, R. Stone and X. Daniela, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," international Journal of Computer Science, vol. 1, no. 4, September 2009.

[2] D. I. Saputra and D. L. Hakim, "ImplementasiAlgoritma Gaussian Naive Bayes Classifier Untuk Prediksi Potensi Tsunami Berbasis Mikrokontroler," Journal of Electrical Engineering and Information Technology, vol. 2, no. 20, December 2022.

[3] J. P. Sianipar, R. E. S. and C. S. , "Waves with Multi-Sensor System Based on Web Application Using Naive Bayes Algorithm," e-Proceeding of Engineering, vol. 9, no. 5, pp. 6183-6188, 2021.

[4] H. D. and S. L. , "Applied Logistic Regression," vol. 2, 2000.

[5] A. A, "Categorical Data Analysis John Wiley and Sons," 1990.

[6] R. Supriyadi, W. Gata, N. Maulidah and A. fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," Jurnal Ilmiah Ekonomi dan Bisnis, vol. 13, no. 2, pp. 67-75, December 2020.

[7] L. and J. , "Data Cleaning in Python : the Ultimate Guide," 4 Febuary 2020.

[8] S. Paul, "Hyperparameter Optimization in Machine Learning Models," August 2018.

Halaman ini sengaja dibiarkan kosong