Perbandingan Algoritma Klasifikasi untuk Prediksi Penyakit Diabetes Menggunakan Orange

Bayu Nurcahyono
Magister Teknik Informatika, Universitas Pamulang
E-Mail: mirmatheorara@gmail.com

***Abstrak*— Diabetes mellitus adalah penyakit kronis yang mempengaruhi jutaan orang di seluruh dunia, dan prediksi dini memiliki dampak besar dalam pencegahan dan pengelolaan penyakit ini. Penelitian ini bertujuan untuk membandingkan kinerja beberapa algoritma klasifikasi yang berbeda dalam prediksi penyakit diabetes. Metode penelitian ini meliputi tiga tahap utama: pengumpulan data, pembuatan model, dan evaluasi. Pembuatan model dilakukan menggunakan aplikasi Orange, dengan empat algoritma klasifikasi yang dibandingkan, yaitu Logistic Regression, kNN, Naive Bayes, dan Decision Tree. Data yang digunakan dalam penelitian ini mencakup atribut Glucose, Age, BMI, Insulin, dan Pregnancies. Berdasarkan dataset diabetes dari Kaggle, model Logistic Regression menunjukkan performa terbaik dengan akurasi 76.7%, sedangkan model kNN memiliki performa terendah dengan akurasi 70.3%.**

***Kata Kunci*— Prediksi diabetes, kNN, Logistic Regression, Naive Bayes, Decision Tree.**

I. PENDAHULUAN

Diabetes mellitus, juga dikenal sebagai diabetes, merupakan masalah kesehatan global yang terus berkembang. Menurut data Organisasi Kesehatan Dunia, pada tahun 2019, lebih dari 420 juta orang di seluruh dunia menderita diabetes, dan jumlah ini diperkirakan akan terus meningkat dalam beberapa dekade mendatang [1].

Diabetes merupakan penyakit kronis yang mempengaruhi cara tubuh mengatur kadar gula darah dan jika tidak dikontrol dengan baik dapat menyebabkan banyak komplikasi serius, seperti kerusakan ginjal, penyakit jantung dan komplikasi lainnya. Memprediksi penyakit diabetes penting dalam upaya pencegahan dan pengendalian penyakit ini. Dengan prediksi yang akurat, individu berisiko tinggi dapat diidentifikasi dengan cepat sehingga intervensi medis dan perubahan gaya hidup yang tepat dapat diterapkan [2]. Pendekatan yang efektif untuk memprediksi diabetes adalah dengan menggunakan algoritma yang mengklasifikasikan data klinis, termasuk atribut seperti glukosa, usia, BMI (indeks massa tubuh), insulin, dan riwayat kehamilan pada wanita [3].

Penelitian ini bertujuan untuk membandingkan kinerja berbagai algoritma klasifikasi dalam konteks prediksi diabetes. Kami akan menganalisis algoritma seperti Naive Bayes, Decision Trees, Logistic Regression, dan kNN.Hasil penelitian ini diharapkan dapat memberikan informasi berharga bagi para profesional kesehatan, peneliti dan praktisi kesehatan dalam memilih algoritma yang paling tepat untuk memprediksi diagnosis diabetes berdasarkan karakteristik pasien. Seiring dengan kemajuan teknologi dan pemrosesan data, pendekatan ini dapat menjadi alat penting dalam memerangi diabetes dan meningkatkan kualitas hidup pasien.

II. METODE PENELITIAN

Metode yang digunakan dalam tulisan ini adalah: (1) pengumpulan data (2) pembuatan model dan pengukuran metrik di aplikasi Orange (3) evaluasi model.

A. Pengumpulan Data

Kumpulan data ini berasal dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal[3]. Tujuan dari kumpulan data ini adalah untuk memprediksi secara diagnostik apakah seorang pasien menderita diabetes atau tidak, berdasarkan pengukuran diagnostik tertentu yang disertakan dalam kumpulan data tersebut. Beberapa kendala ditempatkan pada pemilihan contoh-contoh ini dari database yang lebih besar. Secara khusus, semua pasien di sini adalah wanita berusia minimal 21 tahun keturunan India Pima. Data yang digunakan sudah dalam kondisi bersih dan tervalidasi, sehingga tidak diperlukan tahapan pemrosesan data.

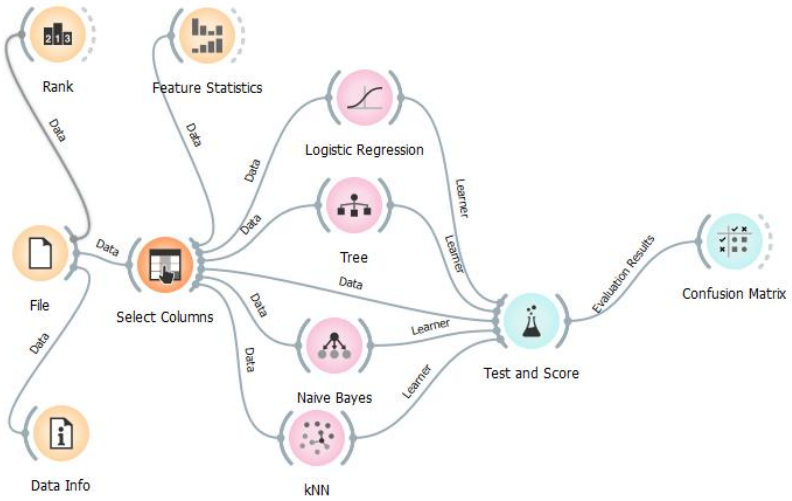
Jumlah sampel sebanyak 768 data dan terdiri dari 8 atribut yaitu: Glucose, Age, BMI, Insulin, Pregnancies, Skin Thickness, Diabetes Pedigree Function, dan Blood Pressure. Setelah dilakukan seleksi atribut, hanya akan diambil 5 besar atribut yang paling berpengaruh yaitu: Glucose, Age, BMI, Insulin, dan Pregnancies berdasarkan *Information Gain, Gini, Gain Ratio, dan ANOVA.*

		#	Info. gain	Gain ratio	Gini	ANOVA
1	Glucose		0.170	0.085	0.101	213.162
2	Age		0.081	0.041	0.048	46.141
3	BMI		0.079	0.039	0.044	71.772
4	Insulin		0.055	0.030	0.031	13.281
5	Pregnancies		0.043	0.021	0.028	39.670
6	SkinThickness		0.036	0.018	0.022	4.304
7	DiabetesPedigreeFunction		0.022	0.011	0.014	23.871
8	BloodPressure		0.015	0.007	0.009	3.257

Gambar 1.
Atribut yang mempengaruhi diabetes

B. Pembuatan Model di Aplikasi Orange dan Pengukuran Metrik

Algoritma yang akan dibandingkan adalah: (1) Logistic Regression (2) kNN (3) Naive Bayes (4) Decision Tree. *Logistic Regression* adalah algoritma pembelajaran mesin yang digunakan untuk memecahkan masalah klasifikasi dengan memodelkan probabilitas kejadian suatu peristiwa dengan memperhatikan faktor-faktor prediktor yang berkaitan. Algoritma *kNN* adalah algoritma klasifikasi yang mendasarkan pada mayoritas kategori dari k- tetangga terdekat. Algoritma *Naive Bayes* adalah algoritma yang mempelajari probabilitas suatu objek dengan ciri-ciri tertentu yang termasuk dalam kelompok/kelas tertentu. *Decision Tree* merupakan salah satu cara data processing dalam memprediksi masa depan dengan cara membangun klasifikasi atau regresi model dalam bentuk struktur pohon.



Gambar 2.
Model Aplikasi Data Orange

Metrik yang dipergunakan untuk pengukuran adalah *Accuracy*. Metrik ini mengukur % kebenaran prediksi tiap algoritma terhadap jumlah keseluruhan data. Rumus dari Accuracy ini adalah:

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

C. Analisa metrik

Analisa Metrik dilakukan dengan merangking algoritma mana yang memiliki tingkat Accuracy dari yang tertinggi sampai terendah.

III. HASIL DAN PEMBAHASAN

A. Confusion Matrix

Dari model yang dibangun dengan mempergunakan masing-masing algoritma, didapatkan confusion matrix sebagai berikut.

Tabel 2.
Model *Confussion Matrix*

Logistic Regression

Actual \ Predicted	0	1	Σ
0	441	59	500
1	120	148	268
Σ	561	207	768

Naive Bayes

Actual \ Predicted	0	1	Σ
0	389	111	500
1	78	190	268
Σ	467	301	768

Decision Tree

Actual \ Predicted	0	1	Σ
0	396	104	500
1	123	145	268
Σ	519	249	768

kNN

Actual \ Predicted	0	1	Σ
0	393	107	500
1	121	147	268
Σ	514	254	768

B. Metrik

Metrix yang dihasilkan adalah sebagai berikut:

Model	AUC	CA	F1	Prec	Recall
Logistic Regression	0.827	0.767	0.759	0.761	0.767
Naive Bayes	0.828	0.754	0.757	0.763	0.754
Tree	0.657	0.704	0.702	0.700	0.704
kNN	0.763	0.703	0.701	0.700	0.703

Gambar 3.
Model *Test and Score*

Dari metriks di atas didapatkan bahwa jika diurutkan berdasarkan % kebenaran prediksi / *Accuracy*/ CA yang dihasilkan masing-masing model:

a. #1 adalah Logistic Regression dengan nilai CA 0.767 atau 76.7%
b. #2 adalah Naive Bayes dengan nilai CA 0.754 atau 75.4%
c. #3 adalah Decision Tree dengan nilai CA 0.704 atau 70.4 %
d. #4 adalah kNN dengan nilai CA 0.703 atau 70.3%

IV. KESIMPULAN

Kesimpulan yang diambil dalam penelitian ini adalah:

1. Tidak semua algoritma klasifikasi memberikan nilai Accuracy yang sama untuk domain permasalahan yang sama.
2. Model terbaik dalam memprediksi penyakit diabetes berdasarkan dataset dari kaggle adalah Logistic Regression dengan nilai Accuracy 76.7%.
3. Model terburuk dalam memprediksi penyakit diabetes berdasarkan dataset dari kaggle adalah kNN dengan nilai Accuracy 70.3%.

DAFTAR PUSTAKA

[1] WHO, “Diabetes,” https://www.who.int/news-room/fact-sheets/detail/diabetes.
[2] O. Emilia, Y. S. Prabandari, and Supriyati, *Promosi Kesehatan dalam Lingkup Kesehatan Reproduksi*. Yogyakarta: UGM PRESS, 2019.
[3] A. Pramudyantoro, E. Utami, and D. Ariatmanto, “Penggabungan K-Nearest Neighbors Dan LightGBM Untuk Prediksi Diabetes Pada Dataset Pima Indians: Menggunakan Pendekatan Exploratory Data Analysis,” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, pp. 1133–1144, 2024, doi: 10.29100/jipi.v9i3.4966.