



Prediksi Penyakit Diabetes Menggunakan Algoritma *Naive Bayes* Berbasis *Data Mining* dengan Software Orange

Fitri Yanti¹, Abdul Majid², Abelli Fikri³, Alvis Juliandry⁴, Ardiansyah⁵, Fikri Ariansyah⁶

¹⁻⁶Teknik Informatika, Universitas Pamulang, Jl. Raya Puspitex no. 10, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten, Indonesia, 15310
Email: ²abdulmajid08180881@gmail.com

Abstrak–Diabetes mellitus adalah penyakit kronis dengan prevalensi yang terus meningkat secara global dan memerlukan deteksi dini untuk mencegah komplikasi serius. Penelitian ini bertujuan mengembangkan model prediksi diabetes menggunakan algoritma *Naive Bayes* dengan pendekatan *data mining*. Dataset yang digunakan adalah *Pima Indians Diabetes Database* yang terdiri dari 768 *instances* dengan 8 fitur medis. Metode penelitian meliputi *preprocessing* data (penanganan *missing values* dan normalisasi) serta implementasi algoritma menggunakan *software Orange Data Mining*. Evaluasi model menggunakan *10-Fold Cross-Validation* menghasilkan akurasi sebesar 76.3%, *precision* 71.2%, *recall* 68.5%, dan *F1-score* 69.8%. Nilai AUC (*Area under Curve*) mencapai 0.82, mengindikasikan kemampuan diskriminasi model yang sangat baik (*excellent*). Model ini berpotensi membantu tenaga medis dalam melakukan *screening* awal pasien berisiko diabetes.

Kata Kunci: data mining, diabetes, Naive Bayes, Orange, prediksi

Abstract–*Diabetes mellitus is a chronic disease with an increasing global prevalence and requires early detection to prevent serious complications. This study aims to develop a diabetes prediction model using the Naive Bayes algorithm with a data mining approach. The dataset used is the Pima Indians Diabetes Database, which consists of 768 instances with 8 medical features. The research methods include data preprocessing (handling missing values and normalization) and algorithm implementation using the Orange Data Mining software. Model evaluation using 10-Fold Cross-Validation resulted in an accuracy of 76.3%, precision of 71.2%, recall of 68.5%, and an F1-score of 69.8%. The AUC (Area under Curve) value reached 0.82, indicating excellent discriminatory ability. This model has the potential to assist medical personnel in conducting early screening for patients at risk of diabetes.*

Keywords: data mining, diabetes, Naive Bayes, Orange, prediction

1. PENDAHULUAN

Diabetes mellitus merupakan penyakit metabolik kronis yang ditandai dengan peningkatan kadar gula darah. Menurut WHO, penyakit ini menjadi salah satu penyebab utama kematian di dunia. Deteksi dini sangat krusial karena diabetes yang tidak tertangani dapat memicu komplikasi seperti penyakit jantung, gagal ginjal, dan stroke. Namun, diagnosis konvensional seringkali memakan waktu dan biaya. Oleh karena itu, diperlukan sistem prediksi berbasis komputer untuk membantu *screening* awal.

Data mining menawarkan solusi melalui ekstraksi pola dari kumpulan data besar. Salah satu teknik klasifikasi yang populer adalah *Naive Bayes*, yang dikenal memiliki kecepatan komputasi tinggi dan performa baik meskipun dengan jumlah data terbatas. Penelitian ini menggunakan *Pima Indians Diabetes Database*, dataset standar yang telah teruji validitasnya dalam penelitian *machine learning* medis. Tujuan penelitian adalah mengimplementasikan dan mengevaluasi performa algoritma *Naive Bayes* menggunakan *software Orange Data Mining* untuk memprediksi status diabetes.

2. METODOLOGI PENELITIAN

2.1 Dataset

Data diperoleh dari Kaggle (*Pima Indians Diabetes Database*) yang bersumber dari *National Institute of Diabetes and Digestive and Kidney Diseases*. Dataset terdiri dari 768 data pasien wanita keturunan Pima Indian berusia minimal 21 tahun. Terdapat 8 variabel prediktor (*Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age*) dan 1



variabel target (*Outcome*: 0 untuk non-diabetes, 1 untuk diabetes). Distribusi kelas tidak seimbang dengan 500 data non-diabetes dan 268 data diabetes.

2.2 Algoritma Naïve Bayes

Metode yang digunakan adalah Gaussian Naive Bayes, varian yang cocok untuk fitur kontinu dengan asumsi distribusi normal. Algoritma ini menghitung probabilitas posterior berdasarkan Teorema Bayes dengan asumsi independensi antar fitur. Rumus dasar yang digunakan adalah:

$$P(C|X) = P(X|C) \times P(C) / P(X)$$

Dimana:

- $P(C|X)$ = Probabilitas posterior (probabilitas kelas C given data X)
- $P(X|C)$ = Likelihood (probabilitas data X given kelas C)
- $P(C)$ = Prior probability (probabilitas awal kelas C)
- $P(X)$ = Evidence (probabilitas data X)

2.3 Preprocessing Data

Sebelum pemodelan, dilakukan tahapan *preprocessing*:

1. Penanganan *Missing Values*: Nilai 0 pada fitur seperti *Glucose* dan *BMI* dianggap sebagai *missing value* dan diganti menggunakan teknik *mean imputation*.
2. Normalisasi: Menggunakan *Min-Max Normalization* untuk menyamakan skala fitur agar algoritma bekerja optimal.

2.4 Implementasi dan Evaluasi

Implementasi dilakukan menggunakan *Orange Data Mining* versi 3 dengan *workflow* visual. Evaluasi model menggunakan metode *10-Fold Cross-Validation* untuk menghindari *overfitting*. Metrik evaluasi meliputi Akurasi, *Precision*, *Recall*, *F1-Score*, dan AUC.

Metrik	Nilai	Presentase	Interpretasi
AUC	0.818	81.8%	Excellent discrimination
CA	0.736	73.6%	Akurasi Baik
F1	0.739	73.9%	Balance precision-recall
Precision	0.745	74.5%	Tingkat False positive rendah
Recall	0.736	73.6%	Sensitivity cukup baik
MCC	0.436	-	Korelasi positif sedang

3. ANALISA DAN PEMBAHASAN

3.1 Eksplorasi Data

Analisis korelasi menunjukkan bahwa *Glucose* memiliki korelasi positif tertinggi (0.47) terhadap target diabetes, diikuti oleh BMI (0.29) dan Usia (0.24). Visualisasi *scatter plot* juga memperlihatkan bahwa penderita diabetes cenderung berkumpul pada area dengan kadar glukosa dan BMI tinggi.

3.2 Performa Model

Berdasarkan pengujian *10-Fold Cross-Validation*, diperoleh hasil performa model sebagai berikut:

Kelebihan Model:

1. AUC Excellent (81.8%): Kemampuan diskriminasi sangat baik



JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 8, Januari Tahun 2026
ISSN 3025-0919 (media online)
Hal 2194-2196

2. Akurasi Baik (73.6%): Lebih dari 7 dari 10 prediksi benar
3. Balance Metrics: Precision dan Recall seimbang (~74%)
4. Konsistensi: Hasil stabil antara 5-fold dan 10-fold
5. Efisiensi: Cepat dalam training dan prediksi
6. Interpretability: Mudah dijelaskan secara probabilistik

Kelemahan Model:

1. Recall 73.6%: 26.4% pasien diabetes tidak terdeteksi
2. False Negative Rate tinggi: Berbahaya dalam konteks medis
3. Asumsi Independensi: Naive Bayes asumsikan fitur independen
4. Akurasi lebih rendah: Dibanding Random Forest/SVM

3.3 Analisis Hasil

Meskipun akurasi cukup tinggi, nilai *Recall* sebesar 68.5% - 73.6% menunjukkan masih adanya kasus *false negative* (pasien diabetes terdeteksi sehat). Dalam konteks medis, hal ini menjadi limitasi karena pasien berisiko tidak terdeteksi dini. Namun, dengan *precision* 74.5%, model memiliki tingkat *false alarm* yang dapat diterima untuk sistem *screening*. Analisis *Feature Importance* mengonfirmasi bahwa *Glucose* adalah kontributor terkuat dalam prediksi, sejalan dengan fakta medis.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma *Naive Bayes* untuk prediksi diabetes menggunakan *Orange Data Mining* dengan hasil yang baik. Model menghasilkan akurasi 76.3% dan AUC 0.82, membuktikan bahwa algoritma sederhana ini efektif untuk *screening* awal. *Glucose* dan *BMI* teridentifikasi sebagai faktor risiko utama. Meskipun demikian, model ini memiliki *recall* moderat sehingga disarankan penggunaannya sebagai sistem pendukung keputusan (*decision support*), bukan pengganti diagnosis medis. Penelitian selanjutnya disarankan menggunakan metode *Ensemble* atau *Deep Learning* serta penambahan fitur gaya hidup untuk meningkatkan sensitivitas model.

REFERENCES

- Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research*, 14, 2349–2353.
- Han, J. W., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.
- Kaggle. (2016). Pima Indians Diabetes Database. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, 104–116. <https://doi.org/10.1016/j.csbj.2016.12.005>
- Orange Data Mining. (2024). Naive Bayes classifier documentation.
- Rish, I. (2001). An empirical study of the Naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 41–46.
- World Health Organization. (2023). Diabetes. <https://www.who.int/news-room/fact-sheets/detail/diabetes>