Analisis Perbandingan Algoritma Decision Tree, K-NN, Naive Bayes, dan XGBoost Memprediksi Penyakit Diabetes Menggunakan Data Behavioral Risk Factor Surveillance System

Lily Pebriana. S[1], Fitra Salam S. Nagalay[2], Ubaidillah Ubaidillah[3]
[1,2]Program Studi Sistem Informasi, Universitas Wira Buana
[3]Program Studi Ilmu Komputer, Universitas Sumatera Selatan
Email : [1*]pebriana.lily@gmail.com, [2]fitrasalamsn@gmail.com, [3]ubai@uss.ac.id

Abstrak

Penyakit diabetes melitus merupakan masalah kesehatan global yang memerlukan deteksi dini secara akurat. Penelitian ini bertujuan untuk membandingkan kinerja empat algoritma *Machine Learning*, *Decision Tree, K-Nearest Neighbor* (K-NN), *Naive Bayes*, dan *XGBoost* dalam memprediksi risiko diabetes berdasarkan indikator gaya hidup. Data yang digunakan bersumber dari dataset sekunder *Behavioral Risk Factor Surveillance System* (BRFSS) tahun 2015 yang memiliki karakteristik kelas tidak seimbang (*imbalanced data*). Evaluasi model dilakukan menggunakan pembagian data uji sebesar 20%, dengan metrik pengujian yang mencakup Akurasi, *Precision, Recall, F1-Score*, serta analisis *Confusion Matrix*. Hasil penelitian menunjukkan bahwa algoritma *XGBoost* mencapai tingkat akurasi keseluruhan tertinggi sebesar 86,71%. Namun, akibat ketidakseimbangan dataset, model tersebut cenderung gagal mendeteksi kelas minoritas sehingga menghasilkan nilai *Recall* yang sangat rendah (17,80%). Sebaliknya, algoritma *Naive Bayes* dengan tingkat akurasi 77,20% justru berhasil mencatatkan kinerja diagnostik terbaik, dibuktikan dengan nilai *Recall* tertinggi (56,55%) dan *F1-Score* (40,62%). Kesimpulan dari penelitian ini menegaskan bahwa untuk tujuan skrining medis, di mana kesalahan *False Negative* sangat berisiko fatal, *Naive Bayes* merupakan model klasifikasi yang paling fungsional dan dapat diandalkan dibandingkan model yang sekadar mengandalkan akurasi tinggi.

Kata kunci: *Diabetes, Klasifikasi, Machine Learning, Imbalanced Data, Naive Bayes, XGBoost*

1. Pendahuluan

Penyakit diabetes melitus merupakan salah satu masalah kesehatan masyarakat global yang prevalensinya terus meningkat setiap tahun. Diagnosis dini dan identifikasi faktor risiko gaya hidup sangat krusial untuk mencegah komplikasi yang lebih parah. Seiring dengan perkembangan teknologi Sistem Informasi di bidang kesehatan, pemanfaatan Machine Learning telah menjadi pendekatan yang efektif dalam menggali pola tersembunyi dari rekam medis maupun data survei kesehatan skala besar.

Penelitian ini memanfaatkan Behavioral Risk Factor Surveillance System (BRFSS), sebuah sistem pengumpulan data survei kesehatan berbasis telepon terbesar di dunia yang dikelola oleh Centers for Disease Control and Prevention (CDC). Meskipun data survei memberikan gambaran komprehensif terkait perilaku pasien, tantangan utama dalam menganalisis data ini adalah tingginya dimensi fitur dan ketidakseimbangan kelas (imbalanced data) antara pasien sehat dan pasien terdiagnosis diabetes.

Penelitian-penelitian sebelumnya

umumnya hanya menerapkan algoritma tunggal atau berfokus pada algoritma klasifikasi klasik dalam mendeteksi diabetes. Untuk mengisi celah penelitian tersebut, studi ini melakukan analisis perbandingan kinerja secara komprehensif antara tiga algoritma klasifikasi klasik yang mewakili pendekatan berbeda, yakni Decision Tree (berbasis pohon keputusan), K-Nearest Neighbor (berbasis metrik jarak), dan Naive Bayes (berbasis probabilitas), dengan algoritma modern berbasis ensemble learning, yaitu Extreme Gradient Boosting (XGBoost). Tujuan utama penelitian ini adalah untuk mengevaluasi model klasifikasi mana yang paling optimal, kuat, dan akurat dalam memprediksi risiko diabetes berdasarkan indikator gaya hidup.

2. Kajian Pustaka

Penelitian mengenai penerapan Machine Learning untuk deteksi dini penyakit diabetes telah banyak dilakukan dengan berbagai pendekatan algoritma dan variasi dataset. Sebagian besar penelitian terdahulu berfokus pada penggunaan algoritma klasifikasi tunggal atau komparasi sesama algoritma klasik.

Algoritma Decision Tree, K-Nearest Neighbor (K-NN), dan Naive Bayes merupakan metode yang paling umum dievaluasi dalam ranah informatika medis karena kemudahan interpretasinya. Penelitian oleh Hasan dkk. [1] menunjukkan bahwa Naive Bayes memiliki keunggulan dalam kecepatan komputasi dan sensitivitas klinis saat menangani dataset rekam medis yang memiliki atribut probabilitas independen. Sementara itu, studi komparatif yang dilakukan oleh Rahman dan Utama [2] menyimpulkan bahwa algoritma K-NN sangat bergantung pada teknik normalisasi data, di mana standarisasi fitur kardiovaskular secara signifikan meningkatkan akurasi deteksi diabetes.

Dalam beberapa tahun terakhir, tren penelitian mulai bergeser ke arah penggunaan algoritma berbasis ensemble learning untuk meningkatkan performa prediktif. Chen dkk. [3] mengimplementasikan Extreme Gradient Boosting (XGBoost) pada dataset kesehatan publik dan melaporkan bahwa XGBoost secara konsisten mengungguli algoritma berbasis pohon keputusan konvensional dalam hal akurasi keseluruhan (overall accuracy) dan Area Under the Curve (AUC). Hal senada juga ditemukan oleh Singh dkk. [4] yang menyoroti ketangguhan algoritma boosting dalam menangani data yang memiliki banyak noise atau nilai yang hilang (missing values).

Meskipun demikian, tantangan utama dalam klasifikasi data kesehatan berskala masif, seperti Behavioral Risk Factor Surveillance System (BRFSS), adalah distribusi kelas yang sangat tidak seimbang (imbalanced data). Analisis yang dilakukan oleh Wibowo [5] pada data survei kesehatan menyoroti fenomena paradoks akurasi, di mana model Machine Learning dapat mencapai akurasi di atas 85% dengan sekadar memprediksi kelas mayoritas (pasien sehat), namun berujung pada nilai False Negative yang fatal secara medis.Jangan lupa melakukan sitasi seperti sebelumnya [2].

3. Metode Penelitian

Penelitian ini menggunakan kerangka kerja Knowledge Discovery in Databases (KDD) yang meliputi pemahaman domain data [6], Exploratory Data Analysis (EDA), pra-pemrosesan, pemodelan, hingga evaluasi hasil.

3.1 Dataset

Data yang digunakan dalam penelitian ini adalah dataset sekunder "Diabetes Health Indicators" yang bersumber dari survei BRFSS tahun 2015 [7]. Dataset ini terdiri dari 253.680 baris data observasi dan 23 atribut (kolom), yang mencakup 22 variabel independen terkait indikator gaya hidup dan rekam medis dasar, serta 1 variabel dependen (Diabetes_binary) sebagai target klasifikasi.

3.2 Exploratory Data Analysis (EDA)

Dataset ini memiliki distribusi kelas

yang sangat tidak seimbang. Terdapat 86,1% pasien yang tidak mengidap diabetes (kelas mayoritas) berbanding 13,9% pasien yang terdiagnosis diabetes (kelas minoritas). Kondisi ini akan menjadi perhatian khusus dalam proses evaluasi model, di mana metrik seperti Precision dan Recall akan lebih relevan dibandingkan sekadar Akurasi.

Distribusi Indeks Massa Tubuh (BMI): Analisis kepadatan distribusi menunjukkan bahwa pasien yang terdiagnosis diabetes memiliki kecenderungan nilai BMI yang lebih tinggi (bergeser ke arah kanan/obesitas) dibandingkan dengan kelompok pasien sehat.

Analisis Korelasi Fitur: Berdasarkan matriks korelasi heatmap terhadap 10 fitur teratas, atribut yang memiliki pengaruh linear positif paling kuat terhadap variabel target diabetes adalah kondisi Kesehatan Umum atau GenHlth (r = 0,29), Riwayat Tekanan Darah Tinggi atau HighBP (r = 0,26), Kesulitan Berjalan atau DiffWalk (r = 0,22), dan BMI (r = 0,22). Sebaliknya, tingkat pendidikan (Education) dan pendapatan (Income) memiliki korelasi negatif, yang mengindikasikan bahwa kelas sosial ekonomi yang lebih tinggi memiliki risiko diabetes yang sedikit lebih rendah.

3.3 Pra-pemrosesan Data

Berdasarkan hasil pengecekan, tidak ditemukan adanya data yang hilang (missing values) pada dataset. Proses selanjutnya adalah pembagian data (data splitting) menjadi data latih (training set) sebesar 80% dan data uji (testing set) sebesar 20%. Mengingat algoritma seperti K-NN sangat sensitif terhadap skala rentang nilai antar atribut, dilakukan standardisasi data menggunakan StandardScaler sehingga seluruh fitur memiliki skala distribusi yang seragam dengan nilai mean 0 dan standar deviasi 1.

3.4 Skenario Pemodelan

Empat algoritma klasifikasi dilatih menggunakan training set yang telah distandardisasi. Parameter yang digunakan meliputi [8]:

- Decision Tree: Menggunakan metrik pemisahan standar berbasis gini impurity.
- K-Nearest Neighbor (K-NN): Diatur pada jumlah ketetanggaan terdekat K=5.
- Naive Bayes: Menggunakan pendekatan Gaussian Naive Bayes yang mengasumsikan distribusi fitur secara normal.
- XGBoost: Menggunakan pengaturan standar dengan fungsi objektif logloss untuk klasifikasi biner.

3.5 Evaluasi Kinerja Model

Mengingat kondisi dataset yang tidak seimbang, evaluasi kinerja model tidak hanya mengandalkan matriks Confusion Matrix dan tingkat Akurasi. Evaluasi juga dilakukan dengan mengukur Precision (kemampuan model meminimalkan False Positive), Recall (kemampuan model meminimalkan False Negative), dan F1-Score (rata-rata harmonis dari Precision dan Recall).

4. Hasil dan Pembahasan

Penelitian ini membandingkan empat algoritma klasifikasi (Decision Tree, K-Nearest Neighbor, Naive Bayes, dan XGBoost) menggunakan data uji sebesar 20% dari total dataset survei BRFSS. Evaluasi kinerja tidak hanya dititikberatkan pada tingkat akurasi secara keseluruhan, melainkan juga memperhatikan kemampuan model dalam mengenali kelas minoritas (pasien diabetes) melalui metrik Precision, Recall, dan F1-Score, serta analisis Confusion Matrix.

Pengujian model menghasilkan metrik kinerja yang bervariasi untuk masing-masing algoritma. Ringkasan komparasi kinerja keempat algoritma dapat dilihat pada Tabel 1.

Tabel 1. Komparasi Kinerja Algoritma Klasifikasi

Algoritma	*Accuracy*	*Precision*	Recall
XGBoost	0.8671 (86.71%)	0.5567	0.1780
K-Nearest Neighbor	0.8468 (84.68%)	0.3952	0.2096
Decision Tree	0.7986 (79.86%)	0.2963	0.3351
Naive Bayes	0.7720 (77.20%)	0.3169	0.5655

Merujuk pada Tabel 1 dan grafik perbandingan akurasi, algoritma boosting modern, yaitu XGBoost, menunjukkan dominasi yang signifikan pada metrik akurasi dengan nilai tertinggi mencapai 86.71%, disusul oleh K-NN sebesar 84.68%. Sebaliknya, algoritma klasik Naive Bayes mencatatkan akurasi terendah di angka 77.20%.

Meskipun demikian, evaluasi komprehensif pada dataset yang memiliki ketidakseimbangan kelas (imbalanced data) tidak dapat disandarkan pada akurasi semata. Perhatian khusus harus diberikan pada nilai Recall (sensitivitas), yang merepresentasikan persentase pasien diabetes yang berhasil dideteksi dengan benar oleh sistem dari keseluruhan pasien yang aktualnya menderita diabetes. Pada metrik ini, Naive Bayes justru mengungguli seluruh algoritma lain dengan nilai Recall tertinggi sebesar 56.55% dan F1-Score tertinggi sebesar 40.62%.

Untuk membedah lebih dalam mengenai letak kesalahan prediksi tiap algoritma, analisis Confusion Matrix dilakukan. Visualisasi matriks kebingungan dari keempat model mengilustrasikan distribusi True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

- XGBoost: Model ini sangat unggul dalam mengklasifikasikan kelas mayoritas. Hal ini dibuktikan dengan jumlah TN tertinggi (42.747 kasus sehat diprediksi benar). Model ini sangat berhati-hati dalam memberikan label diabetes, sehingga nilai FP sangat rendah (hanya 992). Namun, dampaknya adalah model ini gagal mendeteksi mayoritas pasien yang sebenarnya sakit, ditunjukkan dengan angka FN yang sangat tinggi (5.751 kasus).
- K-Nearest Neighbor (K-NN): Kinerja K-NN berada di tengah-tengah. Model ini berhasil menebak 41.494 TN dan 1.467 TP, dengan tingkat kesalahan FN mencapai 5.530 kasus.
- Decision Tree: Model berbasis pohon keputusan ini mulai menunjukkan peningkatan dalam mendeteksi kelas minoritas dengan memprediksi benar 2.345 kasus diabetes (TP). Konsekuensinya, jumlah tebakan salah pada orang sehat (FP) meningkat menjadi 5.567 kasus.
- Naive Bayes: Algoritma probabilitas ini menunjukkan karakteristik yang bertolak belakang dengan XGBoost. Naive Bayes berhasil mengidentifikasi jumlah pasien diabetes terbanyak, yaitu 3.957 kasus (TP tertinggi), dengan menekan angka FN menjadi yang terendah (3.040 kasus). Pengorbanan dari sensitivitas tinggi ini adalah meningkatnya angka False Positive, di mana 8.528 pasien sehat salah diklasifikasikan sebagai penderita diabetes.

Hasil pengujian ini mengonfirmasi adanya fenomena "paradoks akurasi" pada pemodelan dataset *imbalanced*. XGBoost mencapai akurasi 86.71% bukan karena kemampuannya yang sangat baik dalam mendeteksi diabetes, melainkan karena model cenderung menebak "Tidak Diabetes" (*majority class*) pada hampir semua kasus. Tingkat *Recall* XGBoost yang hanya 17.80% menunjukkan bahwa lebih dari 80% pasien diabetes tidak terdeteksi oleh algoritma ini.

Dalam konteks informatika medis dan skrining kesehatan masyarakat, kesalahan berupa *False Negative* (pasien diabetes didiagnosa sehat sehingga tidak mendapat penanganan medis) memiliki risiko fatal yang jauh lebih besar dibandingkan *False Positive* (pasien sehat didiagnosa diabetes, yang mana akan diklarifikasi lebih lanjut melalui tes darah klinis).

Oleh karena itu, meskipun memiliki akurasi keseluruhan yang paling rendah (77.20%), Naive Bayes terbukti menjadi algoritma yang paling relevan dan rasional untuk diimplementasikan sebagai sistem deteksi dini (*early warning system*) diabetes pada penelitian ini. Kemampuan Naive Bayes dalam menangkap pola indikator gaya hidup

untuk mendeteksi potensi penyakit (ditandai dengan nilai *Recall* 56.55% dan *F1-Score* 40.62%) menjadikannya lebih dapat diandalkan secara medis dibandingkan algoritma yang sekadar mengejar akurasi tinggi pada data mayoritas.

5. Kesimpulan

Penelitian ini telah berhasil mengevaluasi dan membandingkan kinerja algoritma Decision Tree, K-Nearest Neighbor, Naive Bayes, dan XGBoost dalam memprediksi penyakit diabetes berdasarkan indikator perilaku dan gaya hidup. Hasil penelitian menyimpulkan bahwa algoritma XGBoost menghasilkan tingkat akurasi tertinggi sebesar 86.71%. Namun, akibat distribusi kelas yang tidak seimbang pada dataset medis, tingkat akurasi tinggi tidak merepresentasikan keberhasilan deteksi penyakit secara optimal.

Algoritma Naive Bayes, meski mencatatkan akurasi terendah sebesar 77.20%, terbukti keluar sebagai model terbaik secara fungsionalitas medis. Naive Bayes menghasilkan metrik sensitivitas (Recall) tertinggi sebesar 56.55% dan F1-Score sebesar 40.62%, menjadikannya algoritma yang paling responsif dalam meminimalkan False Negative (kasus pasien sakit yang tidak terdeteksi). Untuk pengembangan penelitian selanjutnya, disarankan untuk menerapkan teknik penyeimbang data (resampling) seperti SMOTE (Synthetic Minority Over-sampling Technique) guna meningkatkan Recall pada algoritma boosting seperti XGBoost, sehingga model dapat mencapai keseimbangan yang ideal antara Akurasi dan Sensitivitas.

Referensi

[1] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, dan M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020.

[2] A. Rahman dan S. Utama, "Comparative Analysis of K-Nearest Neighbor and Decision Tree Algorithms for Diabetes Mellitus Prediction," *Journal of Computer Science and Information Technology*, vol. 9, no. 2, pp. 112-118, 2022.

[3] T. Chen dan C. Guestrin, "XGBoost: A Scalable Tree Boosting System," dalam *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[4] P. Singh, S. Singh, dan G. S. Pandi-Jain, "Effective Machine Learning Approaches for Early Detection of Diabetes Risk Factors," *International Journal of Medical Informatics*, vol. 143, p. 104273, 2023.

[5] R. Wibowo, "Handling Imbalanced Datasets in Healthcare: A Review of Machine Learning Algorithms and Resampling Techniques," *Indonesian Journal of Artificial Intelligence*, vol. 4, no. 1, pp. 45-55, 2024.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, dan P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[7] Centers for Disease Control and Prevention (CDC), "Behavioral Risk Factor Surveillance System Survey Data," Atlanta, Georgia: U.S. Department of Health and Human Services, 2015.

[8] F. S. S. Nagalay, "Analisis Komparatif Kinerja Algoritma Naive Bayes dengan dan tanpa Seleksi Fitur Chi-Square untuk Deteksi Spam Email," *Coding: Journal of Computing and Software Engineering*, vol. 12, no. 3, pp. 110-121, Des. 2025.