Penerapan Machine Learning dalam Mendeteksi Penyakit Diabetes Dengan Algoritma Random Forest dan Naïve Bayes

Disusun Oleh :

Agung Prasetyo (20220801109)

Fandy Ilham Maulana (20220801020)

TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS ESA UNGGUL

2025

Pendahuluan

Diabetes melitus, yang lebih sering disebut sebagai diabetes, adalah kondisi kesehatan jangka panjang yang muncul ketika pankreas tidak menghasilkan insulin dalam jumlah yang memadai atau ketika tubuh tidak dapat memanfaatkan insulin yang tersedia dengan baik(Jian et al., 2021). Terdapat dua kategori utama dari diabetes, yaitu tipe 1 dan tipe 2. Pada diabetes tipe 1, yang juga dikenal sebagai diabetes insulin-dependent atau diabetes masa kanak-kanak, terjadi kekurangan dalam produksi insulin, sehingga individu yang mengalami hal ini memerlukan suntikan insulin setiap hari. Sementara itu, diabetes tipe 2, atau yang disebut diabetes non-insulin-dependent dan muncul pada usia dewasa, ditandai dengan ketidakmampuan tubuh untuk memanfaatkan insulin secara efisien(Jian et al., 2021).

Beragam faktor dapat memicu terjadinya diabetes. Contohnya adalah diabetes melitus tipe 1, yang bisa timbul akibat respons autoimun yang merusak sel-sel penghasil insulin di pankreas, yang dikenal sebagai sel beta. Di sisi lain, diabetes tipe 2 lebih sering disebabkan oleh faktor-faktor seperti usia, riwayat keluarga yang menderita diabetes, hipertensi, kadar trigliserida yang tinggi, serta adanya penyakit jantung atau riwayat stroke(Jian et al., 2021).

Berdasarkan informasi dari Organisasi Kesehatan Dunia (WHO), jumlah orang yang menderita diabetes telah meningkat secara dramatis dalam beberapa dekade terakhir dan diperkirakan akan terus bertambah di masa mendatang(Arminarahmah & Mahalisa, 2024). Diabetes melitus telah menjadi epidemi kesehatan global yang pesat, dengan proyeksi bahwa jumlah penderita diabetes akan meningkat dari 463 juta (yang menunjukkan prevalensi global sebesar 9,3%) pada tahun 2019 menjadi 700 juta (prevalensi global 10,9%) pada tahun 2045(Low, n.d.).

Melihat banyaknya kasus diabetes yang ada, dibutuhkan langkah-langkah awal untuk diagnosis dini penyakit ini dengan memprediksi risiko pada mereka yang berpotensi menderita diabetes. Prediksi terkait diabetes dapat dilakukan dengan mengumpulkan data dalam jumlah besar mengenai individu yang menderita diabetes, menyimpannya dalam basis data, dan menganalisisnya untuk mengidentifikasi pola-pola tertentu yang dapat membantu dalam deteksi dini penyakit ini(Nur Ikhromr et al., 2023). Salah satu cara untuk meningkatkan proses diagnosis

dan pemantauan penyakit adalah dengan memanfaatkan kemajuan teknologi informasi dan pembelajaran mesin(Arminarahmah & Mahalisa, 2024). Pembelajaran mesin merupakan satu cabang dari kecerdasan buatan dan ilmu komputer yang berfokus pada pemanfaatan data dan algoritma untuk meniru cara manusia belajar(Jian et al., 2021).

Pembelajaran mesin sendiri terbagi dalam dua kategori utama, yaitu pembelajaran terawasi dan pembelajaran tidak terawasi(Jian et al., 2021). Pembelajaran terawasi ditandai dengan penggunaan data berlabel dalam proses pelatihan algoritma dan dapat digunakan untuk keperluan klasifikasi atau regresi. Dalam tugas klasifikasi, tujuannya adalah untuk menentukan kelas atau kategori mana yang sesuai untuk setiap data yang belum diketahui sebagai bagian dari proses prediksi atau diagnosis.

Penelitian ini bertujuan mengimplementasikan model klasifikasi Naïve Bayes dan Random Forest, mengevaluasi performa keduanya, dan membandingkan hasilnya untuk menentukan algoritma yang paling efektif dalam deteksi dini potensi penyakit diabetes. Naive Bayes merupakan metode klasifikasi yang mengandalkan penghitungan probabilitas. Penentuan kelas suatu data dilakukan dengan membandingkan nilai kemungkinan dari suatu sampel terhadap kelas yang berbeda. Algoritma Naive Bayes adalah teknik pembelajaran Bayesian yang diketahui sangat berguna dalam beragam aplikasi. Naive Bayes termasuk dalam kategori supervised learning. Teknik ini dikenal memiliki tingkat akurasi yang baik dengan perhitungan yang relatif sederhana(Nur Ikhromr et al., 2023). Di sisi lain, Random Forest merupakan perkembangan dari metode pohon keputusan yang terdiri dari beberapa pohon keputusan; setiap pohon tersebut dilatih dengan sampel yang berbeda, dan setiap atribut digabungkan menjadi pohon yang diambil dari sekumpulan atribut yang acak(Setiawan et al., 2024). Algoritma Random Forest memiliki kelebihan karena model yang dihasilkan lebih stabil(Teknika & Ria Supriyatna, n.d.). Hal ini disebabkan oleh proses pembentukan pohon keputusan yang dilakukan tanpa pemangkasan dan secara independen menggunakan data acak, sehingga mengurangi kemungkinan overfitting. Oleh karena itu, Random Forest memiliki akurasi yang tinggi(Teknika & Ria Supriyatna, n.d.).

Penelitian ini diharapkan menghasilkan model deteksi diabetes yang efektif dan akurat, serta memberikan rekomendasi praktis bagi sektor kesehatan dalam melakukan deteksi dini penyakit diabetes secara lebih cepat dan efisien

Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen berbasis machine learning untuk mengevaluasi kinerja algoritma Naïve Bayes dan Random Forest dalam mendeteksi penyakit diabetes. Pendekatan ini dipilih karena memungkinkan analisis data secara objektif melalui penerapan teknik statistik dan pengukuran performa model klasifikasi secara terukur.

Data yang digunakan dalam penelitian ini berasal dari survei kesehatan yang dikembangkan oleh *Centers for Disease Control and Prevention (CDC)*. Dataset yang digunakan telah disiapkan dalam kondisi seimbang, yaitu 50% penderita diabetes dan 50% non-diabetes, untuk mengatasi permasalahan ketidakseimbangan kelas yang umum terjadi pada data medis. Data asli dari CDC berjumlah 253.680 entri, dengan distribusi 218.334 data non-diabetes dan 35.346 data diabetes, yang menunjukkan ketimpangan signifikan antar kelas.

Tahap preprocessing dimulai dengan membaca data dari file CSV, lalu memisahkan fitur prediktor dan label target. Seluruh fitur numerik dinormalisasi menggunakan StandardScaler guna menyamakan skala antar fitur, sehingga proses pelatihan model menjadi lebih optimal. Dataset hasil normalisasi ini kemudian digabung kembali dengan label target dan disimpan sebagai data bersih untuk mendukung dokumentasi dan *reproducibility*.

Data kemudian dibagi secara stratifikasi menjadi 60% data latih dan 40% data uji, untuk menjaga distribusi kelas yang seimbang dalam kedua subset. Untuk mengatasi kemungkinan ketidakseimbangan pada data latih, khususnya selama pelatihan model Naïve Bayes, diterapkan teknik Synthetic Minority Oversampling Technique (SMOTE). Penerapan SMOTE bertujuan agar model tidak bias terhadap kelas mayoritas dan mampu mengidentifikasi pola pada kelas minoritas dengan lebih baik.

Setelah pelatihan awal, model Random Forest digunakan untuk melakukan proses seleksi fitur berdasarkan tingkat kepentingan (*feature importance*). Lima fitur paling berpengaruh kemudian dipilih dan digunakan untuk melatih ulang kedua model, yaitu Naïve Bayes dan Random Forest. Pendekatan ini dilakukan untuk menguji apakah penggunaan fitur yang lebih sedikit namun lebih relevan dapat mempertahankan atau meningkatkan akurasi model serta menyederhanakan proses komputasi.

Evaluasi model dilakukan pada data uji menggunakan metrik klasifikasi seperti akurasi, presisi, recall, F1-score, dan ROC-AUC. Metrik-metrik ini memberikan gambaran menyeluruh mengenai kemampuan model dalam mengklasifikasikan kasus positif dan negatif secara seimbang. Seluruh tahapan analisis dilaksanakan menggunakan bahasa pemrograman Python dengan pustaka machine learning seperti Scikit-learn. Hasil evaluasi dari kedua algoritma kemudian dibandingkan untuk menentukan pendekatan yang paling efektif dalam mendeteksi dini penyakit diabetes.

HASIL DAN PEMBAHASAN

Penelitian yang dilakukan menggunakan dua model klasifikasi, yaitu *Random Forest* dan *Naïve Bayes* untuk mendeteksi potensi untuk mendeteksi potensi penyakit diabetes berdasarkan dataset BRFSS 2015. Setelah dilakukan pelatihan dan pengujian model dengan pembagian data stratified split (60% training dan 40% testing), diperoleh hasil metrik evaluasi sebagai berikut:\

Metrik Evaluasi	Random Forest	Naïve Bayes
Precision	0.72	0.73
Recall	0.79	0.76
F1-Score	0.75	0.74
Support	21.208	21.208

Dari hasil tersebut, dapat dilihat bahwa Random Forest unggul dalam nilai Recall (0.79) dan F1-Score (0.75), sedangkan Naïve Bayes sedikit lebih tinggi dalam Precision (0.73).

Selain itu, hasil Feature Importance menggunakan Random Forest menunjukkan fitur-fitur yang paling berpengaruh dalam deteksi diabetes, yaitu:

1. BMI_Age
2. GenHlth
3. BMI
4. Age
5. CardioRisk

Hasil evaluasi menunjukkan bahwa Random Forest memiliki performa yang lebih baik dibandingkan dengan Naïve Bayes dalam mendeteksi kasus diabetes, terutama dalam metrik Recall yang sangat penting pada konteks deteksi dini penyakit. Recall mengukur seberapa banyak kasus positif (diabetes) berhasil terdeteksi oleh model. Nilai Recall yang tinggi (0.79) pada Random Forest menunjukkan kemampuan model ini dalam menangkap sebagian besar kasus diabetes, sehingga mengurangi risiko false negative yang dapat berbahaya pada diagnosa medis.

F1-Score juga cukup tinggi (0.75) pada Random Forest, yang berarti model ini seimbang antara Precision dan Recall. Hal ini penting untuk memastikan bahwa deteksi dini penyakit tidak hanya sensitif tetapi juga relevan dengan kasus sebenarnya (positif).

Sebaliknya, meskipun Precision pada Naïve Bayes sedikit lebih tinggi, nilai Recall-nya lebih rendah dibandingkan Random Forest. Ini menunjukkan bahwa model Naïve Bayes cenderung lebih konservatif dalam memprediksi positif, yang dalam konteks deteksi dini diabetes kurang ideal.

Hasil Feature Importance yang diperoleh dari Random Forest mengindikasikan bahwa fitur BMI_Age (kombinasi indeks massa tubuh dan usia) menjadi prediktor paling signifikan. Hal ini wajar, mengingat usia dan BMI sering menjadi faktor risiko utama dalam perkembangan diabetes tipe 2.

Selain itu, fitur GenHlth (persepsi kesehatan umum) juga memiliki kontribusi penting, yang mencerminkan bagaimana faktor persepsi subjektif dapat terkait dengan risiko diabetes.

Fitur BMI, Age, dan CardioRisk juga memiliki nilai penting yang tinggi dalam model, yang sejalan dengan temuan sebelumnya bahwa faktor usia, obesitas, dan risiko kardiovaskular adalah komponen utama dalam deteksi risiko diabetes.

Berdasarkan hasil evaluasi, Random Forest dipilih sebagai model terbaik untuk deteksi dini diabetes dengan nilai ROC AUC sebesar 0.82, yang menunjukkan kemampuan model dalam memisahkan kelas positif dan negatif secara efektif. Selain itu, model ini juga memiliki stabilitas tinggi melalui proses ensemble learning yang menggabungkan banyak pohon keputusan, sehingga risiko overfitting dapat diminimalisir.

Naïve Bayes tetap relevan untuk aplikasi yang membutuhkan model yang ringan dan cepat diimplementasikan, namun dengan masalah pada performa Recall.

Daftar Pustaka

Arminarahmah, N., & Mahalisa, G. (2024). Implementasi Model Machine Learning pada Klasifikasi Status Penyakit Diabetes Berbasis Streamlit. *Smart Comp: Jurnalnya Orang Pintar Komputer*, *13*(3). https://doi.org/10.30591/smartcomp.v13i3.5866

Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A machine learning approach to predicting diabetes complications. *Healthcare (Switzerland)*, *9*(12). https://doi.org/10.3390/healthcare9121712

Low, L. L. (n.d.). *Brief Overview of ML Methods*. https://doi.org/10.17605/OSF

Nur Ikhromr, F., Sugiyarto, I., Faddillah, U., & Sudarsono, B. (2023). IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA

NAIVES BAYES DAN K-NEAREST NEIGHBOR IMPLEMENTATION OF DATA MINING TO PREDICT DIABETES DISEASE USING NAIVES BAYES AND K-NEAREST NEIGHBOR ALGORITHMS. *Journal of Information Technology and Computer Science (INTECOMS)*, *6*(1).

Setiawan, A., Hadryan Nst, Z., Khairi, Z., & Efrizoni, L. (2024). KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST. In *Jurnal Informatika & Rekayasa Elektronika)* (Vol. 7, Issue 2). http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

Teknika, J., & Ria Supriyatna, A. (n.d.). Teknika 17 (1): 163-172 Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *IJCCS*, *x, No.x*, 1–5.