Maliki Interdisciplinary Journal (MIJ) eISSN: 3024-8140
Volume 4, Issue November, 2026 pp. 2101-2112 http://urj.uin-malang.ac.id/index.php/mij/index

Klasifikasi risiko diabetes menggunakan algoritma naive bayes berdasarkan dataset pima indians diabetes

Qusay Mutawali
Program Studi Teknik Informatika, Universitas Islam Negeri Maulana Malik Ibrahim Malang
qmutawali@gmail.com

Kata Kunci:
diabetes mellitus, Naive Bayes, klasifikasi, Pima Indians Diabetes, data mining

Keywords:
diabetes mellitus, Naive Bayes, classification, Pima Indians Diabetes, data mining

A B S T R A K

Diabetes mellitus merupakan penyakit tidak menular dengan prevalensi yang terus meningkat, sehingga deteksi dini menjadi krusial untuk mencegah komplikasi lebih lanjut. Metode diagnosis konvensional yang bergantung pada pemeriksaan laboratorium sering kali terkendala waktu, biaya, dan keterbatasan akses fasilitas kesehatan. Penelitian ini bertujuan membangun dan mengevaluasi model klasifikasi risiko diabetes menggunakan algoritma Gaussian Naive Bayes pada dataset Pima Indians Diabetes yang terdiri atas 768 data pasien dengan delapan variabel klinis. Tahapan penelitian meliputi praproses data untuk menangani nilai nol yang tidak wajar secara medis, pembagian data latih dan uji dengan skema stratified split (80:20), pelatihan model, serta evaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan Area Under Curve (AUC). Hasil penelitian menunjukkan model mencapai akurasi 70,13%, presisi 56,67%, recall 62,96%, F1-score 59,65%, dan AUC 0,765. Variabel Glucose dan BMI teridentifikasi sebagai penentu paling berpengaruh terhadap hasil klasifikasi, sejalan dengan pemahaman klinis mengenai faktor risiko diabetes tipe dua. Penelitian ini menyimpulkan bahwa Gaussian Naive Bayes, meskipun sederhana secara komputasi, tetap layak digunakan sebagai instrumen skrining awal risiko diabetes, dengan rekomendasi penerapan teknik penyeimbangan kelas pada penelitian lanjutan guna meningkatkan sensitivitas model.

A B S T R A C T

Diabetes mellitus is a non-communicable disease with a continuously rising prevalence, making early detection essential to prevent further complications. Conventional diagnostic methods that rely on laboratory examination are often constrained by time, cost, and limited access to healthcare facilities. This study aims to build and evaluate a diabetes risk classification model using the Gaussian Naive Bayes algorithm on the Pima Indians Diabetes dataset, consisting of 768 patient records with eight clinical variables. The research stages include data preprocessing to handle medically implausible zero values, splitting the data into training and testing sets using an 80:20 stratified scheme, model training, and evaluation using accuracy, precision, recall, F1-score, and Area Under Curve (AUC) metrics. The results show that the model achieved an accuracy of 70.13%, precision of 56.67%, recall of 62.96%, F1-score of 59.65%, and an AUC of 0.765. Glucose and BMI were identified as the most influential variables in the classification outcome, consistent with clinical understanding of type 2 diabetes risk factors. This study concludes that Gaussian Naive Bayes, despite its computational simplicity, remains a viable instrument for early diabetes risk screening, with a recommendation to apply class-balancing techniques in future research to improve model sensitivity.

This is an open access article under the CC BY-NC-SA license.
Copyright © 2023 by Author. Published by Universitas Islam Negeri Maulana Malik Ibrahim Malang.

PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit tidak menular yang prevalensinya terus meningkat secara global (American Diabetes Association, 2023), ditandai dengan gangguan metabolisme glukosa akibat ketidakcukupan produksi insulin maupun resistensi insulin pada jaringan tubuh. Kondisi ini apabila tidak terdeteksi sejak dini berpotensi memicu komplikasi serius seperti kerusakan ginjal, gangguan penglihatan, hingga penyakit kardiovaskular. Deteksi dini terhadap risiko diabetes menjadi langkah krusial dalam upaya pencegahan, mengingat penanganan yang terlambat cenderung berdampak pada beban biaya kesehatan yang lebih besar serta menurunnya kualitas hidup penderita.Proses diagnosis diabetes secara konvensional umumnya bergantung pada pemeriksaan laboratorium dan interpretasi klinis oleh tenaga medis, yang meskipun akurat, memerlukan waktu, biaya, serta akses terhadap fasilitas kesehatan yang memadai. Keterbatasan ini menjadi tantangan tersendiri, terutama di wilayah dengan akses layanan kesehatan yang belum merata. Di sisi lain, perkembangan teknologi pengolahan data telah membuka peluang baru dalam bidang kesehatan, khususnya melalui pemanfaatan teknik penambangan data (data mining) untuk membangun sistem pendukung keputusan yang mampu mengidentifikasi pola risiko penyakit berdasarkan riwayat data pasien.

Salah satu pendekatan yang banyak diadopsi dalam klasifikasi risiko penyakit adalah algoritma Naive Bayes, sebuah metode probabilistik yang bekerja berdasarkan teorema Bayes dengan asumsi independensi antar fitur. Popularitas algoritma ini tidak lepas dari kesederhanaan struktur komputasinya, kecepatan proses pelatihan model, serta kemampuannya memberikan hasil klasifikasi yang kompetitif meskipun dengan jumlah data yang terbatas. Karakteristik tersebut menjadikan Naive Bayes relevan diterapkan pada kasus-kasus klasifikasi biner seperti penentuan status berisiko atau tidak berisiko diabetes.Penelitian terkait penerapan Naive Bayes pada kasus diabetes bukanlah hal baru. Mahendra et al. (2025) misalnya, mengkaji penerapan teknik data mining untuk mendeteksi diabetes mellitus dengan memanfaatkan algoritma Naive Bayes, yang menegaskan bahwa pendekatan berbasis probabilitas ini cukup efektif digunakan sebagai alat bantu identifikasi awal. Penelitian ini menempatkan diri sebagai kelanjutan dari eksplorasi tersebut, dengan memanfaatkan dataset Pima Indians Diabetes yang berisi rekam variabel klinis seperti kadar glukosa, tekanan darah, indeks massa tubuh, dan riwayat kehamilan sebagai basis pembentukan model klasifikasi.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk membangun serta mengevaluasi model klasifikasi risiko diabetes menggunakan algoritma Gaussian Naive Bayes pada dataset Pima Indians Diabetes, sekaligus menganalisis kontribusi masing-masing variabel klinis terhadap akurasi prediksi yang dihasilkan. Kontribusi penelitian ini diharapkan dapat memperkaya khazanah studi klasifikasi penyakit berbasis metode statistik sederhana, sekaligus menjadi rujukan awal bagi pengembangan sistem skrining risiko diabetes yang ringan secara komputasi namun tetap andal.

Kajian Pustaka

Diabetes Mellitus dan Faktor Risikonya

Diabetes mellitus didefinisikan sebagai kelompok gangguan metabolik yang ditandai dengan hiperglikemia kronis akibat kelainan sekresi insulin, kerja insulin, atau kombinasi keduanya (International Diabetes Federation, 2021). Beberapa faktor yang umum dikaitkan dengan peningkatan risiko diabetes antara lain kadar glukosa darah, indeks massa tubuh (BMI), usia, tekanan darah, riwayat kehamilan, serta faktor keturunan yang direpresentasikan melalui fungsi silsilah diabetes (diabetes pedigree function). Variabel-variabel inilah yang lazim dijadikan atribut dalam pembangunan model prediksi berbasis data klinis, termasuk pada dataset Pima Indians Diabetes yang digunakan secara luas dalam riset klasifikasi penyakit.

Penambangan Data dalam Bidang Kesehatan

Penambangan data (data mining) merujuk pada rangkaian proses ekstraksi pola dan pengetahuan tersembunyi dari kumpulan data berskala besar. Dalam ranah kesehatan, teknik ini banyak dimanfaatkan untuk mengidentifikasi pola tersembunyi antara variabel klinis pasien dengan status suatu penyakit, sehingga dapat mendukung pengambilan keputusan medis yang lebih cepat dan berbasis bukti. Salah satu cabang penambangan data yang relevan untuk kasus ini adalah klasifikasi, yaitu proses pengelompokan data ke dalam kategori tertentu berdasarkan pola yang dipelajari dari data latih.

Algoritma Naive Bayes

Naive Bayes merupakan algoritma klasifikasi probabilistik yang berlandaskan pada teorema Bayes dengan asumsi bahwa setiap fitur bersifat saling bebas secara kondisional terhadap kelas target (Bishop, 2006). Meskipun asumsi independensi ini jarang terpenuhi sepenuhnya dalam data dunia nyata, algoritma ini tetap menunjukkan kinerja yang memadai pada berbagai kasus klasifikasi, termasuk deteksi penyakit (Rish, 2001). Varian Gaussian Naive Bayes secara khusus dirancang untuk menangani fitur bertipe kontinu dengan mengasumsikan distribusi normal pada tiap fitur di dalam kelasnya, sehingga cocok diterapkan pada dataset klinis seperti Pima Indians Diabetes yang sebagian besar atributnya berupa nilai numerik berkelanjutan.

Penelitian Terdahulu

Beberapa penelitian sebelumnya telah mengeksplorasi penerapan algoritma klasifikasi pada kasus diabetes maupun sindrom metabolik terkait. Mahendra et al. (2025) menerapkan algoritma Naive Bayes untuk mendeteksi diabetes mellitus, menunjukkan bahwa pendekatan probabilistik ini layak digunakan sebagai model deteksi awal berbasis data rekam klinis. Sebagai pembanding metode, Nurussakinah dan Faisal (2023) mengkaji penerapan algoritma decision tree untuk mengklasifikasikan penyakit diabetes, yang menghasilkan model berbasis aturan pohon keputusan dengan tingkat interpretabilitas yang tinggi, meskipun struktur pohonnya berpotensi menjadi kompleks seiring bertambahnya jumlah fitur. Sementara itu, Fauziyah et al. (2024) melakukan

optimalisasi performa algoritma Gaussian Naive Bayes pada kasus prediksi metabolic syndrome dengan menerapkan teknik SMOTE untuk mengatasi permasalahan ketidakseimbangan kelas pada data, sebuah isu yang juga relevan diperhatikan mengingat distribusi kelas pada dataset Pima Indians Diabetes cenderung tidak seimbang antara kelompok berisiko dan tidak berisiko.Ketiga penelitian tersebut menegaskan bahwa metode klasifikasi probabilistik maupun berbasis aturan sama-sama memiliki peran penting dalam pengembangan sistem deteksi dini penyakit metabolik, sekaligus membuka ruang bagi penelitian ini untuk mengevaluasi lebih lanjut kinerja Gaussian Naive Bayes secara spesifik pada dataset Pima Indians Diabetes, termasuk mengidentifikasi variabel klinis yang paling berkontribusi terhadap hasil klasifikasi.

Metodologi Penelitian

Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental, di mana proses klasifikasi dilakukan secara terstruktur melalui tahapan pengumpulan data, praproses, pelatihan model, hingga evaluasi kinerja. Pendekatan eksperimental dipilih karena penelitian ini bertujuan mengukur secara objektif tingkat akurasi algoritma Gaussian Naive Bayes dalam mengklasifikasikan status risiko diabetes berdasarkan variabel klinis yang tersedia.

Sumber dan Karakteristik Data

Data yang digunakan dalam penelitian ini bersumber dari Pima Indians Diabetes Database (Smith et al., 1988), sebuah kumpulan data klinis yang telah lazim dijadikan acuan dalam riset klasifikasi penyakit diabetes. Dataset ini terdiri atas 768 baris data pasien perempuan dengan delapan variabel prediktor dan satu variabel target, sebagaimana dirangkum pada Tabel 1.

Tabel 1. Variabel Penelitian pada Dataset Pima Indians Diabetes

No	Nama Variabel	Keterangan	Tipe Data
1	Pregnancies	Jumlah kehamilan yang pernah dialami	Numerik (diskrit)
2	Glucose	Konsentrasi glukosa plasma hasil uji toleransi glukosa oral 2 jam	Numerik (kontinu)
3	BloodPressure	Tekanan darah diastolik (mm Hg)	Numerik (kontinu)
4	SkinThickness	Ketebalan lipatan kulit trisep (mm)	Numerik (kontinu)

No	Nama Variabel	Keterangan	Tipe Data
5	Insulin	Kadar insulin serum 2 jam (mu U/ml)	Numerik (kontinu)
6	BMI	Indeks massa tubuh (kg/m²)	Numerik (kontinu)
7	DiabetesPedigreeFunction	Fungsi silsilah yang merepresentasikan riwayat genetik diabetes	Numerik (kontinu)
8	Age	Usia pasien (tahun)	Numerik (diskrit)
9	Outcome	Status target: 1 = berisiko diabetes, 0 = tidak berisiko	Kategorikal (biner)

Tahap Praproses Data

Sebelum data digunakan untuk pelatihan model, dilakukan sejumlah tahapan praproses guna memastikan kualitas data. Tahapan tersebut meliputi pemeriksaan nilai kosong atau bernilai nol yang secara medis tidak wajar pada variabel tertentu (misalnya kadar glukosa atau tekanan darah bernilai nol), yang kemudian ditangani melalui teknik imputasi menggunakan nilai median atau rata-rata sesuai distribusi masing-masing variabel. Selanjutnya, dilakukan pemeriksaan distribusi kelas target untuk mengidentifikasi potensi ketidakseimbangan data, mengingat isu ini turut menjadi perhatian dalam penelitian Fauziyah et al. (2024) pada kasus serupa. Data kemudian dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan teknik stratified split, guna menjaga proporsi kelas yang seimbang pada kedua subset data.

Algoritma Klasifikasi

Algoritma yang digunakan dalam penelitian ini adalah Gaussian Naive Bayes, varian dari Naive Bayes yang dirancang untuk menangani fitur bertipe numerik kontinu dengan mengasumsikan bahwa nilai setiap fitur pada masing-masing kelas mengikuti distribusi normal (Gaussian). Model ini bekerja dengan menghitung probabilitas posterior suatu instans termasuk ke dalam kelas tertentu berdasarkan probabilitas prior kelas dan fungsi kepadatan probabilitas Gaussian dari tiap fitur, kemudian memilih kelas dengan nilai probabilitas tertinggi sebagai hasil prediksi akhir.

Skema Evaluasi

Kinerja model dievaluasi menggunakan confusion matrix sebagai dasar penghitungan sejumlah metrik, yaitu akurasi (accuracy), presisi (precision), sensitivitas atau recall, dan F1-score. Keempat metrik ini dipilih agar evaluasi tidak hanya berpatokan pada akurasi keseluruhan, melainkan juga mempertimbangkan kemampuan model dalam mengenali

kelas positif (berisiko diabetes) secara proporsional, mengingat potensi ketidakseimbangan distribusi kelas pada dataset yang digunakan.

Perangkat dan Lingkungan Eksperimen

Seluruh proses eksperimen, mulai dari praproses data hingga evaluasi model, dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka scikit-learn (Pedregosa et al., 2011) untuk implementasi algoritma Gaussian Naive Bayes, serta pustaka pandas dan numpy untuk manipulasi data.

Hasil

Karakteristik dan Kualitas Data Dataset Pima Indians Diabetes

Pemeriksaan awal terhadap 768 rekam data pasien mengungkap ketimpangan proporsi kelas target, dengan 500 sampel (65,10%) tergolong tidak berisiko diabetes dan 268 sampel (34,90%) tergolong berisiko diabetes, sebagaimana tercermin pada distribusi variabel Outcome.Penelusuran lebih lanjut menemukan sejumlah nilai nol yang secara fisiologis mustahil terjadi pada beberapa variabel klinis, dengan sebaran sebagai berikut:

a) Insulin: 374 baris (48,70%) tercatat bernilai nol, proporsi tertinggi di antara seluruh variabel

b) SkinThickness: 227 baris (29,56%) mengandung anomali serupa

c) BloodPressure: 35 baris terindikasi bernilai nol

d) BMI: 11 baris memerlukan koreksi

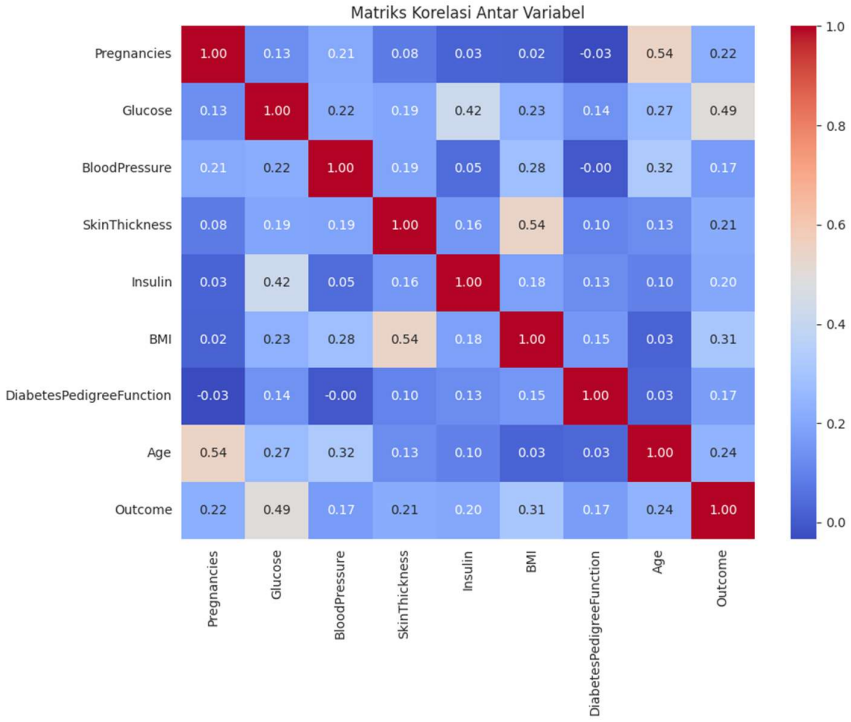
e) Glucose: 5 baris menunjukkan kejanggalan nilai

Seluruh nilai anomali tersebut kemudian disubstitusi menggunakan nilai median masing-masing kolom, menghasilkan dataset bersih tanpa distorsi numerik yang dapat mengaburkan pola klasifikasi.

Keterkaitan Antar Variabel Klinis

Pemetaan korelasi menunjukkan bahwa Glucose merupakan penentu paling kuat terhadap status Outcome (r=0,49), disusul BMI (r=0,31), Age (r=0,24), dan Pregnancies

(r=0,22), sebagaimana divisualisasikan pada Gambar 1.



Gambar 1. Matriks korelasi antar variabel klinis pada dataset Pima Indians Diabetes

Sebaliknya, DiabetesPedigreeFunction dan BloodPressure menunjukkan keterkaitan yang jauh lebih lemah terhadap penentuan status risiko.

Kinerja Klasifikasi Gaussian Naive Bayes

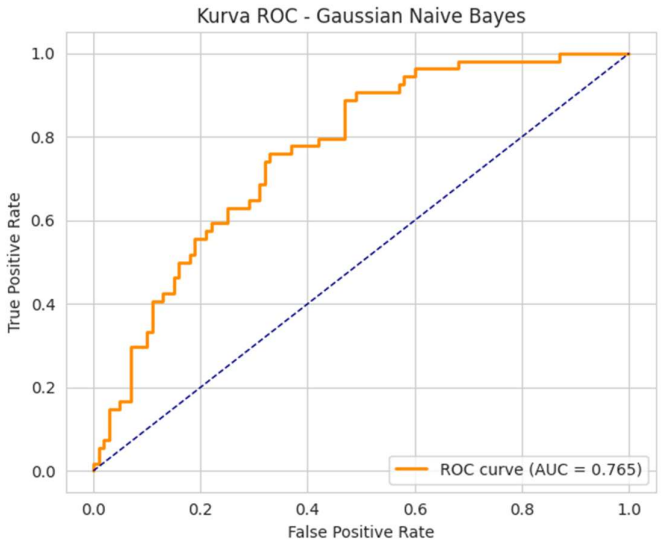
Setelah data terbagi menjadi 614 sampel latih dan 154 sampel uji melalui skema stratified split, model Gaussian Naive Bayes menghasilkan capaian sebagaimana dirangkum pada Tabel 2.

Tabel 2. Rekapitulasi Kinerja Model Gaussian Naive Bayes

Metrik	**Nilai**
Akurasi	70,13
Presisi	56,67
Recall	62,96
F1-Score	59,65

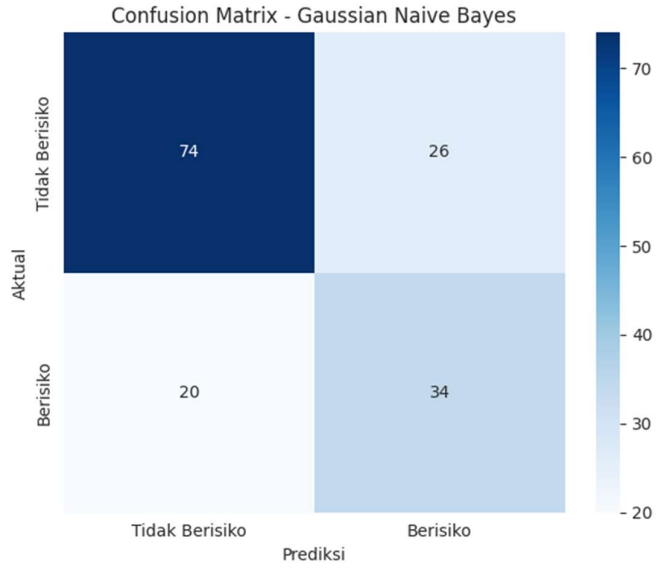
AUC	0,765

Model mencatat akurasi sebesar 70,13%, presisi 56,67%, recall 62,96%, F1-score 59,65%, serta nilai Area Under Curve (AUC) sebesar 0,765 sebagaimana tervisualisasi pada kurva ROC di Gambar 2.



Gambar 2. Kurva ROC hasil klasifikasi model Gaussian Naive Bayes pada data uji, dengan nilai Area Under Curve (AUC) sebesar 0,765

Rincian confusion matrix pada Gambar 3 menunjukkan model berhasil mengenali 74 dari 100 pasien pada kelompok tidak berisiko dan 34 dari 54 pasien pada kelompok berisiko, dengan 26 kasus salah klasifikasi berupa false positive dan 20 kasus berupa false negative.



Gambar 3. Confusion matrix hasil klasifikasi model Gaussian Naive Bayes pada 154 data uji, menunjukkan sebaran prediksi terhadap kelas aktual berisiko dan tidak berisiko diabetes

Kontribusi Variabel terhadap Pembentukan Keputusan

Analisis nilai rata-rata fitur pada masing-masing kelas menunjukkan bahwa Glucose memiliki selisih rata-rata paling besar antara kelompok berisiko dan tidak berisiko, diikuti oleh BMI dan Insulin, mengindikasikan ketiga variabel tersebut menjadi penentu utama arah keputusan klasifikasi model.

Diskusi

Interpretasi Kinerja Model dalam Konteks Studi Sejenis

Capaian akurasi 70,13% pada penelitian ini menegaskan bahwa algoritma probabilistik sesederhana Naive Bayes tetap kompetitif untuk kasus deteksi dini diabetes, selaras dengan temuan Mahendra et al. (2025) yang turut memvalidasi kelayakan Naive Bayes sebagai model deteksi awal tanpa memerlukan proses penyetelan parameter yang rumit. Kesenjangan antara presisi kelompok berisiko (56,67%) dan kelompok tidak berisiko (79%) mengindikasikan model masih cenderung menghasilkan false positive dalam jumlah tidak sedikit, sebuah pola yang lazim muncul ketika distribusi kelas pada data pelatihan tidak proporsional.Dibandingkan dengan pendekatan berbasis pohon keputusan yang diterapkan Nurussakinah dan Faisal (2023), Gaussian Naive Bayes unggul dari sisi kesederhanaan komputasi dan kecepatan proses pelatihan, meskipun struktur decision tree menawarkan keterbacaan aturan klasifikasi yang lebih eksplisit bagi pengguna non-teknis. Perbedaan karakteristik ini menegaskan bahwa pemilihan algoritma pada konteks skrining kesehatan idealnya mempertimbangkan keseimbangan antara efisiensi komputasi dan kebutuhan interpretabilitas oleh tenaga medis.

Isu Ketidakseimbangan Kelas dan Implikasinya

Selisih performa presisi antar kelas dalam penelitian ini memperkuat kekhawatiran yang diangkat Fauziyah et al. (2024) mengenai konsekuensi ketidakseimbangan distribusi kelas terhadap kinerja klasifikasi kasus penyakit metabolik. Mengingat proporsi kelompok berisiko hanya mencapai sepertiga dari keseluruhan data, penerapan teknik penyeimbangan seperti SMOTE (Chawla et al., 2002) sebagaimana dipraktikkan pada penelitian rujukan tersebut berpotensi menekan angka false negative, sebuah konsekuensi yang secara klinis jauh lebih merugikan dibandingkan false positive karena berisiko menunda penanganan pasien yang sesungguhnya membutuhkan intervensi dini.

Relevansi Klinis Variabel Dominan

Dominasi Glucose dan BMI sebagai penentu utama klasifikasi konsisten dengan pemahaman medis bahwa kadar glukosa darah dan indeks massa tubuh merupakan dua indikator klinis paling representatif dalam penapisan risiko diabetes tipe dua. Temuan ini memberi implikasi praktis bahwa instrumen skrining sederhana yang berfokus pada

kedua variabel tersebut berpotensi memberikan estimasi risiko awal yang cukup andal, bahkan sebelum pemeriksaan laboratorium yang lebih menyeluruh dilakukan.

Keterbatasan Penelitian

Penelitian ini memiliki sejumlah keterbatasan yang perlu diperhatikan pada pengembangan selanjutnya:

a) Ukuran dan Cakupan Data: Dataset yang digunakan berasal dari satu populasi spesifik (perempuan keturunan Pima), sehingga generalisasi hasil ke populasi dengan karakteristik demografis berbeda perlu dilakukan secara hati-hati

b) Ketidakseimbangan Kelas: Proporsi kelas berisiko yang lebih kecil berpotensi menyebabkan bias model terhadap kelas mayoritas, meskipun stratified split telah diterapkan untuk meminimalkan dampaknya

c) Penanganan Nilai Hilang: Strategi imputasi median, meskipun praktis, berpotensi mengaburkan variasi asli data dibandingkan metode imputasi yang lebih kompleks seperti regresi atau k-nearest neighbor

d) Asumsi Independensi Fitur: Asumsi independensi antar fitur pada Naive Bayes secara teoretis tidak sepenuhnya sesuai dengan korelasi antar variabel klinis yang teridentifikasi pada tahap eksplorasi data

Kesimpulan dan Rekomendasi

Kesimpulan

Penelitian ini berhasil membangun serta mengevaluasi model klasifikasi risiko diabetes berbasis algoritma Gaussian Naive Bayes pada dataset Pima Indians Diabetes, dengan capaian akurasi 70,13%, presisi 56,67%, recall 62,96%, F1-score 59,65%, dan AUC 0,765. Hasil ini menegaskan bahwa pendekatan probabilistik sederhana tetap relevan digunakan sebagai instrumen deteksi dini risiko diabetes, terutama pada konteks yang mengutamakan efisiensi komputasi.

Temuan kunci penelitian ini meliputi:

a. Identifikasi bahwa Glucose dan BMI merupakan variabel paling berkontribusi terhadap keputusan klasifikasi model

b. Konfirmasi bahwa ketidakseimbangan distribusi kelas berdampak pada kesenjangan presisi antara kelompok berisiko dan tidak berisiko

c. Validasi bahwa Gaussian Naive Bayes tetap kompetitif dibandingkan algoritma pembanding seperti decision tree pada kasus klasifikasi serupa

d. Penegasan urgensi penanganan ketidakseimbangan kelas sebagaimana disarankan

pada penelitian rujukan sejenis.

Rekomendasi untuk Praktisi

Berdasarkan temuan penelitian, berikut rekomendasi yang dapat dipertimbangkan:

a. Penerapan Model sebagai Alat Skrining Awal: Fasilitas kesehatan tingkat pertama dapat memanfaatkan model berbasis Naive Bayes sebagai instrumen penapisan awal yang ringan secara komputasi sebelum pemeriksaan laboratorium lanjutan

b. Penanganan Ketidakseimbangan Data: Praktisi disarankan menerapkan teknik penyeimbangan kelas seperti SMOTE atau oversampling guna meningkatkan sensitivitas model terhadap kelompok berisiko

c. Prioritas pada Variabel Kunci: Program skrining sederhana dapat difokuskan pada pengukuran Glucose dan BMI sebagai indikator awal yang paling informatif

d. Evaluasi Berkala Model: Model klasifikasi perlu dievaluasi ulang secara berkala menggunakan data terbaru guna menjaga relevansi dan akurasi prediksi seiring perubahan karakteristik populasi.

Arah Penelitian Masa Depan

Untuk penelitian selanjutnya, direkomendasikan:

a. Eksplorasi penerapan teknik penyeimbangan data seperti SMOTE untuk mengevaluasi peningkatan sensitivitas model terhadap kelas berisiko

b. Perbandingan sistematis antara Gaussian Naive Bayes dengan algoritma klasifikasi lain seperti decision tree, random forest, atau support vector machine pada dataset yang sama

c. Perluasan penelitian menggunakan dataset dengan cakupan populasi yang lebih beragam guna meningkatkan generalisabilitas model

d. Pengembangan sistem skrining berbasis web atau aplikasi mobile yang mengintegrasikan model klasifikasi ini untuk kebutuhan deteksi dini di layanan kesehatan primer

Secara keseluruhan, penelitian ini menyajikan bukti empiris bahwa algoritma Gaussian Naive Bayes, meskipun sederhana secara struktur komputasi, mampu memberikan kontribusi nyata dalam upaya deteksi dini risiko diabetes berbasis data klinis, sekaligus membuka ruang pengembangan lebih lanjut menuju sistem skrining kesehatan yang

lebih akurat dan aplikatif.

Daftar Pustaka

American Diabetes Association. (2023). 2. Classification and diagnosis of diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*, *46*(Supplement_1), S19–S40.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Fauziyah, N. J., Rahmania, F., Daniyal, M., & Sari, N. F. A. T. (2024). Analisis dan optimalisasi performa algoritma Gaussian Naive Bayes pada prediksi metabolic syndrome menggunakan SMOTE. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(2), 112–122. https://repository.uin-malang.ac.id/20732/

International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). International Diabetes Federation.

Mahendra, Y. H., Kusumawati, R., & Imamudin, M. (2025). Analisis data mining untuk deteksi diabetes mellitus menggunakan Naïve Bayes. *Jurnal Computer Science and Information Technology (CoSciTech)*, *6*(1), 39–44. https://repository.uin-malang.ac.id/25515/

Nurussakinah, N., & Faisal, M. (2023). Klasifikasi penyakit diabetes menggunakan algoritma decision tree. *Jurnal Informatika*, *10*(2), 143–149. https://repository.uin-malang.ac.id/17886/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*, 41–46.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, 261–265.