VOL. 03. NO. 01 JANUARI 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

ANALISIS KLASIFIKASI RISIKO DIABETES MENGGUNAKAN GAUSSIAN NAÏVE BAYES PADA DATA MEDIS

Firli Retnosari[1], Riri Narasati[2], Cep Lukman Rohmat[3], Rosidin[4].

Program Studi Teknik Informatika[124]
Program Studi Manajemen Informatika[3]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
firliretnosari12@gmail.com

(*) Corresponding Author : firliretnosari12@gmail.com
Published : 17 Juli 2026

Abstract—*Diabetes mellitus is a chronic metabolic disease with a prevalence that continues to increase each year, making early detection an important aspect of prevention. The utilization of machine learning in medical data offers opportunities to develop risk prediction systems that are both accurate and efficient. This study proposes the use of the Naïve Bayes Classifier algorithm as a simple, fast, and effective probabilistic classification method for predicting diabetes risk based on medical and lifestyle features. The dataset used consists of public medical data containing 100,000 samples and 31 attributes, including clinical indicators (glucose levels, HbA1c, blood pressure), demographic information, body mass index, physical activity, medical history, and various lifestyle-related factors. The research stages include data cleaning, feature normalization, correlation analysis, data splitting, model training, and performance evaluation using accuracy, precision, recall, and F1-score. The results indicate that Naïve Bayes achieves very high performance with an accuracy of 0.9980, precision of 1.0000, recall of 0.9968, and an F1-score of 0.9984. The most influential features in predicting diabetes risk are HbA1c, postprandial glucose, and fasting glucose. These findings demonstrate that Naïve Bayes is a reliable and efficient method for diabetes risk prediction and has strong potential for implementation in clinical decision-support systems for early screening. Future research is recommended to compare other algorithms, incorporate additional behavioral features, and validate the model using real clinical datasets.*

Keywords: *Diabetes Prediction, Naïve Bayes Classifier, Machine Learning, Medical Data, Digital Health.*

Abstrak—Diabetes melitus merupakan penyakit metabolik kronis dengan prevalensi yang meningkat setiap tahun, sehingga deteksi dini menjadi aspek penting dalam upaya pencegahan. Pemanfaatan machine learning pada data medis membuka peluang untuk mengembangkan sistem prediksi risiko yang akurat dan efisien. Penelitian ini mengusulkan penggunaan algoritma Naïve Bayes Classifier sebagai metode klasifikasi probabilistik yang sederhana, cepat, dan efektif untuk memprediksi risiko diabetes berdasarkan fitur medis dan gaya hidup. Dataset penelitian merupakan data medis publik berisi 100.000 sampel dan 31 atribut, meliputi indikator klinis (glukosa, HbA1c, tekanan darah), demografi, indeks massa tubuh, aktivitas fisik, riwayat penyakit, serta faktor gaya hidup lainnya. Tahapan penelitian meliputi pembersihan data, normalisasi fitur, analisis korelasi, pembagian data, pelatihan model, dan evaluasi performa menggunakan accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa Naïve Bayes memberikan performa sangat tinggi dengan akurasi 0,9980, precision 1,0000, recall 0,9968, dan F1-score 0,9984. Fitur paling berpengaruh terhadap risiko diabetes adalah HbA1c, glucose_postprandial, dan glucose_fasting. Temuan ini menunjukkan bahwa Naïve Bayes merupakan metode yang andal dan efisien untuk prediksi risiko diabetes, serta berpotensi diterapkan dalam sistem pendukung keputusan klinis untuk skrining awal. Penelitian selanjutnya disarankan membandingkan algoritma lain, menambah fitur perilaku, serta menguji model pada dataset klinis nyata.

Kata Kunci: *Diabetes Prediction, Naïve Bayes Classifier, Machine Learning, Medical Data, Digital Health.*

INTRODUCTION

Perkembangan teknologi informasi dalam bidang kesehatan telah mendorong transformasi signifikan dalam cara institusi medis mengelola data dan mengambil keputusan klinis. Digitalisasi data kesehatan memungkinkan analisis yang lebih cepat, akurat, dan efisien, terutama melalui pemanfaatan machine learning. Dalam konteks

layanan kesehatan preventif, machine learning telah menjadi alat penting untuk memprediksi risiko penyakit kronis berdasarkan pola historis dan parameter medis. Salah satu penyakit yang memerlukan deteksi dini berbasis teknologi adalah diabetes melitus, mengingat prevalensinya yang terus meningkat dan dampak jangka panjangnya terhadap kesehatan masyarakat. Diabetes melitus merupakan penyakit metabolik yang ditandai oleh tingginya kadar glukosa darah dan dapat menyebabkan komplikasi serius seperti gagal ginjal, gangguan jantung, neuropati, dan retinopati jika tidak ditangani dengan tepat.

Tren peningkatan jumlah penderita setiap tahun menunjukkan bahwa strategi deteksi dini masih belum optimal. Pemeriksaan manual membutuhkan waktu, biaya, serta tenaga medis terlatih, sehingga sering tidak dapat dilakukan secara rutin oleh seluruh lapisan masyarakat. Keterbatasan ini menjadi tantangan bagi negara berkembang, termasuk Indonesia, yang masih menghadapi kesenjangan akses layanan kesehatan dan rendahnya kesadaran masyarakat dalam melakukan pemeriksaan mandiri. Permasalahan tersebut menegaskan perlunya pendekatan prediktif berbasis teknologi yang mampu mengolah data medis dalam jumlah besar dan menghasilkan estimasi risiko secara objektif.

Machine learning menjadi solusi potensial karena dapat mengidentifikasi pola kompleks yang tidak mudah dilihat melalui analisis konvensional. Sistem prediksi otomatis dapat berfungsi sebagai alat bantu skrining awal, sehingga individu berisiko tinggi dapat segera memperoleh pemeriksaan lanjutan. Pendekatan ini sejalan dengan prioritas kesehatan modern yang menekankan pencegahan dan intervensi dini untuk menekan angka komplikasi diabetes. Penelitian ini bertujuan membangun model prediksi risiko diabetes menggunakan algoritma Naïve Bayes Classifier, sebuah metode berbasis probabilistik yang dikenal sederhana, cepat, dan efektif dalam tugas klasifikasi. Model dikembangkan dengan memanfaatkan dataset medis publik yang berisi variabel-variabel penting seperti usia, indeks massa tubuh (BMI), tekanan darah, kadar glukosa, HbA1c, kebiasaan merokok, aktivitas fisik, serta riwayat keluarga diabetes. Variabel-variabel tersebut menjadi indikator penting yang secara klinis berkaitan dengan perkembangan diabetes. Dengan data yang representatif, model memiliki potensi generalisasi yang baik untuk diterapkan dalam berbagai skenario. Tahapan penelitian meliputi pengumpulan data, pra- pemrosesan untuk mengatasi nilai hilang dan ketidakkonsistenan, transformasi data, pembagian dataset menjadi data latih dan data uji, pelatihan model Naïve Bayes, serta evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Pra-pemrosesan merupakan langkah krusial karena kualitas data sangat memengaruhi performa model. Evaluasi dilakukan untuk memastikan bahwa model mampu membedakan kategori risiko secara akurat dan konsisten.

Pendekatan ini juga memungkinkan identifikasi atribut yang paling berpengaruh dalam menentukan risiko diabetes sehingga hasil model dapat diinterpretasikan oleh tenaga kesehatan. Secara teoretis, Naïve Bayes menjadi pilihan algoritma yang tepat untuk prediksi medis karena kemampuannya mengolah data numerik maupun kategorikal, proses komputasi yang ringan, serta performa stabil pada dataset berskala besar maupun kecil. Meskipun algoritma ini mengasumsikan independensi antar fitur, banyak penelitian membuktikan bahwa Naïve Bayes tetap mampu memberikan performa yang kompetitif bahkan pada data dengan korelasi antarvariabel. Selain itu, kejelasan mekanisme kerja algoritma menjadikannya mudah dipahami oleh praktisi kesehatan, sehingga meningkatkan potensi adopsinya dalam sistem informasi medis berbasis teknologi.

Penelitian ini tidak hanya memberikan kontribusi teoretis, tetapi juga nilai praktis. Model prediksi risiko diabetes dapat digunakan sebagai alat skrining awal untuk membantu tenaga kesehatan mengidentifikasi individu dengan potensi risiko lebih tinggi. Dalam situasi fasilitas kesehatan terbatas, sistem semacam ini membantu mengarahkan intervensi dan pemeriksaan lanjutan secara lebih efektif. Selain itu, pemanfaatan model

prediktif dapat meningkatkan kesadaran masyarakat terhadap faktor risiko pribadi dan mendorong perubahan perilaku menuju gaya hidup yang lebih sehat. Secara keseluruhan, penelitian ini berkontribusi dalam pengembangan informatika kesehatan melalui integrasi machine learning dalam sistem prediksi medis. Dengan mengembangkan model berbasis Naïve Bayes yang ringan, mudah diimplementasikan, dan memiliki performa baik, penelitian ini mendukung upaya digitalisasi layanan kesehatan dan peningkatan kapasitas deteksi dini diabetes. Model yang dihasilkan diharapkan dapat menjadi langkah awal dalam membangun sistem kesehatan yang lebih preventif, terukur, dan responsif terhadap kebutuha masyarakat. kronis yang ditandai dengan tingginya kadar glukosa darah akibat gangguan produksi atau penggunaan insulin. Berbagai studi menunjukkan bahwa faktor usia, indeks massa tubuh, tekanan darah, kadar glukosa, HbA1c, serta riwayat keluarga memiliki hubungan signifikan dengan peningkatan risiko diabetes (Rifky et al., 2024; Budiargo, 2025). Variabel-variabel ini menjadi dasar utama dalam pembangunan model prediksi berbasis data karena keterkaitannya yang kuat dengan perkembangan kondisi metabolik pasien. Dataset medis yang memuat indikator tersebut banyak digunakan dalam penelitian machine learning untuk mendukung sistem deteksi dini dan klasifikasi risiko penyakit.

Machine learning telah banyak dimanfaatkan dalam bidang kesehatan karena kemampuannya mengolah data kompleks dan menghasilkan pola prediksi yang tidak mudah diidentifikasi secara manual. Di antara berbagai algoritma yang digunakan, Naïve Bayes menjadi salah satu metode yang paling banyak diterapkan dalam prediksi diabetes karena sederhana, efisien, dan memiliki performa stabil pada data berukuran besar maupun kecil (Prihamtini, 2025). Algoritma ini bekerja berdasarkan Teorema Bayes dengan mengasumsikan independensi antar fitur, sehingga perhitungannya cepat dan tidak memerlukan komputasi tinggi. Keunggulan tersebut menjadikan Naïve Bayes efektif diterapkan pada data medis yang bersifat multidimensional.

Beberapa penelitian terdahulu menegaskan keandalan Naïve Bayes dalam melakukan klasifikasi penyakit kronis. Model probabilistik ini mampu memberikan hasil prediksi yang kompetitif meskipun distribusi data tidak sepenuhnya memenuhi asumsi independensi antarvariabel (Zuhri et al., 2025). Penelitian lain juga menegaskan bahwa Naïve Bayes unggul dalam memproses data medis karena mudah diinterpretasikan serta memberikan gambaran probabilistik yang bermanfaat bagi tenaga kesehatan dalam memahami faktor risiko pasien. Hal ini menjadikannya sebagai salah satu metode yang tepat untuk sistem pendukung keputusan medis berbasis data.

Selain algoritma klasifikasi, proses evaluasi model juga menjadi bagian penting dalam

penelitian machine learning. Metrik seperti akurasi, presisi, recall, dan F1-score digunakan secara luas untuk menilai performa model dalam membedakan kategori diabetes dan non- diabetes (Putra & Hermawan, 2024). Evaluasi tersebut penting untuk memastikan bahwa model yang dihasilkan mampu bekerja secara konsisten pada data baru serta dapat digunakan sebagai alat skrining awal yang membantu tenaga kesehatan dalam pengambilan keputusan. Penelitian ini juga memanfaatkan confusion matrix untuk melihat distribusi prediksi benar dan salah sehingga hasil evaluasi dapat dianalisis secara lebih mendalam.

Berdasarkan kajian teori dan penelitian sebelumnya, penggunaan Naïve Bayes dalam prediksi risiko diabetes dinilai tepat karena stabilitas performanya, kemampuannya dalam mengolah data klinis secara efisien, serta kecocokannya untuk implementasi dalam sistem kesehatan digital. Kombinasi antara data medis yang representatif, preprocessing yang tepat, dan evaluasi model yang komprehensif memungkinkan penelitian ini menghasilkan model prediksi yang

akurat dan dapat diimplementasikan dalam layanan skrining kesehatan modern.

MATERIALS AND METHODS

Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi machine learning untuk memprediksi risiko diabetes berdasarkan data medis. Dataset diperoleh dari sumber publik dan berisi variabel penting seperti usia, indeks massa tubuh, tekanan darah, kadar glukosa, HbA1c, serta riwayat kesehatan lainnya yang relevan. Data yang telah diperoleh kemudian disimpan dalam format terstruktur untuk diproses lebih lanjut.

Tahap awal penelitian dilakukan dengan preprocessing data yang meliputi pembersihan data, penghapusan nilai duplikat, penanganan missing value, serta normalisasi nilai numerik agar seluruh variabel berada dalam rentang yang seragam. Proses ini dilakukan untuk memastikan bahwa data tidak mengandung noise atau ketidakkonsistenan yang dapat memengaruhi performa model. Selain itu, transformasi data dilakukan pada variabel kategorikal dengan metode encoding sehingga seluruh atribut dapat dibaca oleh algoritma.

Dataset kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji dengan proporsi 80:20 untuk memastikan bahwa model dapat dilatih secara optimal sekaligus diuji kemampuannya pada data baru. Proses pelatihan dilakukan menggunakan algoritma Naïve Bayes Classifier yang dipilih karena kemampuan komputasinya yang cepat serta efektivitasnya dalam pemodelan probabilistik. Selama proses pelatihan, model belajar mengenali hubungan antar fitur seperti kadar glukosa dan usia dengan kategori risiko diabetes.

Setelah model selesai dilatih, langkah selanjutnya adalah melakukan pengujian menggunakan data uji. Evaluasi performa model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menilai sejauh mana model mampu melakukan klasifikasi dengan benar. Confusion matrix digunakan sebagai alat tambahan untuk menganalisis jenis kesalahan prediksi yang terjadi, seperti false positive atau false negative, sehingga model dapat dievaluasi secara lebih komprehensif.

Seluruh proses penelitian, mulai dari preprocessing hingga evaluasi model, dilakukan menggunakan bahasa pemrograman Python dengan pustaka seperti Pandas, NumPy, dan Scikit-learn. Metode ini diharapkan mampu menghasilkan model prediksi diabetes yang akurat, stabil, dan dapat digunaka sebagai dasar pengembangan sistem pendukung keputusan dalam bidang kesehatan digital.

RESULTS AND DISCUSSION

Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes Classifier mampu melakukan klasifikasi risiko diabetes secara efektif berdasarkan variabel medis yang digunakan. Model menghasilkan akurasi tinggi dengan nilai konsisten pada metrik lain seperti precision, recall, dan F1-score, sehingga dapat disimpulkan bahwa Naïve Bayes bekerja stabil dalam mengenali pola risiko pada dataset kesehatan. Kinerja model yang baik mencerminkan bahwa variabel seperti kadar glukosa, usia, indeks massa tubuh, tekanan darah, dan HbA1c memiliki pengaruh kuat terhadap proses klasifikasi.

Analisis distribusi data menunjukkan bahwa jumlah individu yang diklasifikasikan sebagai tidak berisiko lebih dominGaanmbar 1 Distribusi Risiko Diabetes (variable dibandingkan kategori berisiko. Kondisi ini sejalan dengan karakteristik dataset medis yang biasanya memiliki proporsi kelas yang tidak seimbang. Pada model Naïve Bayes, ketidakseimbangan ini masih dapat ditangani secara optimal target)

Gambar ini menggambarkan penyebaran duparobabilitas antar fitur dengan baik. Visualisasi ini penting kelas pada variabel target Diagnosed_Diabetesse,bagai langkah awal untuk memahami struktur data dan yaitu Tidak Diabetes dan Diabetes. Dari totmalemprediksi potensi bias model sebelum dilakukan proses

100.000 sampel, kelompok Diabetes mendominasi dengan jumlah sekitar 60.000 data, sementara kategori Tidak Diabetes berjumlah sekitar 40.000 data. Perbedaan proporsi ini menunjukkan bahwa dataset memiliki ketidakseimbangan kelas, meskipun tidak terlalu ekstrem. Dominasi kelas

Diabetes memperlihatkan bahwa model memiliki lebih banyak informasi untuk mempelajari pola pada kelas ini, sehingga akurasi prediksi terhadap kategori Diabetes cenderung lebih tinggi dibandingkan kelas Tidak Diabetes. Namun demikian, performa prediksi pada kelas minoritas tetap terjaga karena algoritma Naïve Bayes mampu menangani distribusi probabilitas antar fitur dengan baik. Visualisasi ini penting sebagai langkah awal untuk memahami struktur data dan memprediksi potensi bias model sebelum dilakukan proses pemodelan lebih lanjut.

Confusion matrix pada gambar menunjukkan performa model dalam membedakan dua kelas, yaitu 0 (Tidak Diabetes) dan 1 (Diabetes) berdasarkan data uji. Matriks ini memberikan informasi detail mengenai jumlah prediksi benar dan salah untuk masing-masing kelas. Berdasarkan visualisasi:

True Negative (TN) = 8000 Artinya, terdapat 8000 sampel yang sebenarnya tidak diabetes dan berhasil diprediksi dengan benar oleh model. False Positive (FP) = 0 Tidak ada sampel tidak diabetes yang salah diprediksi sebagai diabetes. Ini menunjukkan model tidak menghSasilkan kesalahan tipe false positive. False Negative (FN) = 39 Sebanyak 39 sampel yang sebenarnya diabetes justru diprediksi sebagai tidak diabetes. Ini merupakan satu-satunya bentuk kesalahan yang muncul dalam jumlah kecil. True Positive (TP) = 11.961 Model berhasil mengidentifikasi 11.961 sampel diabetes secara benar. Secara keseluruhan, matriks ini menunjukkan bahwa model memiliki kinerja yang sangat baik. Prediksi benar pada kedua kelas mendominasi dengan tingkat kesalahan yang sangat rendah. Meskipun terdapat 39 kesalahan pada kelas diabetes (false negative), jumlah tersebut masih dalam batas wajar mengingat ukuran dataset yang besar.

Tidak ditemukannya false positive juga memperlihatkan bahwa Gambar ini menggambarkan penyebaran dua kelasmodel sangat berhati-hati dan akurat dalam menghindari salah pada variabel target Diagnosed_Diabetes, yaitu Tidakdeteksi pada individu yang tidak memiliki diabetes. Kinerja ini Diabetes dan Diabetes. Dari total 100.000 sampel,memperkuat kesimpulan bahwa algoritma Naïve Bayes mampu kelompok Diabetes mendominasi dengan jumlahmelakukan klasifikasi risiko diabetes secara efisien dan

sekitar 60.000 data, sementara kategori Tidak Diabetes berjumlah sekitar 40.000 data.

konsisten, bahkan pada dataset berdimensi besar.

Perbedaan proporsi ini menunjukkan bahwa dataset memiliki ketidakseimbangan kelas, meskipun tidak terlalu ekstrem. Dominasi kelas
Diabetes memperlihatkan bahwa model memiliki lebih banyak informasi untuk mempelajari pola pada kelas ini, sehingga akurasi prediksi terhadap kategori Diabetes cenderung lebih tinggi dibandingkan kelas Tidak Diabetes. Namun demikian, performa prediksi pada kelas minoritas tetap terjaga karena algoritma Naïve Bayes mampu menangani distribusi

Gambar ROC Curve tersebut menunjukkan kemampuan model dalam membedakan antara pasien yang memiliki diabetes dan yang tidak. Kurva berwarna oranye yang tampak melengkung tajam ke arah kiri atas menggambarkan bahwa model memiliki tingkat True Positive Rate (TPR) yang tinggi meskipun False Positive Rate (FPR) tetap rendah. Hal ini menandakan bahwa model mampu mendeteksi individu yang benar-benar mengalami diabetes secara efektif tanpa banyak melakukan kesalahan dalam mengklasifikasikan individu sehat sebagai penderita diabetes.
Nilai AUC sebesar 0.9216 memperkuat hal tersebut, karena nilai ini berada pada kategori sangat baik (excellent classifier) dan menunjukkan bahwa model memiliki kemampuan diskriminatif yang tinggi. Semakin besar nilai AUC, semakin efektif model dalam memisahkan dua kelas, dan pada grafik ini terlihat bahwa performanya jauh di atas garis diagonal acuan yang mewakili tebakan acak. Dengan demikian, ROC Curve memberikan bukti bahwa model Naïve Bayes yang digunakan sangat handal dan konsisten dalam memprediksi risiko diabetes berdasarkan data medis yang dianalisis

Gambar tersebut menggambarkan perbedaan rata-rata nilai fitur numerik antara dua kelompok, yaitu

individu dengan diabetes (kelas 1) dan individu non-diabetes (kelas 0), setelah melalui proses standardisasi. Dari visualisasi terlihat bahwa beberapa fitur klinis menunjukkan selisih rata-rata yang sangat

besar antara kedua kelompok.
Fitur HbA1c, glucose_postprandial, dan glucose_fasting menempati urutan paling atas dengan perbedaan nilai tertinggi, menandakan bahwa ketiga variabel ini merupakan indikator yang paling membedakan pasien diabetes dari non- diabetes.
Hal ini sejalan dengan karakteristik klinis bahwa peningkatan kadar glukosa darah dan HbA1c merupakan tanda umum dari gangguan regulasi gula darah. Selain itu, fitur seperti riwayat keluarga diabetes, usia, BMI, dan tekanan darah sistolik juga menunjukkan

perbedaan yang

cukup signifikan, yang mengindikasikan bahwa faktor hereditas, usia lanjut, dan kondisi metabolik turut berperan dalam risiko diabetes. Sebaliknya, beberapa fitur lain seperti

diet_score, hdl_cholesterol,

dan physical_activity_minutes_per_week memiliki perbedaan rata-rata yang negatif atau sangat kecil, menunjukkan bahwa variabel tersebut tidak terlalu membedakan antara kedua kelompok dalam dataset ini. Secara keseluruhan, gambar ini memberikan gambaran jelas mengenai fitur-fitur yang memiliki pengaruh terbesar dalam pemisahan kelas diabetes dan non- diabetes, serta membantu dalam memahami prioritas variabel yang paling relevan dalam proses klasifikasi.

CONCLUSION

Penelitian ini berhasil melakukan pemodelan prediksi risiko diabetes menggunakan algoritma Naïve Bayes Classifier berdasarkan data medis yang telah melalui proses pra-pemrosesan, normalisasi, dan eksplorasi data secara menyeluruh. Berdasarkan hasil pengujian, model mampu menghasilkan nilai akurasi yang sangat tinggi dengan didukung oleh metrik evaluasi seperti precision, recall, dan F1-score yang stabil. Hal ini menunjukkan bahwa algoritma Naïve Bayes efektif digunakan untuk mengolah data klinis berdimensi besar dan mampu menangkap pola-pola penting yang berkaitan dengan risiko diabetes.
Hasil analisis fitur menunjukkan bahwa variabel medis seperti HbA1c, glucose_fasting, dan glucose_postprandial merupakan indikator paling kuat dalam membedakan individu diabetes dan non-diabetes, sedangkan variabel lain seperti BMI, usia, tekanan darah, serta riwayat keluarga

memberikan kontribusi tambahan yang signifikan. Keberhasilan model dalam mengklasifikasikan risiko diabetes juga diperkuat oleh visualisasi confusion matrix dan nilai AUC yang tinggi, yang menegaskan kemampuan diskriminatif model dalam membedakan kedua kelas dengan tingkat kesalahan yang sangat rendah.
Kelebihan penelitian ini terletak pada penggunaan proses pra-pemrosesan data yang komprehensif dan analisis fitur yang mendalam sehingga model mampu bekerja secara optimal. Selain itu, penggunaan Naïve Bayes memberikan keuntungan dari sisi efisiensi komputasi dan interpretabilitas, sehingga model mudah diterapkan dalam konteks klinis sebagai alat bantu skrining awal.

Namun, penelitian ini masih memiliki beberapa keterbatasan, khususnya terkait ketidakseimbangan jumlah sampel antar kelas dan belum dilakukannya perbandingan performa dengan algoritma machine learning lainnya. Selain itu, model belum mempertimbangkan interaksi antar fitur secara lebih kompleks, sehingga potensi pola yang lebih dalam mungkin belum sepenuhnya tertangkap.

Untuk penelitian selanjutnya, model dapat dikembangkan dengan menyeimbangkan data menggunakan teknik seperti SMOTE, melakukan perbandingan performa dengan algoritma lain seperti Random Forest, XGBoost, atau Neural Network, serta mempertimbangkan integrasi fitur klinis tambahan untuk meningkatkan akurasi prediksi. Pengembangan sistem berbasis aplikasi atau dashboard klinis juga dapat menjadi langkah lanjutan agar model dapat digunakan secara

langsung oleh tenaga kesehatan dalam proses deteksi dini risiko diabetes.

REFERENCE

[1] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, "Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, 2023, doi: 10.51519/journalisi.v5i3.539.

[2] K. D. Indarwati and H. Februariyanti, "Analisis Sentimen Terhadap Kualitas Pelayanan Aplikasi Gojek Menggunakan Metode Naive Bayes Classifier," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 10, no. 1, 2023, doi: 10.35957/jatisi.v10i1.2643.

[3] A. Nurian, "Analisis Sentimen Ulasan Pengguna Aplikasi Google Play Menggunakan Naïve Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 11, no. 3s1, pp. 829–835, 2023, doi: 10.23960/jitet.v11i3s1.3348.

[4] H. Firda, "Comparison of Rating-based and Inset Lexicon-based Labeling in Sentiment Analysis using SVM (Case Study: GoBiz Application Reviews on Google Play Store)," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, no. 2, p. 516, 2025, doi: 10.32520/stmsi.v14i2.4795.

[5] R. Vincent, I. Maulana, and O. Komarudin, "Perbandingan Klasifikasi Naive Bayes dan Support Vector Machine Dalam Analisis Sentimen Dengan Multiclass Di Twitter," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 4, 2024, doi: 10.36040/jati.v7i4.7152.