

## Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Diabetes di Puskesmas Tegal Sari

Husein Nur Ansyari<sup>1</sup>, Sriani<sup>2</sup>, Armansyah<sup>3</sup>

<sup>1,2,3</sup> Program Studi Pendidikan Ilmu Komputer, Universitas Islam Negeri Sumatera Utara

e-mail: [ansyarinurhusein32@gmail.com](mailto:ansyarinurhusein32@gmail.com)

### Abstrak

Diabetes merupakan penyakit metabolik kronis yang ditandai dengan peningkatan kadar glukosa darah dan berpotensi menyebabkan komplikasi serius apabila tidak terdeteksi secara dini. Penelitian ini bertujuan untuk menerapkan algoritma Naïve Bayes dalam mengklasifikasikan status diabetes pasien di Puskesmas Tegal Sari Medan guna mendukung skrining awal yang lebih efektif dan akurat. Populasi dalam penelitian ini adalah seluruh pasien yang tercatat dalam rekam medis Puskesmas Tegal Sari. Teknik pengambilan sampel yang digunakan adalah purposive sampling berdasarkan ketersediaan data rekam medis yang lengkap, sehingga diperoleh sampel sebanyak 399 data pasien. Metode yang diterapkan adalah Gaussian Naïve Bayes dengan pendekatan supervised learning, di mana data dibagi menggunakan teknik holdout sebesar 80% untuk data latih (319 data) dan 20% untuk data uji (80 data). Tahapan penelitian meliputi preprocessing data yang mencakup penanganan missing values, label encoding pada variabel jenis kelamin, serta normalisasi data menggunakan Min-Max Scaling. Variabel yang digunakan sebagai prediktor meliputi glukosa, tekanan darah, ketebalan kulit, insulin, BMI, jenis kelamin, dan usia. Hasil pengujian menunjukkan bahwa model Naïve Bayes mencapai akurasi sebesar 82,50%, presisi 70%, recall 80,77%, dan F1-Score 75%. Analisis korelasi mengidentifikasi glukosa ( $r=0,46$ ) dan BMI ( $r=0,33$ ) sebagai variabel dengan pengaruh paling signifikan terhadap klasifikasi diabetes. Berdasarkan temuan tersebut, dapat disimpulkan bahwa algoritma Naïve Bayes mampu mengklasifikasikan status diabetes dengan tingkat akurasi yang baik dan berpotensi diimplementasikan sebagai alat bantu skrining diabetes di fasilitas kesehatan tingkat dasar.

**Kata Kunci:** Diabetes, Naïve Bayes, Klasifikasi, Data Mining, Puskesmas

### Abstract

Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels and can lead to serious complications if not detected early. This study aims to apply the Naïve Bayes algorithm to classify patients' diabetes status at the Tegal Sari Health Center (Puskesmas) in Medan in order to support more effective and accurate early screening. The population in this study consisted of all patients recorded in the medical records of Puskesmas

Tegal Sari. The sampling technique used was purposive sampling based on the availability of complete medical record data, resulting in a sample of 399 patient records. The method applied was Gaussian Naïve Bayes with a supervised learning approach, in which the data were split using a holdout technique of 80% for training data (319 records) and 20% for testing data (80 records). The research stages included data preprocessing, comprising the handling of missing values, label encoding for the gender variable, and data normalization using Min-Max Scaling. The variables used as predictors included glucose, blood pressure, skin thickness, insulin, BMI, gender, and age. The test results showed that the Naïve Bayes model achieved an accuracy of 82.50%, precision of 70%, recall of 80.77%, and F1-score of 75%. Correlation analysis identified glucose ( $r = 0.46$ ) and BMI ( $r = 0.33$ ) as the variables with the most significant influence on diabetes classification. Based on these findings, it can be concluded that the Naïve Bayes algorithm is able to classify diabetes status with a good level of accuracy and has the potential to be implemented as a diabetes screening tool in primary healthcare facilities.

**Keywords :** Diabetes, Naïve Bayes, Classification, Data Mining, Community Health Center

## LATAR BELAKANG

Perkembangan teknologi data mining dan machine learning memberikan peluang dalam mendukung proses deteksi dan klasifikasi penyakit secara lebih efektif. Diabetes Mellitus merupakan salah satu penyakit kronis dengan prevalensi yang terus meningkat sehingga diperlukan metode klasifikasi yang mampu menghasilkan prediksi secara akurat. Pemanfaatan dataset Behavioral Risk Factor Surveillance System (BRFSS) memungkinkan analisis terhadap berbagai faktor kesehatan, perilaku, dan demografi yang berkaitan dengan diabetes. Beberapa algoritma machine learning, seperti Naïve Bayes, C4.5, dan Random Forest, telah banyak digunakan dalam klasifikasi penyakit karena memiliki karakteristik dan tingkat kinerja yang berbeda. Selain pemilihan algoritma, konfigurasi hyperparameter juga dapat memengaruhi hasil klasifikasi. Oleh karena itu, Particle Swarm Optimization (PSO) dapat digunakan sebagai pendekatan optimasi untuk memperoleh konfigurasi model yang lebih sesuai dan meningkatkan kinerja klasifikasi diabetes (Maulana & Eliyani, 2025).

Klasifikasi terhadap penyakit diabetes secara medis masih menghadapi berbagai kesulitan, terutama akibat reduksi data dan adanya fitur yang tidak relevan dalam data medis. Klasifikasi terhadap pasien diabetes dapat dilakukan dengan mengumpulkan informasi penderita yang disimpan dalam basis data, kemudian diolah hingga membentuk suatu pola tertentu yang hasilnya dapat digunakan untuk mendiagnosa penyakit diabetes (Somantri, 2022).

Diabetes melitus adalah penyakit gangguan metabolisme kronis yang ditandai dengan kondisi hiperglikemia akibat gangguan pemanfaatan glukosa darah oleh tubuh. Penelitian oleh (Irawan et al., 2023) menerapkan algoritma Naïve Bayes dan K-Nearest Neighbor untuk mengklasifikasikan penderita diabetes menggunakan data publik dari Kaggle sebanyak 2000 data dengan delapan atribut prediktor, meliputi kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, berat badan, riwayat keluarga, dan usia. Hasil pengujian menunjukkan bahwa

algoritma K-Nearest Neighbor memberikan performa yang lebih unggul dibandingkan Naïve Bayes, dengan akurasi sebesar 92,6% berbanding 76,6%, sehingga K-Nearest Neighbor dinilai lebih andal dalam mendiagnosis penyakit diabetes pada penelitian tersebut.

Penelitian oleh (Nuryamin & Risyda, 2025) menerapkan algoritma Decision Tree C4.5 dan Naïve Bayes untuk memprediksi penyakit diabetes melitus menggunakan dataset publik dari Kaggle sebanyak 769 data dengan delapan atribut, meliputi Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age. Data dibagi dengan rasio 70:30 untuk pelatihan dan pengujian menggunakan aplikasi RapidMiner. Hasil pengujian menunjukkan bahwa algoritma Naïve Bayes memperoleh akurasi sebesar 73,48%, lebih tinggi dibandingkan algoritma Decision Tree yang hanya mencapai 70,87%, sehingga Naïve Bayes dinilai lebih efektif dalam memodelkan pola untuk memprediksi penyakit diabetes melitus pada dataset tersebut.

Diabetes mellitus merupakan salah satu penyakit kronis yang terus mengalami peningkatan jumlah penderita dan dapat menimbulkan berbagai komplikasi serius, seperti penyakit jantung, stroke, gagal ginjal, hingga infeksi pada kaki yang berisiko menyebabkan amputasi. Selain pemilihan algoritma, penggunaan metode sampling juga dapat memengaruhi hasil klasifikasi. Oleh karena itu, diperlukan perbandingan beberapa algoritma dengan metode sampling yang berbeda untuk memperoleh model yang mampu memberikan hasil prediksi diabetes secara lebih akurat dan andal (Sari & Linda Larasati, 2024).

Penelitian oleh (Veronica Agustin et al., 2023) menerapkan algoritma Naïve Bayes untuk mengklasifikasikan penyakit diabetes pada perempuan menggunakan dataset dari Kaggle sebanyak 767 data dengan sembilan atribut, meliputi pregnancies, glucose, blood pressure, insulin, BMI, diabetes pedigree function, dan age. Data diolah menggunakan tools RapidMiner dengan tahapan Knowledge Discovery on Database (KDD) dan pembagian data latih 60% serta data uji 40%. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes menghasilkan nilai akurasi sebesar 78,50%, precision 85,24%, recall 83,64%, serta nilai AUC sebesar 0,855 yang termasuk kategori baik (good). Penelitian ini menyimpulkan bahwa Naïve Bayes dapat digunakan untuk mengklasifikasikan penyakit diabetes pada perempuan guna mendukung deteksi dini, karena prosesnya cepat dan metodenya mudah dipahami meskipun tingkat akurasi masih berpotensi ditingkatkan dengan metode lain seperti ensemble.

Penelitian oleh (Darma Jonhar, 2026) menerapkan dan membandingkan algoritma K-Nearest Neighbor (k-NN) dan Naïve Bayes untuk mengklasifikasikan risiko diabetes berdasarkan faktor gaya hidup menggunakan dataset Diabetes Health Indicators dari Kaggle (bersumber dari BRFSS) sebanyak sekitar 200.000 baris data dengan 22 atribut, meliputi indikator perilaku seperti aktivitas fisik, konsumsi buah dan sayur, kebiasaan merokok, konsumsi alkohol berat, serta indikator klinis seperti tekanan darah tinggi dan kolesterol. Penelitian ini menegaskan bahwa pemilihan model klasifikasi tidak hanya perlu mempertimbangkan akurasi keseluruhan, tetapi juga performa pada tiap kelas, khususnya pada dataset kesehatan yang cenderung tidak seimbang.

Berdasarkan permasalahan tersebut, diperlukan suatu pendekatan yang dapat membantu proses identifikasi penyakit diabetes secara lebih cepat dan sistematis. Salah satu metode yang dapat diterapkan adalah Naïve Bayes, yaitu metode klasifikasi yang memanfaatkan probabilitas berdasarkan karakteristik atau gejala yang dimiliki pasien.

Penerapan metode tersebut dalam sistem berbasis web diharapkan dapat memberikan informasi awal mengenai kemungkinan seseorang mengalami diabetes berdasarkan gejala yang dimasukkan oleh pengguna. Penelitian ini bertujuan untuk menerapkan metode Naïve Bayes dalam melakukan klasifikasi diagnosis penyakit diabetes serta mengembangkan sistem berbasis web yang dapat membantu pengguna memperoleh informasi awal mengenai kondisi diabetes. Dengan adanya sistem tersebut, proses deteksi awal dapat dilakukan secara lebih praktis sehingga pengguna dapat memperoleh informasi yang selanjutnya dapat dijadikan pertimbangan untuk melakukan konsultasi dan pemeriksaan lebih lanjut kepada tenaga kesehatan (Nurwijayanti, 2023).

### **Penelitian Terdahulu**

Beberapa penelitian terdahulu telah membuktikan efektivitas algoritma Naïve Bayes dalam klasifikasi penyakit diabetes dengan berbagai pendekatan. (Nurul Anisa, 2022) menerapkan Gaussian Naïve Bayes pada dataset Kaggle yang memuat 390 data pasien dengan variabel glukosa, kolesterol, HDL, tekanan darah, usia, berat badan, dan jenis kelamin, menghasilkan akurasi 92,3% pada data latih dan 91,6% pada data uji dengan pembagian 60:40. Penelitian ini menegaskan bahwa Naïve Bayes unggul dibandingkan K-Nearest Neighbor yang hanya menghasilkan akurasi 91%, sekaligus menunjukkan bahwa glukosa merupakan variabel dengan korelasi langsung paling kuat terhadap status diabetes. (aditya samsuri et al., 2025) mengembangkan penelitian lebih lanjut dengan menambahkan teknik forward selection pada algoritma Naïve Bayes menggunakan 768 data dari Kaggle melalui pendekatan KDD dan tools RapidMiner.

Hasilnya menunjukkan bahwa penerapan seleksi fitur mampu meningkatkan akurasi model secara signifikan dari 67,32% menjadi 75,01%, dengan precision 76,10% dan recall 90,00%, sekaligus mengonfirmasi bahwa glukosa merupakan atribut dengan kontribusi terbesar terhadap prediksi diabetes. Sementara itu, (Amri et al., 2025) melakukan perbandingan antara Naïve Bayes dan Random Forest pada 266 data rekam medis nyata dari Klinik Citra Sejati menggunakan Python dan Scikit-learn. Penelitian ini menemukan bahwa meskipun Random Forest mencapai akurasi lebih tinggi (89,97%) dengan AUC 0,93, Naïve Bayes tetap relevan sebagai model awal dengan akurasi 85,97% dan AUC 0,72, khususnya karena kecepatan komputasinya dan kemudahan interpretasi dalam konteks klinis.

Penelitian oleh (Khasanah et al., 2022) mengklasifikasikan diagnosis diabetes melitus pada 130 data pasien RS Dirgahayu Samarinda periode 2018–2021 menggunakan algoritma Naïve Bayes Classifier, dengan enam variabel yaitu usia, jenis kelamin, status merokok, kadar glukosa, tekanan darah, dan kelas diagnosis ("Ya"/"Tidak"). Penelitian ini menguji empat proporsi pembagian data training dan testing, yaitu 60:40, 70:30, 80:20, dan 90:10, untuk membandingkan tingkat akurasi model pada tiap proporsi. Hasil penelitian menunjukkan bahwa proporsi data testing 40% dan 20% menghasilkan akurasi tertinggi sebesar 92,31%, sedangkan proporsi 30% dan 10% masing-masing menghasilkan akurasi 89,74% dan 84,62%. Penelitian ini membuktikan bahwa algoritma Naïve Bayes Classifier efektif digunakan untuk klasifikasi diagnosis diabetes melitus berbasis data rekam medis rumah sakit, meskipun tingkat akurasi dapat bervariasi tergantung proporsi pembagian data yang digunakan.

Penelitian oleh (Muhammad et al., 2023) membandingkan algoritma Naïve Bayes dan Decision Tree untuk mengklasifikasikan penyakit diabetes menggunakan dataset UCI dari

pasien Rumah Sakit Sylhet, Bangladesh, sebanyak 520 data dengan 17 atribut, meliputi usia, jenis kelamin, poliuria, polidipsia, penurunan berat badan mendadak, kelemahan, dan gejala klinis lainnya. Data dibagi menjadi 300 data training dan 220 data testing, kemudian diproses menggunakan aplikasi Weka dan dievaluasi menggunakan confusion matrix dengan metrik akurasi, precision, recall, dan F-measure. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memperoleh tingkat akurasi sebesar 96,36%, precision 0,967, dan recall 0,964, jauh lebih unggul dibandingkan algoritma Naïve Bayes yang hanya mencapai akurasi 90,45%, precision 0,908, dan recall 0,905. Penelitian ini menyimpulkan bahwa Decision Tree lebih efektif dibandingkan Naïve Bayes dalam memprediksi penyakit diabetes berdasarkan gejala klinis pasien.

Penelitian oleh (Mahalisa & Arminarahmah, 2025) mengevaluasi efektivitas penggabungan Gaussian Naïve Bayes (GNB) dengan Analisis Komponen Utama (PCA) dalam klasifikasi penyakit diabetes, dengan tujuan menilai apakah reduksi dimensi menggunakan PCA dapat meningkatkan kinerja model GNB. Penelitian ini melalui tahapan pengumpulan data, pra-pemrosesan, seleksi fitur, standarisasi data, reduksi dimensi dengan PCA, pelatihan dan pengujian model, hingga evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Penelitian ini menyimpulkan bahwa PCA tidak selalu memberikan manfaat dalam meningkatkan performa klasifikasi diabetes menggunakan Gaussian Naïve Bayes, karena efektivitasnya dipengaruhi oleh karakteristik data dan pemilihan hiperparameter.

Penelitian oleh (Azizah et al., 2025) menerapkan algoritma Naïve Bayes untuk mengklasifikasikan penyakit diabetes menggunakan dataset dari Kaggle sebanyak 100.000 data yang terbagi ke dalam dua kelas, yaitu Diabetes sebanyak 8.500 data dan Non-Diabetes sebanyak 91.500 data. Penelitian ini menerapkan tahapan pra-pemrosesan meliputi pembersihan data dan pembagian dataset dengan rasio 90:10 untuk data training dan testing, kemudian dievaluasi menggunakan confusion matrix dengan metrik akurasi, precision, dan recall. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mencapai akurasi sebesar 90,66%, dengan class precision untuk kelas Non-Diabetes sebesar 96,70% dan kelas Diabetes sebesar 46,51%, serta class recall masing-masing sebesar 92,96% dan 65,88%. Penelitian ini menyimpulkan bahwa Naïve Bayes merupakan pendekatan yang reliabel dan efektif untuk klasifikasi penyakit diabetes, meskipun performa pada kelas minoritas (Diabetes) masih dapat ditingkatkan melalui teknik data balancing pada dataset yang tidak seimbang.

Penelitian oleh (Azizah et al., 2025) membandingkan tiga algoritma klasifikasi, yaitu K-Nearest Neighbor (K-NN), Naïve Bayes, dan Decision Tree, untuk mengklasifikasikan penderita diabetes melitus menggunakan dataset dari Kaggle sebanyak 769 data dengan delapan atribut, meliputi Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age. Penelitian ini diolah menggunakan RapidMiner dengan tahapan data cleaning dan normalisasi pada proses pra-pemrosesan. Hasil penelitian menunjukkan bahwa algoritma K-NN memperoleh akurasi tertinggi sebesar 80,34% dengan precision 82,62% dan recall 88,40%, diikuti oleh Decision Tree dengan akurasi 77,43% dan recall tertinggi 97,60%, sementara Naïve Bayes menghasilkan akurasi terendah yaitu 77,34% dengan precision 75,08% dan recall 84,20%.

Penelitian oleh (Asyiah, 2026) menerapkan dan mengevaluasi performa algoritma Gaussian Naïve Bayes dalam mendukung deteksi dini penyakit diabetes menggunakan Pima

Indians Diabetes Dataset dari Kaggle sebanyak 768 data dengan delapan atribut, meliputi Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age. Penelitian ini menyimpulkan bahwa Naïve Bayes memiliki performa yang baik dalam mengklasifikasikan risiko diabetes meskipun merupakan metode yang sederhana, sehingga model tersebut diimplementasikan ke dalam aplikasi berbasis web sebagai alat bantu skrining awal bagi tenaga kesehatan maupun masyarakat.

Penelitian oleh (Ani Kusuma Putri et al., 2026) menerapkan algoritma Gaussian Naïve Bayes untuk mengklasifikasikan tingkat keparahan penyakit Infeksi Saluran Pernapasan Akut (ISPA) menggunakan dataset dari UPT Puskesmas Bontang Barat sebanyak 1.501 data pasien periode 2024–2025. Penelitian ini melalui tahapan pre-processing berupa data selection, cleaning, dan transformation, kemudian divalidasi menggunakan metode K-Fold Cross Validation dengan K=10 dan dievaluasi menggunakan confusion matrix. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memperoleh rata-rata nilai accuracy sebesar 75,00%, precision 78,12%, recall 75,00%, dan F1-score 76,36%. Penelitian ini menyimpulkan bahwa Naïve Bayes cukup andal dalam mendukung proses diagnosis ISPA dan menunjukkan kinerja yang relatif lebih baik dibandingkan beberapa algoritma lain pada penelitian-penelitian terdahulu untuk berbagai kasus penyakit.

Semua penelitian tersebut secara konsisten menunjukkan bahwa Naïve Bayes merupakan algoritma yang andal untuk klasifikasi diabetes, namun masih terdapat gap penelitian pada penggunaannya terhadap data rekam medis lokal dari fasilitas kesehatan tingkat dasar seperti Puskesmas, dengan variabel klinis spesifik dan tahapan preprocessing yang komprehensif, sebagaimana yang menjadi fokus penelitian ini.

### **Persamaan Matematika**

#### **Naïve Bayes**

Algoritma Naïve Bayes merupakan algoritma klasifikasi statistika yang digunakan untuk memprediksi probabilitas keanggotaan suatu kelas, dengan kecepatan dan akurasi yang tinggi terutama pada basis data berukuran besar, serta memiliki kemampuan klasifikasi yang serupa dengan neural network dan decision tree. Perhitungan probabilitas pada algoritma ini didasarkan pada Teorema Bayes, yang dirumuskan sebagai berikut:

$$P(C_i | X) = \frac{P(X | C_i) \cdot P(C_i)}{P(X)} \dots\dots\dots (1)$$

Dengan  $P(H|X)$  adalah probabilitas hipotesis H berdasarkan kondisi X (probabilitas posterior),  $P(X|H)$  adalah probabilitas kemunculan X berdasarkan kondisi pada hipotesis H,  $P(H)$  adalah probabilitas awal hipotesis H (probabilitas prior) yang tidak dipengaruhi oleh data apapun, dan  $P(X)$  adalah probabilitas kemunculan X (Marito Putri & Nurina Sari, 2022).

#### **Gaussian Naïve Bayes**

Algoritma Naïve Bayes merupakan metode klasifikasi statistik yang digunakan untuk memperkirakan kemungkinan sebuah data tergolong dalam kelas tertentu berdasarkan perhitungan probabilitas. Algoritma ini didasarkan pada Teorema Bayes yang dikembangkan oleh Thomas Bayes dan dikenal karena kemampuannya memproses basis data besar dengan cepat dan akurat. Dalam machine learning, Naïve Bayes sering diterapkan karena mampu memberikan akurasi tinggi melalui perhitungan yang sederhana. Algoritma ini pada dasarnya

hanya mendukung atribut dengan tipe data diskrit atau yang telah didiskritisasi, sehingga tidak mendukung atribut dengan nilai kontinu (numerik) — namun varian Gaussian Naïve Bayes dikembangkan khusus untuk mengatasi keterbatasan ini dengan mengasumsikan bahwa nilai atribut kontinu mengikuti distribusi normal (Gaussian) (Handayani et al., 2025) .

$$P(X_k|C_i) = \left( \frac{1}{\sqrt{(2\pi)\sigma_i k^2}} \right) \frac{e}{x} p \left( -\frac{((X_k - \mu_i k)^2)}{(2\sigma_i k^2)} \right) \dots\dots\dots (2)$$

## METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen berbasis machine learning untuk mengklasifikasikan status diabetes pasien. Pendekatan kuantitatif dipilih karena penelitian ini melibatkan pengolahan data numerik rekam medis pasien secara sistematis menggunakan algoritma Gaussian Naïve Bayes. Alur penelitian dirancang mengikuti tahapan yang terstruktur mulai dari pengumpulan data, praproses, pelatihan model, hingga evaluasi, sehingga pembaca dapat mengikuti dan mereplikasi setiap langkah penelitian ini dengan baik.

Populasi dalam penelitian ini adalah seluruh pasien yang tercatat dalam rekam medis di Puskesmas Tegal Sari, yang berlokasi di Jalan Tegal Sari No. 71, Kelurahan Tegal Sari, Kecamatan Medan Timur, Kota Medan, Sumatera Utara. Pengumpulan data dilaksanakan pada semester genap tahun ajaran 2023/2024, dengan jadwal penelitian yang mencakup tahap perencanaan, pengumpulan data, analisis dan perancangan, penerapan, serta pengujian yang berlangsung dari bulan Mei hingga Agustus.

Teknik pengambilan sampel yang digunakan adalah purposive sampling, yaitu pengambilan sampel berdasarkan ketersediaan data rekam medis pasien yang lengkap dan memenuhi kriteria variabel penelitian. Berdasarkan teknik tersebut, diperoleh sampel sebanyak 399 data rekam medis pasien. Seluruh data bersumber dari rekam medis pasien yang telah didiagnosis di Puskesmas Tegal Sari, mencakup informasi klinis pasien saat diagnosis maupun riwayat medis sebelumnya. Selama proses pengumpulan data, kerahasiaan identitas pasien dijaga sepenuhnya dan penggunaan data dilakukan sesuai dengan regulasi dan kebijakan yang berlaku dalam penelitian medis.

Variabel dalam penelitian ini terbagi menjadi dua kelompok, yaitu variabel bebas (independent variables) dan variabel terikat (dependent variable). Variabel bebas yang digunakan sebagai prediktor meliputi enam atribut klinis, yaitu: (1) glukosa, (2) tekanan darah, (3) ketebalan kulit, (4) insulin, (5) BMI (Body Mass Index), dan (6) jenis kelamin. Pemilihan keenam variabel ini didasarkan pada hasil analisis korelasi yang menunjukkan bahwa variabel-variabel tersebut memiliki hubungan yang signifikan atau relevan secara medis terhadap status diabetes pasien, di mana glukosa memiliki korelasi tertinggi dengan variabel target ( $r = 0,46$ ), diikuti oleh BMI ( $r = 0,33$ ) dan insulin ( $r = 0,20$ ). Variabel terikat (dependent variable) dalam penelitian ini adalah variabel Target, yaitu status diabetes pasien yang dikategorikan menjadi dua kelas: diabetes (1) dan tidak diabetes (0).

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh langsung dari rekam medis pasien Puskesmas Tegal Sari Medan. Data tersebut bersifat sekunder dalam konteks penelitian, artinya data telah tersedia dalam sistem rekam medis dan diakses dengan izin dari pihak Puskesmas. Format data awal berupa file .xlsx yang kemudian

dikonversi menjadi format .csv untuk keperluan pemrosesan dalam model. Dataset terdiri dari 399 baris data dengan 16 atribut, yang kemudian disaring menjadi 7 atribut penting untuk digunakan dalam proses analisis, yaitu enam variabel prediktor dan satu variabel target.

Pengumpulan data dilakukan dengan mengakses rekam medis pasien secara langsung di Puskesmas Tegal Sari. Proses ini melibatkan identifikasi dan pencatatan data klinis yang relevan, meliputi hasil pemeriksaan laboratorium terkait diabetes seperti kadar glukosa darah, tekanan darah, indeks massa tubuh (BMI), ketebalan kulit, kadar insulin, jenis kelamin, dan usia pasien, serta label diagnosis yang menyatakan apakah pasien terdiagnosis diabetes atau tidak. Informasi yang diperoleh tidak hanya mencakup data klinis saat diagnosis, tetapi juga riwayat medis pasien sebelumnya jika tersedia. Seluruh proses pengumpulan data dilakukan dengan memperhatikan aspek kerahasiaan data pasien sesuai etika penelitian medis yang berlaku.

Analisis data dalam penelitian ini menggunakan algoritma Gaussian Naïve Bayes dengan pendekatan supervised learning yang diimplementasikan menggunakan bahasa pemrograman Python melalui library Scikit-learn pada lingkungan Jupyter Notebook. Tahapan analisis data dilakukan secara terstruktur sebagai berikut.

Pertama, praproses data (preprocessing). Tahap ini bertujuan membersihkan dan mempersiapkan data agar siap digunakan dalam analisis. Praproses yang dilakukan meliputi tiga langkah: (1) penanganan missing values, yaitu mengidentifikasi dan menangani nilai yang hilang dalam dataset; (2) encoding data kategorikal, yaitu mengubah variabel jenis kelamin dari format teks menjadi numerik menggunakan metode label encoding, di mana male dikodekan sebagai 1 dan female sebagai 0; dan (3) normalisasi data menggunakan metode Min-Max Scaling untuk mengubah seluruh nilai variabel ke dalam rentang [0, 1] guna memastikan konsistensi skala antar variabel sebelum pelatihan model.

Kedua, pembagian data (data splitting). Dataset yang telah melalui tahap praproses kemudian dibagi menjadi dua kelompok menggunakan metode holdout, yaitu 80% sebagai data latih (training data) yang berjumlah 319 data dan 20% sebagai data uji (testing data) yang berjumlah 80 data. Pembagian ini dilakukan dengan parameter `random_state = 42` untuk memastikan hasil yang konsisten dan dapat direplikasi.

Ketiga, pelatihan model Model Gaussian Naïve Bayes dilatih menggunakan data latih melalui fungsi `GaussianNB()` dari `library` Scikit-learn. Pemilihan varian Gaussian didasarkan pada karakteristik data yang sebagian besar bersifat numerik kontinu. Proses pelatihan melibatkan penghitungan rata-rata ( $\mu$ ) dan standar deviasi ( $\sigma$ ) dari setiap fitur untuk masing-masing kelas menggunakan Persamaan (4.1), yang selanjutnya digunakan untuk menghitung probabilitas *likelihood* setiap variabel berdasarkan fungsi distribusi Gaussian pada Persamaan (2.4).

Keempat, evaluasi model. Performa model dievaluasi menggunakan data uji melalui empat metrik evaluasi yang dihitung berdasarkan Confusion Matrix, yaitu: (1) akurasi, yang mengukur proporsi prediksi benar secara keseluruhan menggunakan Persamaan (4.2); (2) presisi (precision), yang mengukur ketepatan prediksi positif menggunakan Persamaan (4.3); (3) recall, yang mengukur kemampuan model mendeteksi seluruh kasus positif menggunakan Persamaan (4.4); dan (4) F1-Score, yang merupakan harmonic mean dari presisi dan recall menggunakan Persamaan (4.5). Selain evaluasi berbasis perangkat lunak, dilakukan juga

perhitungan manual menggunakan rumus Gaussian Naïve Bayes untuk memvalidasi kebenaran hasil prediksi model secara matematis

## HASIL DAN PEMBAHASAN

### 1) Hasil

Penelitian ini bertujuan mengklasifikasikan status diabetes pasien di Puskesmas Tegal Sari Medan menggunakan algoritma Gaussian Naïve Bayes dengan data rekam medis sebanyak 399 pasien. Sebelum model dibangun, seluruh data melewati tahap praproses yang mencakup penanganan missing values, encoding variabel jenis kelamin, dan normalisasi Min-Max Scaling. Hasil praproses menunjukkan bahwa tidak ditemukan adanya missing values pada dataset, sehingga tidak diperlukan imputasi data. Variabel jenis kelamin berhasil dikodekan menjadi nilai numerik, dan seluruh variabel prediktor telah berhasil dinormalisasi ke dalam rentang [0, 1] sehingga skala antar variabel menjadi seragam dan siap digunakan untuk pelatihan model. Sebelum pelatihan model, dilakukan analisis korelasi antar variabel untuk memahami hubungan linear antara setiap prediktor dengan variabel target. Hasil analisis korelasi disajikan pada Tabel 1 berikut.

**Tabel 1. Nilai Korelasi Variabel Prediktor terhadap Target**

| Variabel        | Nilai Korelasi (r) terhadap Target |
|-----------------|------------------------------------|
| Glukosa         | 0,461                              |
| BMI             | 0,331                              |
| Ketebalan Kulit | 0,213                              |
| Insulin         | 0,198                              |
| Tekanan Darah   | 0,172                              |
| Jenis Kelamin   | —                                  |

Berdasarkan Tabel 1, glukosa merupakan variabel dengan korelasi paling kuat terhadap status diabetes ( $r = 0,461$ ), diikuti oleh BMI ( $r = 0,331$ ). Hal ini mengindikasikan bahwa kadar glukosa darah dan indeks massa tubuh merupakan faktor klinis yang paling dominan dalam menentukan klasifikasi diabetes pada dataset ini. Sebaliknya, tekanan darah menunjukkan korelasi yang lemah ( $r = 0,172$ ), yang berarti variabel tersebut memiliki kontribusi lebih kecil terhadap prediksi dibandingkan variabel lainnya. Temuan ini selaras secara medis karena kadar glukosa darah memang menjadi indikator utama dalam diagnosis diabetes melitus. Berdasarkan pembagian data dengan rasio 80:20, diperoleh 319 data latih dan 80 data uji. Model Gaussian Naïve Bayes dilatih menggunakan 319 data latih dengan menghitung rata-rata dan standar deviasi setiap variabel prediktor untuk masing-masing kelas. Hasil perhitungan parameter distribusi Gaussian untuk kelas positif dan negatif diabetes disajikan pada Tabel 2 dan Tabel 3 berikut.

**Tabel 2. Rata-Rata dan Standar Deviasi Kelas Positif Diabetes**

| Variabel        | Rata-Rata ( $\mu$ ) | Standar Deviasi ( $\sigma$ ) |
|-----------------|---------------------|------------------------------|
| Glukosa         | 141,26              | 31,94                        |
| Tekanan Darah   | 70,82               | 21,49                        |
| Ketebalan Kulit | 22,16               | 17,68                        |
| Insulin         | 100,34              | 138,69                       |
| BMI             | 35,14               | 7,26                         |
| Jenis Kelamin   | 0,53                | 0,50                         |

**Tabel 3. Rata-Rata dan Standar Deviasi Kelas Negatif Diabetes**

| Variabel        | Rata-Rata ( $\mu$ ) | Standar Deviasi ( $\sigma$ ) |
|-----------------|---------------------|------------------------------|
| Glukosa         | 110,64              | 26,14                        |
| Tekanan Darah   | 68,18               | 18,35                        |
| Ketebalan Kulit | 19,66               | 14,89                        |
| Insulin         | 68,79               | 98,86                        |
| BMI             | 30,30               | 7,88                         |
| Jenis Kelamin   | 0,37                | 0,48                         |

Tabel 2 dan Tabel 3 menunjukkan bahwa kelas positif diabetes memiliki nilai rata-rata glukosa (141,26) dan BMI (35,14) yang lebih tinggi dibandingkan kelas negatif diabetes, yaitu glukosa (110,64) dan BMI (30,30). Perbedaan rata-rata yang cukup signifikan pada kedua variabel ini mengonfirmasi bahwa glukosa dan BMI merupakan pembeda utama antara kedua

kelas, sehingga model Gaussian Naïve Bayes dapat memanfaatkan distribusi kedua variabel ini secara optimal dalam proses klasifikasi.

Setelah model dilatih, pengujian dilakukan terhadap 80 data uji. Hasil prediksi model dibandingkan dengan nilai aktual menggunakan *Confusion Matrix* yang disajikan pada Tabel 4 berikut.

**Tabel 4. Confusion Matrix Hasil Pengujian Model**

|                               | <b>Prediksi: Diabetes</b> | <b>Prediksi: Tidak Diabetes</b> |
|-------------------------------|---------------------------|---------------------------------|
| <b>Aktual: Diabetes</b>       | 21 (TP)                   | 5 (FN)                          |
| <b>Aktual: Tidak Diabetes</b> | 9 (FP)                    | 45 (TN)                         |

*Confusion Matrix* pada Tabel 4, model berhasil memprediksi dengan benar sebanyak 66 data dari 80 data uji, yang terdiri atas 21 kasus *True Positive* dan 45 kasus *True Negative*. Terdapat 5 kasus *False Negative*, yaitu pasien yang sebenarnya menderita diabetes tetapi diprediksi tidak diabetes, dan 9 kasus *False Positive*, yaitu pasien yang sebenarnya tidak menderita diabetes tetapi diprediksi diabetes. Hasil perhitungan keempat metrik evaluasi berdasarkan *Confusion Matrix* tersebut disajikan pada Tabel 5 berikut.

**Tabel 5. Hasil Evaluasi Model Gaussian Naïve Bayes**

| <b>Metrik Evaluasi</b> | <b>Nilai</b> |
|------------------------|--------------|
| Akurasi                | 82,50%       |
| Presisi (Precision)    | 70,00%       |
| Recall                 | 80,77%       |
| F1-Score               | 75,00%       |

Tabel 5 menunjukkan bahwa model Gaussian Naïve Bayes mampu mencapai akurasi sebesar 82,50%, yang berarti model memprediksi dengan benar sebanyak 82,50% dari seluruh data uji. Nilai *recall* yang mencapai 80,77% menunjukkan bahwa model mampu mendeteksi lebih dari 80% pasien yang benar-benar menderita diabetes, meskipun nilai presisi sebesar 70,00% mengindikasikan bahwa masih terdapat beberapa pasien sehat yang salah diprediksi sebagai penderita diabetes. F1-Score sebesar 75,00% mencerminkan keseimbangan yang cukup baik antara presisi dan *recall* dalam kondisi distribusi kelas yang tidak sepenuhnya seimbang pada dataset ini.

Hasil penelitian menunjukkan bahwa algoritma Gaussian Naïve Bayes mampu mengklasifikasikan status diabetes pasien di Puskesmas Tegal Sari dengan tingkat akurasi 82,50%, sehingga hipotesis yang diajukan dalam penelitian ini terbukti, yaitu bahwa algoritma

Naïve Bayes mampu mengklasifikasikan status diabetes pasien dengan tingkat akurasi minimal 80%. Capaian ini membuktikan bahwa pendekatan *machine learning* berbasis probabilistik dapat diimplementasikan secara efektif pada data rekam medis dari fasilitas kesehatan tingkat dasar, bukan hanya pada dataset publik yang telah banyak digunakan dalam penelitian sebelumnya.

Temuan bahwa glukosa merupakan variabel dengan kontribusi paling signifikan terhadap klasifikasi diabetes ( $r = 0,461$ ) sejalan dengan landasan medis yang menjadikan kadar glukosa darah sebagai indikator utama dalam diagnosis diabetes melitus. Hal ini juga dikonfirmasi oleh Samsuri, Faqih & Kaslani (2025) yang dalam penelitiannya menggunakan teknik *forward selection* menemukan bahwa glukosa merupakan atribut dengan bobot kontribusi paling besar dalam prediksi diabetes dibandingkan atribut lainnya. Selain itu, BMI yang menempati posisi kedua ( $r = 0,331$ ) turut memperkuat temuan Amri et al. (2025) yang menggunakan data rekam medis nyata dari Klinik Citra Sejati dan menemukan bahwa indeks massa tubuh berperan penting dalam klasifikasi diabetes, mengingat tingginya prevalensi obesitas di antara penderita diabetes tipe 2.

Nilai *recall* sebesar 80,77% yang lebih tinggi dari nilai presisi (70,00%) mencerminkan karakteristik model Gaussian Naïve Bayes yang lebih sensitif dalam mendeteksi kasus positif diabetes daripada menghindari kesalahan prediksi positif. Dalam konteks skrining medis di fasilitas kesehatan tingkat dasar seperti Puskesmas, karakteristik ini justru lebih menguntungkan secara klinis. Kasus *False Negative*, yaitu pasien diabetes yang tidak terdeteksi oleh model, memiliki risiko medis yang jauh lebih serius dibandingkan kasus *False Positive*, karena pasien yang tidak terdeteksi tidak akan mendapatkan penanganan yang diperlukan. Pandangan ini selaras dengan temuan (Pebriana et al., 2026) yang menegaskan bahwa dalam konteks skrining kesehatan masyarakat, model dengan *recall* tinggi lebih fungsional dan dapat diandalkan dibandingkan model yang sekadar mengejar akurasi tinggi namun menghasilkan *recall* yang rendah.

Nilai akurasi 82,50% yang diperoleh dalam penelitian ini berada pada rentang yang kompetitif dibandingkan beberapa penelitian sebelumnya yang menggunakan dataset publik. (Nurul Anisa, 2022) memperoleh akurasi 91,6% menggunakan dataset Kaggle dengan 390 data, namun dataset tersebut memiliki distribusi variabel yang berbeda dan tidak bersumber dari rekam medis lokal. Perbedaan akurasi ini dapat dijelaskan oleh beberapa faktor, antara lain: karakteristik dataset yang berbeda, variabel klinis yang digunakan, jumlah data, serta kondisi distribusi kelas pada masing-masing dataset. Pada penelitian ini, dataset bersumber dari rekam medis nyata Puskesmas Tegal Sari yang memiliki karakteristik lokal spesifik, sehingga distribusi data tidak seideal dataset publik yang telah melalui kurasi ketat. Hal ini menjadi salah satu faktor yang memengaruhi capaian akurasi model, sekaligus menjadi keunggulan penelitian ini dalam hal representasi kondisi nyata di lapangan.

Nilai F1-Score sebesar 75,00% yang dihasilkan menunjukkan bahwa model memiliki ruang untuk peningkatan lebih lanjut. Samsuri, (aditya samsuri et al., 2025) membuktikan bahwa penambahan teknik seleksi fitur berupa *forward selection* pada algoritma Naïve Bayes mampu meningkatkan akurasi dari 67,32% menjadi 75,01% dan *recall* dari 64,27% menjadi 90,00% secara signifikan. Temuan ini membuka peluang bahwa penerapan teknik seleksi fitur atau optimasi serupa pada dataset penelitian ini berpotensi meningkatkan performa model

secara lebih optimal. Di sisi lain, (Somantri, 2022) juga membuktikan bahwa optimasi bobot atribut menggunakan *Genetic Algorithm* mampu meningkatkan akurasi Naïve Bayes dari 87,69% menjadi 88,65%, yang menunjukkan bahwa pendekatan hibridisasi algoritma merupakan arah pengembangan yang menjanjikan untuk penelitian lanjutan.

Secara keseluruhan, penelitian ini memberikan kontribusi nyata bagi pengembangan ilmu pengetahuan, khususnya dalam dua aspek. Pertama, penelitian ini membuktikan bahwa algoritma Gaussian Naïve Bayes dapat diterapkan secara efektif pada data rekam medis lokal dari fasilitas kesehatan tingkat dasar di Indonesia dengan tingkat akurasi yang memadai, mengisi *gap* penelitian yang selama ini lebih banyak menggunakan dataset publik internasional. Kedua, temuan mengenai dominasi variabel glukosa dan BMI sebagai prediktor utama diabetes pada populasi pasien Puskesmas Tegal Sari memberikan informasi berbasis bukti yang dapat dimanfaatkan oleh tenaga kesehatan dalam menetapkan prioritas pemeriksaan klinis untuk skrining dini diabetes di tingkat komunitas. Meskipun demikian, model ini sebaiknya digunakan sebagai alat bantu skrining awal dan bukan sebagai pengganti diagnosis medis, mengingat masih adanya kasus *False Negative* yang memerlukan verifikasi klinis lebih lanjut oleh tenaga medis yang berwenang.

## SIMPULAN

Penelitian ini berhasil mengembangkan sistem klasifikasi diabetes berbasis algoritma Naïve Bayes menggunakan data rekam medis pasien di Puskesmas Tegal Sari. Sistem yang dibangun mampu melakukan klasifikasi secara otomatis terhadap status diabetes pasien berdasarkan enam atribut klinis, yaitu kadar glukosa, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh (BMI), dan usia. Hasil evaluasi model menunjukkan tingkat akurasi sebesar 82,50% pada data uji yang terdiri dari 80 sampel, dengan rincian 66 prediksi benar dan 14 prediksi salah. Dari evaluasi Confusion Matrix, diperoleh nilai True Positive (TP) sebanyak 20 kasus, True Negative (TN) sebanyak 42 kasus, False Positive (FP) sebanyak 12 kasus, dan False Negative (FN) sebanyak 6 kasus. Berdasarkan nilai tersebut, diperoleh presisi sebesar 70%, recall sebesar 80,77%, dan F1-Score sebesar 75%, yang mengindikasikan bahwa model memiliki kemampuan deteksi yang lebih kuat terhadap kasus positif diabetes dibandingkan kemampuan eliminasi kasus negatif. Analisis korelasi antarvariabel mengungkapkan bahwa kadar glukosa merupakan prediktor terkuat terhadap status diabetes dengan nilai korelasi 0,461 terhadap variabel target, diikuti oleh BMI ( $r = 0,331$ ). Temuan ini sejalan dengan literatur medis yang menegaskan peran sentral glukosa darah dan indeks massa tubuh dalam patofisiologi diabetes melitus. Sebaliknya, tekanan darah menunjukkan korelasi yang lebih lemah ( $r = 0,172$ ), mengindikasikan kontribusi yang terbatas pada proses klasifikasi dalam konteks dataset ini. Secara keseluruhan, penelitian ini membuktikan bahwa algoritma Naïve Bayes dengan pendekatan distribusi Gaussian efektif diterapkan pada data klinis kontinu untuk mendukung skrining diabetes. Sistem ini berpotensi diintegrasikan ke dalam sistem informasi Puskesmas sebagai alat bantu skrining awal, namun disarankan untuk tidak menggantikan diagnosis klinis secara menyeluruh mengingat masih terdapatnya kasus False Negative. Peningkatan performa model di masa mendatang dapat dilakukan melalui perluasan dataset, penambahan fitur seperti riwayat keluarga dan pola hidup, serta eksplorasi algoritma komplementer untuk mengurangi kesalahan klasifikasi.

**DAFTAR PUSTAKA**

- Aditya Samsuri, Ahmad Faqih, & Kaslani. (2025). 07+Filda,+4085-Article+Text-16276-1-6-20250415.Cpedit+(1).
- Amri, M. R. A., Permana, E., Pachadria, P. A., & Fitri, S. (2025). Perbandingan Metode Naïve Bayes Dan Random Forest Dalam Memprediksi Penyakit Diabetes Melitus Pada Klinik Citra Sejati. *Jtim: Jurnal Teknologi Informasi Dan Multimedia*, 7(4), 847–858. <https://doi.org/10.35746/Jtim.V7i4.747>
- Ani Kusuma Putri, B., Azhima Yoga Siswa, T., & Yulianto, F. (2026). Implementasi Algoritma Naive Bayes Untuk Klasifikasi Penyakit Ispa. 20(1).
- Asyiah, N. (2026). Bulletin Of Computer Science Research Implementasi Dan Evaluasi Performa Algoritma Naïve Bayes Dalam Deteksi Dini Penyakit Diabetes. *Media Online*, 6(4), 1719–1729. <https://doi.org/10.47065/Bulletincsr.V6i4.1159>
- Azizah, Q. N., Puspitasari, D., Sukmana, S. H., Sikumbang, E. D., & Ramanda, K. (2025). Pendekatan Algoritma Naive Bayes Dalam Memprediksi Penyakit Diabetes. In *Jurnal Infotech* (Vol. 7, Number 2). <http://ejournal.bsi.ac.id/Ejurnal/Index.php/Infotech89>
- Darma Jonhar, T. (2026). Hendri Mahmud Nawawi 4) 1), 2), 3), 4) *Informatika*, Universitas Nusa Mandiri. *Jl. Raya Jatiwaringin*, 3(2).
- Handayani, F., Firdaus, R., Wahyudi, A., Fu'adah Amran, H., Taufiq, R. M., Informatika, T., Komputer, I., & Riau, U. M. (2025). Kombinasi Algoritma Gaussian Naïve Bayes Dan Adaboost Untuk Meningkatkan Akurasi Dalam Klasifikasi Penyakit Diabetes.
- Irawan, F., Suprpti, T., & Bahtiar, A. (2023). Klasifikasi Algoritma Naive Bayes Dan K-Nearest Neighbor Pada Penderita Diabetes. 18, 80.
- Khasanah, L. U., Nasution, Y. N., Deny, F., & Amijaya, T. (2022). Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. 1(1), 41–50. <http://jurnal.fmipa.unmul.ac.id/Index.php/Basis>
- Mahalisa, G., & Arminarahmah, N. (2025). Penerapan Gaussian Naive Bayes Dan Analisis Komponen Utama Dalam Klasifikasi Diabetes. *Technologia : Jurnal Ilmiah*, 16(2), 220. <https://doi.org/10.31602/Tji.V16i2.15942>
- Marito Putry, N., & Nurina Sari, B. (2022). Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus. *Jurnal Sains Dan Manajemen*, 10(1).
- Maulana, R., & Eliyani, E. (2025). Diabetes Classification Algorithm Optimization Using Particle Swarm Optimization On Naïve Bayes, C4.5 And Random Forest. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, 14(4), 499–509. <https://doi.org/10.32736/Sisfokom.V14i4.2431>
- Muhammad, M., Samodro, J., Kunta Biddinika, M., Fadlil, A., Dahlan, A., & Ringroad Selatan, Y. J. (2023). Klasifikasi Penyakit Diabetes Dengan Algoritma Decision Tree Dan Naïve Bayes. 6(2).
- Nurul Anisa, D. (2022). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes. *Dinamika Informatika*, 14(1).
- Nurwijayanti, K. (2023). Klasifikasi Diagnosa Penyakit Diabetes Dengan Metode Naïve Bayes Berbasis Web. *Jurnal Kecerdasan Buatan Dan Teknologi Informasi*, 2(3), 115–121.
- Nuryamin, Y., & Risyda, F. (2025). Jsi (Jurnal Sistem Informasi) Universitas Dirgantara Marsekal Suryadarma Analisis Prediksi Penyakit Diabetes Menggunakan Metode

- Decision Tree C4.5 Dan Naive Bayes.  
<https://www.kaggle.com/code/melikedilekci/diabetes-dataset-for-beginners>
- Pebriana, L. S., Salam Nagalay, F. S., & Ubaidillah, U. (2026). Analisis Perbandingan Algoritma Decision Tree, K-NN, Naive Bayes, Dan Xgboost Memprediksi Penyakit Diabetes Menggunakan Data Behavioral Risk Factor Surveillance System.
- Sari, N. C., & Linda Larasati, T. (2024). Komparasi Algoritma Naïve Bayes Dan Gradient Boosting Untuk Prediksi Pasien Diabetes. *Jurnal Nasional Teknologi Dan Sistem Informasi*, 10(2), 118–125. <https://doi.org/10.25077/teknosi.v10i2.2024.118-125>
- Somantri, O. (2022). An Optimize Weights Naïve Bayes Model For Early Detection Of Diabetes. *Telematika*, 15(1). <https://doi.org/10.35671/telematika.v15i1.1307>
- Veronica Agustin, A., Voutama, A., Singaperbangsa Karawang Hs Ronggo Waluyo, U. J., & Barat, J. (2023). Implementasi Data Mining Klasifikasi Penyakit Diabetes Pada Perempuan Menggunakan Naïve Bayes. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Number 2).