Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 2272 – 2281 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PIPELINE MACHINE LEARNING PREDIKSI RISIKO PENYAKIT DARI DATA KUESIONER

Anggiat Roberto Sinaga[1], Siti Aisyah[2*], Dheo Zakaria Harahap[3], Anindyia Sitorus Pane[4], Sendy Valiza A. Ginting[5]
Universitas Prima Indonesia
anggiatrsinaga@gmail.com
e-mail: [1]anggiatrsinaga@gmail.com, [2]*siti_aisyah@unprimdn.ac.id

Abstract: *This study aims to design a system capable of predicting disease risk levels using the Naive Bayes algorithm and questionnaire data. Respondent data was collected via Google Forms, covering variables such as age, lifestyle, stress levels, sleep quality, physical activity, and family health history. The data underwent preprocessing and was converted into a tabular format before being divided into 80 training data points and 20 testing data points. The Naive Bayes algorithm was used to classify disease risk into low, moderate, and high categories. Test results showed that the model could generate predictions with a high level of accuracy. The trained model was then implemented in a web-based system so that users could easily and quickly determine their disease risk.*

Keywords: *Naive Bayes, Disease Risk Prediction.*

Abstract: Penelitian ini bertujuan untuk memprediksi tingkat risiko penyakit menggunakan algoritma Naive Bayes melalui data kursioner. Pengumpulan data pada penelitian ini dilakukan menggunakan kuesioner kesehatan, mencakup variabel seperti usia, gaya hidup, tingkat stres, kualitas tidur, aktivitas fisik, dan riwayat kesehatan keluarga. Data yang terkumpul kemudian disimpan dalam bentuk tabular sebelum dibagi menjadi 80 data latih dan 20 data uji. Algoritma Naive Bayes digunakan untuk mengklasifikasi risiko penyakit ke dalam kategori rendah, sedang, dan tinggi.. Hasil pengujian menunjukkan bahwa model mampu menghasilkan prediksi dengan tingkat akurasi yang tinggi. Model yang telah dilatih kemudian diimplementasikan dalam sistem berbasis web sehingga pengguna dapat dengan mudah dan cepat mengetahui risiko penyakit mereka.

Keywords: Naive Bayes, Prediksi Risiko Penyakit, Pipeline

PENDAHULUAN

Penyakit tidak menular (Non-Communicable Diseases/NCDs) seperti hipertensi, diabetes, dan penyakit jantung merupakan penyebab kematian terbesar di dunia dan terus meningkat setiap tahun. Penyakit-penyakit ini bersifat kronis, berkembang dalam jangka panjang, dan sering kali tidak menunjukkan gejala yang signifikan pada tahap awal[1] (WHO, 2025). Pengenalan dini serta upaya berbasis identifikasi faktor risiko menjadi penting untuk menekan dampak penyakit dan peningkatan kualitas hidup masyarakat (Artika et al., 2025). Salah satu metode yang paling umum digunakan untuk mengumpulkan data mengenai faktor risiko adalah melalui kuesioner kesehatan yang mencakup informasi mengenai gaya hidup, aktivitas fisik, pola makan, kebiasaan merokok, kualitas tidur, tingkat stres, dan riwayat keluarga. Meskipun mudah dan murah, data kuesioner menghadirkan berbagai tantangan yang signifikan, seperti nilai yang hilang, jawaban yang tidak konsisten, bias jawaban, dan format input yang beragam, sehingga data tersebut tidak dapat langsung digunakan dalam analisis atau prediksi (I. H. Sarker, 2022). Seiring dengan kemajuan teknologi

informasi, Machine Learning telah menjadi pendekatan yang banyak digunakan dalam analisis kesehatan karena mampu mengidentifikasi pola dalam data dan menghasilkan prediksi otomatis.

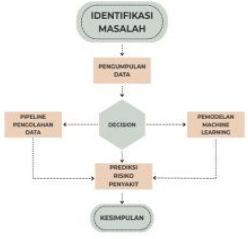
Berbagai penelitian telah menunjukkan bahwa model-model seperti Naive Bayes, Random Forest, dan Support Vector Machine dapat memprediksi penyakit jantung, diabetes, dan hipertensi dengan tingkat akurasi yang cukup tinggi. Banyak penelitian telah menggunakan Naive Bayes untuk mengklasifikasikan diabetes, seperti penelitian yang menggunakan kumpulan data dari UCI Machine Learning Repository untuk memprediksi risiko diabetes stadium awal. Penelitian terkait mengklasifikasikan risiko diabetes dengan akurasi 87,88% (Musababa & Fachrie, 2025). Hasil klasifikasi juga didapatkan pada data penyakit jantung yang menunjukkan bahwa metode ini terbilang efektifk dengan akurasi mencapai 79,92 % (Adam et al., 2024). Naive Byes merupakan metode klasifikasi yang populer untuk Machine Learning dan analisis data medis. Namun, keberhasilan model machine learning sangat bergantung pada kualitas data. Tanpa pra-pemrosesan yang memadai, model rentan terhadap kesalahan prediksi (Amelia et al., 2025).

Berdasarkan urgensi tersebut dilakukan penelitian yang untuk memprediksi resiko penyakit yang mengintegrasikan pipeline dalam pengolahan data dengan model machine learning berbasis web. Studi ini mengembangkan sistem prediksi risiko penyakit berbasis Machine Learning dengan mengintegrasikan pipeline pemrosesan data kuesioner dan model klasifikasi Naive Bayes. Selain itu, penelitian ini juga mempertanyakan bagaimana algoritma Naive Bayes dapat dimanfaatkan untuk mengenali pola faktor risiko pada data kuesioner sehingga mampu memberikan prediksi tingkat risiko penyakit degeneratif dalam kategori rendah, sedang, atau tinggi. serta bagaimana mengembangkan sistem informasi berbasis web yang dapat mengintegrasikan pipeline dan model tersebut sehingga mampu memberikan hasil prediksi risiko penyakit secara otomatis, informatif, dan mudah dipahami oleh pengguna.

METODE

Pengumpulan data pada penelitian ini dilakukan menggunakan kuesioner kesehatan sebagai sumber data utama. Kuesioner digunakan untuk mengumpulkan 17 variabel faktor risiko penyakit degeneratif, yaitu Kode responden, Jenis kelamin, Umur, Riwayat keluarga, frekuensi olahraga, frekuensi konsumsi makanan tinggi gula dan lemak, kebiasaan merokok, status konsumsi alkohol, riwayat gula darah tinggi, obesitas, gejala fisik, riwayat tekanan darah tinggi, frekuensi stres, konsumsi tinggi garam, Jantung, nyeri atau sesak dada dan kadar kolesterol.

Data dikumpulkan dari responden secara langsung melalui penyebaran tautan formulir secara daring. Seluruh responden mengisi data secara mandiri sesuai kondisi kesehatan dan kebiasaan hidup mereka. Data yang terkumpul berjumlah 100 responden selanjutnya disimpan dalam bentuk digital dan dibagi menjadi 2 dataset tabular yaitu, 80 data latih dan 20 data uji, untuk memudahkan proses pengolahan dan analisis. Sebelum digunakan pada tahap selanjutnya, data diperiksa terlebih dahulu untuk memastikan kelengkapan dan konsistensi jawaban, sehingga dapat mengurangi nilai kosong (missing value) dan kesalahan pengisian.



Gambar 1 Diagram Penelitian

1. Tahap ini bertujuan mengidentifikasi permasalahan utama terkait sulitnya deteksi dini penyakit tidak menular akibat kualitas data kuesioner yang rendah serta belum optimalnya pemanfaatan machine learning dalam prediksi risiko penyakit.
2. Data dikumpulkan melalui kuesioner untuk memperoleh informasi faktor risiko penyakit, seperti pola hidup, kebiasaan, dan riwayat kesehatan responden. Data yang diperoleh bersifat mentah.
3. Data kemudian diproses melalui Pipeline Pengolahan Data, yang meliputi, pembersihan data, penanganan missing value, transformasi data, normalisasi, dan pemilihan fitur.
4. Data yang telah diproses digunakan untuk membangun model machine learning menggunakan metode naive bayes yang bertujuan mempelajari pola hubungan antara faktor risiko dan risiko penyakit.
5. Decision dilakukan untuk menilai kelayakan data, performa model, dan kesiapan sistem.
6. Model yang telah dilatih digunakan untuk memprediksi tingkat risiko penyakit tidak menular menjadi kategori risiko rendah, sedang, dan tinggi.
7. Tahap akhir adalah pengujian sistem untuk memastikan website berjalan sesuai fungsi yang direncanakan.

Pipeline pengolahan data

Pipeline pengolahan data merupakan serangkaian tahapan terstruktur yang dirancang untuk mempersiapkan data mentah agar siap digunakan dalam pemodelan machine learning. Dalam konteks data kuesioner kesehatan, pipeline memiliki peran penting karena data yang dikumpulkan sering kali tidak lengkap, seperti nilai kosong (missing value), jawaban tidak konsisten, kesalahan input, serta variasi format data. Tahapan awal pipeline dimulai dengan pembersihan data (data cleaning) untuk memastikan bahwa data kuesioner yang diperoleh dari responden tidak mengandung duplikasi, kesalahan format, atau informasi yang tidak relevan (Santoso & Wijaya, 2026). Selanjutnya, dilakukan penanganan nilai kosong (missing value), yaitu memperbaiki atau mengisi nilai yang hilang dengan teknik yang sesuai dengan karakteristik data (Ramadhan et al., 2024). Tahapan berikutnya adalah transformasi data, di mana jawaban yang awalnya berupa teks dikonversi menjadi format numerik agar dapat dibaca oleh algoritma. Kemudian, pipeline dibagi menjadi data latih dan data uji dengan menggunakan rasio 80:20.

Machine Learning

Machine learning adalah cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk mempelajari pola dari data historis dan menghasilkan prediksi atau keputusan tanpa perlu diprogram secara eksplisit untuk setiap kondisi (Huda & Amalia, 2025). Proses penerapan Naive Bayes meliputi perhitungan probabilitas awal untuk setiap kelas risiko, perhitungan probabilitas setiap fitur dalam setiap kelas, serta penggabungan semua probabilitas tersebut menggunakan Teorema Bayes untuk menghasilkan prediksi akhir. Penelitian ini menggunakan metode Naive Bayes untuk mengklasifikasikan tingkat risiko penyakit ke dalam tiga kategori rendah, sedang, dan tinggi berdasarkan faktor risiko yang diperoleh dari data kuesioner.

HASIL DAN PEMBAHASAN

Pengumpulan Data

Data dalam penelitian ini diperoleh melalui penyebaran kuesioner menggunakan platform Google Form kepada responden. Kuesioner dirancang untuk mengumpulkan informasi dan memahami faktor-faktor risiko penyakit tidak menular seperti penyakit diabetes, hipertensi dan jantung. Kuesioner berisi 17 variabel pertanyaan yang meliputi usia, jenis kelamin, riwayat penyakit keluarga,

frekuensi olahraga, pola makan, frekuensi merokok, frekuensi konsumsi alkohol, kadar gula darah tinggi, berat badan, tingkat kolestrol, tingkat tekanan darah tinggi, dan frekuensi stress. Melalui kuesioner ini telah terkumpul data responden dengan jumlah keseluruhan 100 data. Kemudian, data disusun ke dalam bentuk dataset tabular untuk memudahkan proses pengolahan data lebih lanjut.

Pengolahan Data

Pipeline pengolahan data adalah alur kerja sistematis yang memastikan semua data melalui proses yang sama, sehingga meningkatkan kualitas data sebelum digunakan untuk Machine Learning.

1. Data Cleaning
   Tahapan pertama adalah pembersihan data untuk menghapus duplikasi, menangani nilai yang hilang pada variabel dan menghapus data yang tidak valid.
2. Transformasi Data
   Data kuesioner yang masih dalam bentuk kategori diubah menjadi data numerik yang terstruktur agar dapat diproses menggunakan metode Naive Bayes.
3. Transformasi Menjadi Dataset Tabular
   Setelah melalui tahapan pipeline, dataset memiliki struktur sebagai berikut:

Tabel 1 Tabel Variabel

Variabel	Keterangan
Jenis kelamin	Laki-laki: 1, perempuan: 0
Kode responden	Identitas unik responden (R1,R2,...)
Umur	Usia responden (Numerik)
Riwayat penyakit keluarga	0: tidak 1: ya
Frekuensi olahraga	0: >4 kali, 1: 3-4 kali, 2: 1-2 kali, 3: tidak pernah
Frekuensi konsumsi makanan tinggi gula dan lemak	0: Jarang, 1: sering, 2: sangat sering
Kebiasaan merokok	0: Tidak, 1: Ya
Frekuensi konsumsi alkohol	0: Tidak, 1: Iya
Riwayat gula darah tinggi	Tidak: 0, ya: 1
Obesitas	Ya: 1, tidak: 0
Gejala fisik	Ya: 1, tidak: 0
Riwayat tekanan darah tinggi	Ya: 1 ,tidak: 0
Frekuensi Stres	0: Jarang, 1: sering, 2: sangat sering
Frekuensi konsumsi makanan tinggi garam	0: Jarang, 1: sering, 2: sangat sering
Jantung	Tidak: 0, ya: 1
Nyeri atau sesak dada	Ya: 0, tidak: 1
Kadar Kolesterol	Tidak: 0, ya: 1

Tabel 2 Tabel Kategori

Kategori	Keterangan
Label risiko	Kategori: (Risiko rendah, risiko sedang, risiko tinggi)

4. Pembagian Data
   Proses ini dilakukan untuk memisahkan data ke dalam dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk mengajari algoritma agar dapat mengenali pola, sedangkan data uji digunakan untuk mengevaluasi kinerja model yang telah dibuat. Data yang dikumpulkan dibagi menjadi 80 data latih dan 20 data uji.
5. Penentuan Skor Risiko
   Penentuan tingkat risiko pada

penelitian ini dilakukan dengan menghitung skor total dari setiap responden berdasarkan variabel faktor risiko yang telah ditransformasikan ke dalam bentuk numerik. Proses perhitungan skor menggunakan perangkat lunak Microsoft Excel dengan memanfaatkan fungsi penjumlahan otomatis.

Setiap variabel faktor risiko memiliki nilai bobot yang berbeda sesuai dengan tingkat risiko. Variabel yang digunakan dalam perhitungan skor meliputi riwayat penyakit keluarga, frekuensi olahraga, konsumsi makanan tinggi gula dan lemak, kebiasaan merokok, konsumsi alkohol, riwayat gula darah tinggi, obesitas, gejala fisik, tekanan darah tinggi, tingkat stres, konsumsi garam, riwayat penyakit jantung, nyeri dada, serta kadar kolesterol. Variabel kode responden, umur dan jenis kelamin tidak dimasukkan dalam perhitungan skor karena tidak secara langsung merepresentasikan faktor risiko.

Perhitungan skor dilakukan menggunakan fungsi SUM pada Microsoft Excel dengan menjumlahkan seluruh nilai variabel faktor risiko dalam satu baris data. Rumus tersebut menjumlahkan seluruh variabel faktor risiko yang telah ditransformasikan ke dalam bentuk numerik pada setiap baris responden. Hasil dari fungsi tersebut menghasilkan nilai skor risiko untuk masing-masing responden secara otomatis. Nilai skor minimum yang diperoleh adalah 0 dan nilai skor maksimum adalah 19. Rentang skor kemudian dibagi menjadi tiga kategori tingkat risiko penyakit.

Setelah skor diperoleh, penentuan kategori risiko juga dilakukan menggunakan fungsi logika pada Microsoft Excel dengan menggunakan rumus. Rumus tersebut digunakan untuk mengelompokkan nilai skor ke dalam kategori risiko rendah, sedang, atau tinggi secara otomatis. Proses penentuan skor dan kategori risiko dilakukan secara sistematis menggunakan Microsoft Excel sehingga mengurangi kesalahan perhitungan hasil kategori risiko.

Metode Naive Baiyes

Naive Bayes merupakan metode klasifikasi yang tidak bergantung pada aturan tertentu, melainkan memanfaatkan teori probabilitas dalam matematika untuk menentukan kemungkinan tertinggi dari suatu klasifikasi. Proses ini dilakukan dengan menganalisis frekuensi masing-masing kelas dalam data pelatihan. Sebagai teknik klasifikasi statistik, Naive Bayes digunakan untuk memperkirakan probabilitas suatu data termasuk dalam kelas tertentu. Pendekatan ini didasarkan pada Teorema Bayes dan memiliki performa klasifikasi yang sebanding dengan metode seperti decision tree dan neural network (Azeraf et al., 2021).

Rumus dasar Naive Bayes adalah sebagai berikut:

$$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)}$$

Rumus Teorema Naive bayes:
X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi x (posteriori probability)
P(H) : Probabilitas hipotesis H (prior probability)
P(X|H) : Probabilitas X berdasarkan kondisi tersebut P(X) = Probabilitas dari X

Perhitungan Probabilitas Naive Bayes

Probabilitas prior dihitung berdasarkan jumlah masing-masing kategori pada data latih. Dataset yang digunakan sebagai data latih berjumlah 80 data.

Sehingga probabilitas prior dihitung sebagai berikut:
P(Rendah) = 65 / 80 = 0.8125
P(Sedang) = 15 / 80 = 0.1875

Nilai probabilitas prior ini digunakan sebagai dasar dalam

menghitung probabilitas posterior pada metode Naive Bayes.

Perhitungan Probabilitas Likelihood

Probabilitas likelihood merupakan probabilitas kemunculan suatu nilai variabel terhadap kelas tertentu. Nilai likelihood dihitung menggunakan 80 data latih yang telah diklasifikasikan ke dalam kategori risiko rendah, risiko tinggi dan risiko sedang. Perhitungan likelihood dilakukan untuk setiap variabel terhadap masing-masing kategori risiko. Salah satu variabel yang digunakan adalah Riwayat Penyakit Keluarga dengan nilai 1 (Ya).

Berdasarkan data latih, diperoleh jumlah data sebagai berikut:
Jumlah data risiko rendah = 65
Jumlah data risiko sedang = 15
Jumlah responden riwayat penyakit keluarga = 1 pada kelas risiko rendah = 18 data
Jumlah responden riwayat penyakit keluarga = 1 pada kelas risiko sedang = 20 data

Maka probabilitas likelihood dihitung sebagai berikut:
P(Riwayat = 1 | Rendah) = 16 / 65
P(Riwayat = 1 | Rendah) = 0.246
P(Riwayat = 1 | Sedang) = 11 / 15
P(Riwayat = 1 | Sedang) = 0.733

Jumlah responden dengan riwayat penyakit keluarga pada kelas risiko rendah sebanyak 16 data dari total 65 data. Sedangkan pada kelas risiko sedang sebanyak 11 data dari total 15 data. Nilai probabilitas likelihood untuk setiap variabel kemudian digunakan dalam perhitungan probabilitas posterior pada metode Naive Bayes.

Perhitungan Probabilitas Posterior

Probabilitas posterior merupakan probabilitas akhir yang digunakan untuk menentukan kategori risiko berdasarkan nilai probabilitas prior dan probabilitas likelihood yang telah dihitung sebelumnya. Perhitungan probabilitas posterior pada penelitian ini menggunakan metode Naive Bayes dengan mengalikan probabilitas prior dan probabilitas likelihood dari masing-masing kategori.

Rumus probabilitas posterior yang digunakan adalah sebagai berikut:

$$P(H|X) = P(H) \times P(X|H)$$

Rumus probabilitas posterior:
P(H|X) = probabilitas posterior
P(H) = probabilitas prior
P(X|H) = probabilitas likelihood

Berdasarkan hasil perhitungan probabilitas prior menggunakan 80 data latih diperoleh nilai sebagai berikut:
P(Rendah) = 65 / 80 = 0.8125
P(Sedang) = 15 / 80 = 0.1875
berdasarkan perhitungan probabilitas likelihood untuk variabel riwayat penyakit keluarga diperoleh:
P(Riwayat = 1 | Rendah) = 0.246
P(Riwayat = 1 | Sedang) = 0.733

Selanjutnya dilakukan perhitungan probabilitas posterior sebagai berikut:
Perhitungan untuk kelas risiko rendah:
P(Rendah|X) = P(Rendah) × P(Riwayat = 1 | Rendah) = 0.812× 0.246 = 0.199

Perhitungan untuk kelas risiko sedang:
P(Sedang|X) = P(Sedang) × P(Riwayat = 1 | Sedang)
= 0.1875 × 0.733
= 0.137

Hasil Akhir Risiko

Berdasarkan hasil perhitungan probabilitas posterior menggunakan metode Naive Bayes, diperoleh nilai probabilitas untuk masing-masing kelas risiko. Nilai probabilitas posterior terbesar digunakan sebagai dasar dalam menentukan hasil klasifikasi.

Hasil perhitungan probabilitas posterior diperoleh sebagai berikut:
P(Rendah|X) = 0.199
P(Sedang|X) = 0.137

Berdasarkan hasil tersebut, nilai probabilitas terbesar terdapat pada kelas risiko sedang. Oleh karena itu, data tersebut diklasifikasikan sebagai **risiko rendah**.

Implementasi Model dalam Sistem

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 2272 – 2281 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Informasi Berbasis Web

Pada tahap ini, model Naive Bayes yang dilatih menggunakan 20 data uji telah diintegrasikan ke dalam sistem informasi berbasis web yang disebut "CheckSehat." Sistem ini berfungsi sebagai antarmuka yang memungkinkan pengguna atau administrator untuk melihat prediksi risiko penyakit secara real-time berdasarkan probabilitas yang dihasilkan oleh model tersebut.

Sistem prediksi risiko penyakit ini dibangun sebagai aplikasi berbasis web yang terdiri dari tiga komponen utama, yaitu frontend, backend, dan model machine learning. Bagian frontend, pengguna berinteraksi melalui browser untuk mendaftar, masuk, danmengisi kuesioner kesehatan. Data yang dimasukkan oleh pengguna dikirim ke backend untuk diproses melalui alur kerja yang telah dirancang, termasuk pembersihan data, mengubah nilai jawaban menjadi format numerik, dan mempersiapkan data untuk

dianalisis oleh model.

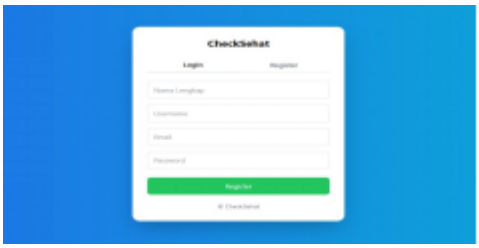
Backend bertanggung jawab untuk menerima input dari formulir, memproses data sesuai dengan alur kerja, dan menjalankan model Naive Bayes yang telah dilatih sebelumnya menggunakan dataset hasil dari proses alur kerja tersebut. Model tersebut

kemudian menghitung probabilitas setiap kategori risiko (rendah, sedang, tinggi) dan mengembalikan hasilnya ke frontend untuk ditampilkan kepada pengguna. Sistem ini juga terhubung ke database untuk menyimpan data pengguna dan riwayat prediksi. Dengan begitu, proses prediksi dapat dilakukan secara otomatis, real-time, dan konsisten.

Gambar 2 Tampilan awal web

Halaman Registrasi Pengguna

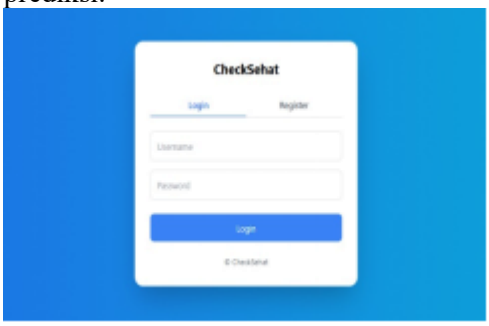
Halaman Registrasi berfungsi sebagai pembuatan akun pengguna. Pada halaman ini, pengguna diminta untuk memasukkan Nama Lengkap, Username, Email, dan Password. Sistem melakukan validasi terhadap Username dan Email untuk memastikan bahwa Username dan Email belum terdaftar. Password disimpan dalam bentuk terenkripsi untuk menjaga keamanan data pengguna.



Gambar 3 Tampilan Registrasi

Halaman Login Pengguna

Setelah proses registrasi telah berhasil, pengguna dapat langsung melakukan login dengan memasukkan Username dan Password yang telah didaftarkan.Sistem akan memverifikasi kecocokan data dengan database menggunakan proses autentikasi yang aman. Jika berhasil, sistem membuat sesi login sehingga pengguna dapat mengakses fitur kuesioner dan riwayat prediksi.



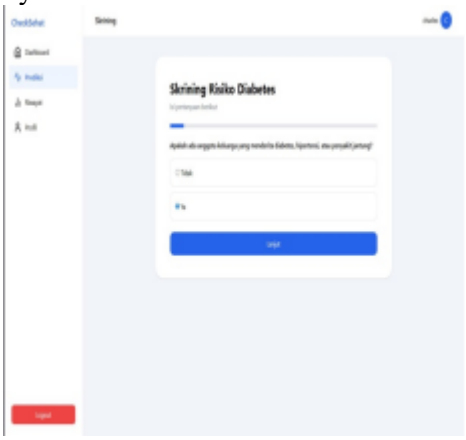
Gambar 4 Tampilan Login

Form Kuesioner Kesehatan

Form kuesioner merupakan bagian inti dari sistem karena berisi 12 pertanyaan faktor risiko yang digunakan sebagai input model. Variabel tersebut meliputi gaya hidup, kebiasaan makan, stres, gejala fisik, riwayat keluarga, dan kondisi penyakit tertentu. Semua

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
April 2026, IX (2): 2272 – 2281 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

pertanyaan dibuat dalam bentuk pilihan (Ya/Tidak) atau (Jarang, Sering, Sangat Sering) sehingga mudah diisi oleh pengguna. Ketika pengguna menekan tombol submit, seluruh jawaban akan dikirim ke backend untuk diproses melalui pipeline. Sistem mengonversi data menjadi nilai numerik (misalnya Ya = 1, Tidak = 0), menyesuaikan format jawaban agar konsisten, dan mempersiapkannya dalam bentuk vektor fitur yang siap dibaca oleh model Naive Bayes.



Gambar 5 Tampilan Pertanyaan

Tampilan Hasil Prediksi Penyakit

Gambar 6 Tampilan Hasil Prediksi

Pada halaman ini, pengguna dapat melihat tingkat risiko penyakit berdasarkan data kuesioner yang telah diisi sebelumnya. Hasil prediksi ditampilkan dalam bentuk kategori risiko dan persentase probabilitas yang menggambarkan keyakinan model terhadap setiap kelas risiko. Ketika data pengguna selesai diproses melalui pipeline dan dianalisis oleh model, sistem akan menghasilkan tiga nilai probabilitas yaitu untuk risiko rendah, risiko sedang, dan risiko tinggi. Pada pengguna yang memperoleh kategori Risiko Tinggi, sistem memberikan informasi tambahan berupa jenis penyakit yang paling mungkin muncul, yaitu penyakit jantung, hipertensi, atau diabetes.Nilai probabilitas ini dihitung berdasarkan perhitungan prior dan likelihood dari algoritma Naive Bayes. Kategori dengan probabilitas tertinggi dianggap sebagai hasil prediksi akhir.

Gambar 7 Tampilan Grafuk Perbandingan

Tampilan Grafik Perbandingan Probabilitas Sistem

Sistem menampilkan hasil prediksi utama kondisi dengan nilai probabilitas tertinggi. Grafik perbandingan probabilitas merupakan bagian dari sistem yang digunakan untuk menyajikan hasil prediksi dalam format visual. Grafik ini dirancang untuk memberikan perbandingan nilai probabilitas untuk setiap penyakit, yang dihitung menggunakan metode Naive Bayes. Berdasarkan pengujian sistem, ditemukan hal-hal berikut:

Diabetes adalah hasil prediksi yang dominan

Ditampilkan dengan label “DOMINAN”

Memiliki nilai probabilitas sebesar 57,58%

Informasi ini merupakan hasil dari perbandingan nilai probabilitas yang dihitung menggunakan metode Naive Bayes terhadap seluruh kategori penyakit.

Selain itu, sistem juga menampilkan detail probabilitas:

Diabetes: 57.58%
Hipertensi: 57.12%
Jantung: 52.48%

Tampilan Riwayat Data Pengguna

Gambar 8 Tampilan data Pengguna

Tampilan riwayat data pengguna merupakan bagian dari sistem yang dirancang untuk menyimpan dan menampilkan hasil prediksi bagi setiap pengguna yang telah mengisi kuesioner. Halaman ini menampilkan data dalam format tabel terstruktur, sehingga memudahkan pemantauan dan analisis hasil prediksi.

Berdasarkan hasil implementasi sistem, ditampilkan sebanyak 20 data uji pengguna yang telah diproses menggunakan metode Naive Bayes. Setiap data menampilkan informasi sebagai berikut:

1. Nama pengguna
2. Umur
3. Status hasil (berisiko)
4. Penyakit dominan
5. Probabilitas masing-masing penyakit (diabetes, hipertensi, jantung)
6. Persentase tertinggi
7. Waktu prediksi

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penelitian ini berhasil membangun sistem prediksi risiko penyakit berbasis machine learning menggunakan metode Naive Bayes dengan pipeline pengolahan data kuesioner. Data yang digunakan sebanyak 100 responden yang dibagi menjadi 80 data latih dan 20 data uji. Proses pipeline meliputi data cleaning, transformasi data, dan pembentukan dataset tabular sehingga data siap digunakan pada proses pemodelan.

Metode Naive Bayes digunakan untuk menghitung probabilitas prior, likelihood, dan posterior berdasarkan data latih. Hasil perhitungan probabilitas tersebut digunakan untuk mengklasifikasikan tingkat risiko penyakit ke dalam kategori risiko rendah, sedang, dan tinggi. Model yang telah dibangun kemudian diimplementasikan ke dalam sistem informasi berbasis web sehingga pengguna dapat memperoleh hasil prediksi secara otomatis.

Sistem yang dikembangkan mampu membantu dalam memprediksi risiko penyakit tidak menular seperti diabetes, hipertensi, dan penyakit jantung berdasarkan faktor risiko dari data kuesioner. Dengan demikian, metode Naive Bayes dan pipeline pengolahan data kuesioner dapat digunakan sebagai pendukung pengambilan keputusan dalam deteksi dini risiko penyakit.

DAFTAR PUSTAKA

[1] World Health Organization, “Noncommunicable diseases (Fact sheet),” WHO, 2025.
[2] Artika, Resti Dwi, et al. "Perancangan Data Pipeline Untuk Analisis Pola Perjalanan dan Permintaan Layanan Transjakarta." Jurnal Informatika dan Teknik Elektro Terapan 13.3S1 (2025).
[3] I. H. Sarker, “A review: Data pre-processing and data augmentation

techniques," Array, vol. 16, p. 100273, 2022.

[4] IBM, IBM SPSS Modeler CRISP-DM Guide. IBM Documentation, 2023.

[5] Musababa, M. A., & Fachrie, M. (2025). Data Streaming Pipeline Model Using DBSTREAM-Based Online Machine Learning for E-Commerce User Segmentation. Journal of Applied Informatics and Computing, 9(6).

[6] Adam, F.A.B, Berliana R, Khoirun N. (2024). Penerapan Metode Naïve Bayes dengan SMOTE pada Sistem Pendukung Keputusan untuk Prediksi Risiko Stroke.

[7] Amelia, R., Rozi, F., Anggraini, D., & Rosyani, P. (2025). Perbandingan Model Machine Learning dalam Prediksi Penyakit Jantung dengan Optimalisasi Fitur Gejala dan Faktor Risiko. Jurnal Pengabdian Masyarakat dan Riset Pendidikan

[8] Yu, X., et al. (2026). Web-based cardiovascular disease risk prediction using machine learning: Integrating feature evaluation and SHAP-based visualizations. Frontiers in Artificial Intelligence.

[9] Rahmada, A., & Susanto, E. R. (2025). Prediksi Risiko Penyakit Jantung Sederhana Menggunakan Algoritma Random Forest Classifier dengan Data Gaya Hidup Siswa. Jurnal Manajemen Informatika Jayakarta.

[10] Hidayat, R., et al. (2024). Implementasi Machine Learning Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine. BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer.

[11] Sihombing, P. R., & Yuliati, I. F. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer.

[12] Pratama, A. S., et al. (2024). "Sistem Pendukung Keputusan Pemilihan Karyawan Terbaik Menggunakan Integrasi Metode AHP dan Algoritma Random Forest." Jurnal Sistim Informasi dan Teknologi.

[13] Santoso, B., & Wijaya, K. (2026). "Optimasi Pipeline Preprocessing pada Data Survei Skala Besar Menggunakan Framework Scikit-Learn." Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI).

[14] Ramadhan, M. F., et al. (2024). "Pengembangan Model Predictive Analytics untuk Penentuan Penerima Beasiswa Menggunakan Pipeline Otomatis." Jurnal Media Informatika Budidarma.

[15] Huda, N., & Amalia, R. (2025). "Implementasi Pipeline Machine Learning untuk Klasifikasi Kepuasan Pelanggan Berdasarkan Data Kuesioner Digital." Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK).

[16] Azeraf, E., Monfrini, E., & Pieczynski, W. (2021). Using the Naïve Bayes as a discriminative model. 2021