IMPLEMENTASI METODE NAÏVE BAYES UNTUK KLASIFIKASI RISIKO PENYAKIT HIPERTENSI

LAPORAN TUGAS AKHIR

Laporan ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh
Gelar Sarjana Strata 1 (S1) pada Program Studi Teknik Informatika
Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

Disusun Oleh :

MUHAMMAD NAFID ZANIS

32602100083

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM SULTAN AGUNG

SEMARANG

2025

IMPLEMENTATION NAÏVE BAYES METHOD FOR CLASSIFICATION OF HYPERTENSION DISEASE RISK

FINAL PROJECT

Proposed to complete the requirement to obtain a bachelor's degree (S1) at Informatics Engineering of Industrial Technology Faculty Sultan Agung Islamic University

Arranged By :

MUHAMMAD NAFID ZANIS

32602100083

MAJORING OF INFORMATICS ENGINEERING

INDUSTRIAL TECHNOLOGY FACULTY

SULTAN AGUNG ISLAMIC UNIVERSITY

SEMARANG

2025

LEMBAR PENGESAHAN
TUGAS AKHIR

IMPLEMENTASI METODE NAÏVE BAYES UNTUK KLASIFIKASI RISIKO PENYAKIT HIPERTENSI

MUHAMMAD NAFID ZANIS
32602100083

Telah dipertahankan di depan tim penguji ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 2 Juni 2025

TIM PENGUJI UJIAN SARJANA :

Moch. Taufik, ST., MIT NIDN. 0622037502 (Ketua Penguji)	.....	04/06/2025
Arief Marwanto, ST., M.Eng., Ph.D NIDN. 0628097501 (Anggota Penguji)	.....	03/06/2025
Ghufron, ST., M.Kom NIDN. 0602079005 (Pembimbing)	.....	03/06/2025

Semarang, 4 Juni 2025

Mengetahui,
Kaprodi Teknik Informatika
Universitas Islam Sultan Agung

Moch. Taufik, ST., MIT
NIDN. 0622037502

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Muhammad Nafid Zanis

NIM : 32602100083

Judul Tugas Akhir : Implementasi Metode Naïve Bayes Untuk Klasifikasi Risiko Penyakit Hipertensi

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 4 Juni 2025

Yang Menyatakan,

Muhammad Nafid Zanis

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : Muhammad Nafid Zanis

NIM : 32602100083

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Alamat Asal : Kudus

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : Implementasi Metode Naïve Bayes Untuk Klasifikasi Risiko Penyakit Hipertensi

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 4 Juni 2025

Yang menyatakan,

Muhammad Nafid Zanis

KATA PENGANTAR

Dengan mengucap syukur ahamdulillah atas kehadirat Allah SWT yang telah memberikan rahmat dan hidayahnya kepada penulis sehingga dapat menyelesaikan Tugas Akhir dengan judul "Implementasi Metode Naïve Bayes Untuk Klasifikasi Risiko Penyakit Hipertensi" ini untuk memenuhi salah satu syarat menyelesaikan studi serta dalam rangka memperoleh gelar sarjana (S1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang.

Tugas Akhir ini disusun dan dibuat dengan adanya bantuan dari berbagai pihak, materi maupun teknis. Oleh karena itu, saya selaku penulis mengucapkan terima kasih kepada :

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, SH., MH., yang mengizinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri UNISSULA Ibu Dr. Ir. Hj. Novi Marlyana ST., MT., IPU., ASEAN.Eng
3. Dosen Pembimbing Bapak Ghufron, ST., M.Kom yang telah membimbing dan memberikan banyak nasehat dan saran.
4. Orang tua Bapak Zaenal Muttaqin dan Ibu Anisah. S.Pd.I yang menjadi *support system* dan mengizinkan untuk menyelesaikan laporan ini.
5. Teman – teman saya Jihan Tri Anggarjita, Rifa Qurotul Laili, Shabrina Isma Rasyida, Siti Nova Romadhani dan Fasikhullisan, ST,. MT dan semua pihak yang tidak dapat penulis sebutkan satu persatu atas dukungan, doa dan bantuannya.

Dengan segala kerendahan hati, penulis menyadari masih terdapat banyak kekurangan dari segi kualitas dan kuantitas maupun ilmu pengetahuan dalam penyusunan laporan sehingga penulis mengharapkan adanya kritikan dan saran yang bersifat membangun demi kesempurnaan laporan di masa mendatang.

Semarang , 4 Juni 2025

Muhammad Nafid Zanis

DAFTAR ISI

COVER ..... **i**
LEMBAR PENGESAHAN PENGUJI ..... **iii**
SURAT PERNYATAAN KEASLIAN TUGAS AKHIR ..... **iv**
LEMBAR PERSETUJUAN PUBLIKASI TUGAS AKHIR ..... **v**
KATA PENGANTAR ..... **vi**
DAFTAR ISI ..... **vii**
DAFTAR GAMBAR ..... **ix**
DAFTAR TABEL ..... **x**
ABSTRAK ..... **xi**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Batasan Masalah 2
1.4 Tujuan Penelitian 3
1.5 Manfaat Penelitian 3
1.6 Sistematika Penulisan 3
BAB II TINJAUAN PUSTAKA DAN DASAR TEORI **4**
2.1 Tinjauan Pustaka 4
2.2 Dasar Teori 6
2.2.1 Klasifikasi 6
2.2.2 Data Mining 6
2.2.3 Naïve Bayes 8
2.2.4 Hipertensi 9
BAB III METODE PENELITIAN **10**
3.1 Metode Penelitian 10
3.1.1 Studi Literatur 11
3.1.2 Pengumpulan Data 11
3.1.3 Pra-Pemrosesan Data 13
3.1.4 Pelatihan Model Naïve Bayes 14

3.1.5 Pengujian Skenario 16
3.1.6 Evaluasi Model 17
3.1.7 Deploy Model 19
3.2 Perancangan Sistem 20
3.3 Perancangan Kebutuhan 21
BAB IV HASIL DAN PEMBAHASAN 24
4.1 Hasil dan Analisis Penelitian 24
4.1.1 Pra-Pemrosesan Data 24
4.1.2 Implementasi Metode Naïve Bayes 26
4.1.3 Pengujian Skenario dataset 39
4.1.4 Evaluasi Model 40
4.2 Deploy Model ke Sistem 44
4.2.1 Menyimpan Model 44
4.2.2 Pembuatan sistem 45
4.2.3 Proses deploy model 47
4.3 Pengujian Sistem 49
BAB V KESIMPULAN DAN SARAN 50
5.1 Kesimpulan 50
5.2 Saran 50
DAFTAR PUSTAKA 51

DAFTAR GAMBAR

Gambar 3. 1 Alur diagram penelitian 10
Gambar 3. 2 Kumpulan data penyakit hipertensi 11
Gambar 3. 3 Alur diagram model Naive Bayes (Surorejo *dkk.*, 2022) 14
Gambar 3. 4 Alur diagram dari sistem yang dibuat 20
Gambar 4. 1 Mengecek *missing value* 24
Gambar 4. 2 Imputasi *missing value* dengan mean dan modus 25
Gambar 4. 3 Identifikasi nilai *outlier* 25
Gambar 4. 4 Penanganan pada nilai *outlier* 26
Gambar 4. 5 Membaca data latih 26
Gambar 4. 6 Tabel confusion matrix dari split data 80:20 41
Gambar 4. 7 Menampilkan data prediksi model yang benar 43
Gambar 4. 8 Menampilkan data prediksi model yang salah 43
Gambar 4. 9 Kode program menyimpan model 44
Gambar 4. 10 Menyimpan model ke google drive 45
Gambar 4. 11 Pembuatan sistem 46
Gambar 4. 12 Tampilan antarmuka pada website streamlit 46
Gambar 4. 13 Repository proyek pada github 47
Gambar 4. 14 Pustaka yang dibutuhkan 47
Gambar 4. 15 Proses deploy model ke streamlit cloud 48
Gambar 4. 16 Tampilan antarmuka website setelah dilakukan proses deploy 48

DAFTAR TABEL

Tabel 3. 1 Ringkasan dari atribut dataset penyakit hipertensi 12
Tabel 3. 2 Rasio *split* dataset 13
Tabel 3. 3 Skenario proporsi data latih dan data uji 17
Tabel 3. 4 Confusion matrix dari evaluasi model 18
Tabel 4. 1 Menghitung probabilitas kelas 27
Tabel 4. 2 Menghitung probabilitas atribut jenis kelamin 27
Tabel 4. 3 Menghitung probabilitas atribut status perokok 28
Tabel 4. 4 Menghitung probabilitas atribut penggunaan obat tekanan darah rendah 28
Tabel 4. 5 Menghitung probabilitas atirbut riwayat diabetes 29
Tabel 4. 6 Menghitung rata-rata dari setiap atribut 30
Tabel 4. 7 Menghitung standar deviasi tiap atribut 30
Tabel 4. 8 Sampel data uji 30
Tabel 4. 9 Parameter variabel umur tiap kelas 31
Tabel 4. 10 Parameter variabel batang rokok tiap kelas 32
Tabel 4. 11 Parameter variabel total kolesterol tiap kelas 34
Tabel 4. 12 Parameter variabel tekanan darah sistolik tiap kelas 35
Tabel 4. 13 Parameter variabel tekanan darah diastolik tiap kelas 36
Tabel 4. 14 Parameter variabel BMI tiap kelas 36
Tabel 4. 15 Parameter variabel detak jantung tiap kelas 37
Tabel 4. 16 Parameter variabel glukosa tiap kelas 37
Tabel 4. 17 Evaluasi model pada skenario dataset yang berbeda 39
Tabel 4. 18 TP, FP, FN, TN dari model tiap kelas 41
Tabel 4. 19 Perhitungan presisi, recall, dan f1-score model tiap kelas 42
Tabel 4. 20 Hasil pengujian menggunakan *blackbox testing* 49

ABSTRAK

Hipertensi merupakan gangguan kesehatan kronis yang ditunjukkan oleh peningkatan tekanan darah dalam arteri dan sering kali tidak disadari oleh penderitanya karena minimnya gejala awal, sehingga disebut sebagai silent killer. Faktor penyebabnya meliputi gaya hidup tidak sehat, seperti stres, obesitas, merokok, konsumsi alkohol, makanan berlemak tinggi, serta faktor usia dan keturunan. Seiring perkembangan teknologi, metode data mining digunakan untuk membantu proses deteksi penyakit, salah satunya melalui teknik klasifikasi. Penelitian ini memanfaatkan algoritma Naïve Bayes untuk mengklasifikasikan tingkat risiko hipertensi. Atribut yang digunakan berupa *sex, age,currentSmoker, cigsPerDay, BPMeds, diabetes, totChol, sysBP, diaBP, heartRate, BMI, glucose*. Akurasi tertinggi didapat pada skenario ketiga mencapai sebesar 89% dengan rasio data latih dan data uji 80:20. Hasil yang diperoleh menunjukkan bahwa penerapan metode Naïve Bayes mampu melakukan klasifikasi dengan performa yang tinggi.

Kata Kunci : Data Mining, Hipertensi, Klasifikasi, Naïve Bayes

ABSTRACT

Hypertension is a chronic health disorder indicated by increased blood pressure in the arteries and is often not realized by sufferers due to minimal early symptoms, so it is called a silent killer. The causative factors include unhealthy lifestyles, such as stress, obesity, smoking, alcohol consumption, high-fat foods, as well as age and hereditary factors. Along with the development of technology, data mining methods are used to help the disease detection process, one of which is through classification techniques. This study uses the Naïve Bayes algorithm to classify the risk level of hypertension. The attributes used are sex, age, currentSmoker, cigsPerDay, BPMeds, diabetes, totChol, sysBP, diaBP, heartRate, BMI, glucose. The highest accuracy was obtained in the third scenario reaching 89% with a training data and test data ratio of 80:20. The results obtained showed that the application of the Naïve Bayes method was able to perform classification with high performance.

Keyword : *Data Mining*, *Hypertension, Classification, Naïve Bayes*

BAB I
PENDAHULUAN

1.1 Latar Belakang

Tekanan darah tinggi atau hipertensi adalah gangguan kronis yang terjadi ketika tekanan darah dalam arteri meningkat secara terus-menerus. Berdasarkan panduan dari *American Heart Association* (AHA), seseorang diklasifikasikan sebagai pengidap penyakit hipertensi apabila tekanan darah sistoliknya tercatat di angka ≥130 mmHg atau tekanan darah diastoliknya tercatat di angka ≥80 mmHg. Hipertensi kerap disebut sebagai *silent killer* karena umumnya penderitanya tidak menyadari bahwa mereka mengidap kondisi tersebut sebelum dilakukan pemeriksaan darah. Kebiasaan hidup yang tidak sehat menjadi penyebab utama terjadinya hipertensi seperti stres, obesitas, merokok, konsumsi alkohol, serta makanan yang mempunyai tinggi lemak. Faktor lain seperti usia, memiliki riwayat hipertensi dalam keluarga, jenis kelamin juga bisa mempengaruhi terjadinya hipertensi. (Setiandari L.O, 2022)

Menurut data dari *World Health Organization* (WHO), penyakit hipertensi menjadi salah satu penyebab utama kematian di dunia dengan jumlah kematian lebih dari 10 juta jiwa setiap tahunnya. Kondisi ini menunjukkan tren peningkatan, khususnya di negara-negara berkembang termasuk Indonesia. Data dari Riset Kesehatan Dasar (Riskesdas) tahun 2018 menunjukkan bahwa tingkat prevalensi hipertensi di Indonesia telah mencapai 34,1%. Angka ini diperkirakan akan terus meningkat secara signifikan apabila upaya pencegahan dan penanganan tidak dilakukan secara optimal. (Yusup dan Rijanto, 2024)

Perkembangan teknologi yang semakin pesat membuat komunitas bidang medis terbantu dengan adanya teknologi yang mampu mendeteksi penyakit secara tepat dan akurat. Alternatif yang digunakan yaitu dengan memanfaatkan beberapa informasi data dan membuat model untuk melakukan prediksi penyakit hipertensi dengan teknik klasifikasi dalam data mining. Sebagai salah satu teknik data mining, klasifikasi bertujuan untuk pengelompokan nilai

variabel yang belum diketahui dengan merujuk pada variabel-variabel yang sudah diketahui sebelumnya. (Riany dan Testiana, 2023)

Dalam beberapa tahun terakhir, teknologi *machine learning* telah berkembang pesat dan digunakan sebagai alat bantu untuk memprediksi penyakit. Salah satu algoritma yang sering digunakan adalah Naïve Bayes. Penelitian yang dilakukan oleh (Rinanda *dkk.*, 2022) melakukan perbandingan antara metode Naïve Bayes dengan KNN mendapatkan hasil akurasi sebesar 75,78% untuk metode Naïve Bayes dibandingkan KNN yang mendapatkan akurasi 74,48%.

Sebagaimana telah diuraikan dalam penelitian sebelumnya mengenai latar belakang masalah, penulis tertarik meneliti Naïve Bayes untuk digunakan sebagai metode klasifikasi tingkat risiko penyakit hipertensi ke dalam klasifikasi risiko rendah atau risiko tinggi.

1.2 Rumusan Masalah

Berdasarkan penjabaran latar belakang diatas, maka dirumuskan masalah utama sebagai berikut yaitu bagaimana merancang sebuah sistem yang dapat melakukan klasifikasi tingkat risiko penyakit hipertensi pada individu ke dalam dua kategori, yaitu risiko rendah dan risiko tinggi dengan menggunakan metode Naïve Bayes.

1.3 Batasan Masalah

1. Kumpulan data yang digunakan diambil dari situs Kaggle berjumlah 4.240 data yang terdiri dari atribut *male, age, current smoker, cigsPerDay, BPMeds, diabetes, totChol, diaBP, BMI, heartRate, glucose,* dan *risk.*
2. Metode yang digunakan adalah *Naïve Bayes.*
3. Label kelas yang digunakan terbatas pada 2 kelas yaitu risiko rendah dan risiko tinggi.
4. Menampilkan *output* dari sistem yaitu menunjukkan hasil klasifikasi risiko rendah dan risiko tinggi.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah membuat suatu sistem yang mampu melakukan klasifikasi risiko penyakit hipertensi ke dalam kategori risiko rendah dan risiko tinggi menggunakan metode *Naïve Bayes*.

1.5 Manfaat Penelitian

1. Sebagai alat bantu alternatif tenaga medis dan masyarakat dalam mengidentifikai risiko penyakit hipertensi lebih cepat dan efisien.
2. Membantu memberikan informasi awal dalam menentukan diagnosis risiko penyakit hipertensi sehingga dapat dilakukan penanganan lebih lanjut

1.6 Sistematika Penulisan

Sistematika penulisan yang akan digunakan oleh peneliti dalam pembuatan laporan tugas akhir adalah sebagai berikut :

BAB I : PENDAHULUAN

Pada BAB I menjelaskan tentang latar belakang, pemilihan judul, rumusan masalah, batasan masalah, tujuan penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II : TINJAUAN PUSTAKA DAN DASAR TEORI

Pada BAB II memuat tentang penelitian terdahulu dan landasan teori yang berkaitan untuk membantu memahami konsep algoritma *Naïve Bayes,* klasifikasi, data mining, dan hipertensi untuk melengkapi penelitian ini.

BAB III : METODE PENELITIAN

Pada BAB III menjelaskan proses penelitian yang dimulai dari pengumpulan data hingga evaluasi hasil klasifikasi.

BAB IV : HASIL DAN ANALISIS PENELITIAN

Pada BAB IV berisi tentang pemaparan hasil penelitian yang dimulai dari hasil akhir sistem, klasifikasi data uji dan akurasi dari sistem.

BAB V : KESIMPULAN DAN SARAN

Pada BAB V merangkum keseluruhan proses penelitian dari awal sampai akhir yang telah dilakukan oleh penulis.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Untuk mendukung penelitian mengenai klasifikasi penyakit hipertensi, terdapat beberapa sumber penelitian terdahulu yang berkaitan baik secara langsung maupun tidak langsung dengan topik penelitian ini. Beberapa di antaranya adalah sebagai berikut :

Penelitian yang berjudul "Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network" (Purwono *dkk.*, 2022) yang bertujuan mengembangkan model prediksi otomatis untuk penyakit hipertensi. Atribut yang digunakan meliputi jenis kelamin, usia, tinggi, berat badan, tekanan darah diastolik, tekanan darah sistolik, denyut jantung, dan BMI dengan menerapkan metode Artificial Neural Network mendapat tingkat akurasi sebesar 85%.

Penelitian yang berjudul "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest Dan Naïve Bayes" (Kharits *dkk.*, 2023) bertujuan untuk mengevaluasi kinerja dua algoritma, yaitu Random Forest dan Naïve Bayes. Atribut yang digunakan yaitu age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, dan thal. Berdasarkan hasil penelitian, algoritma Naïve Bayes menunjukkan tingkat akurasi yang rendah dibandingkan Random Forest dengan akurasi yang diperoleh 100% , sementara Naïve Bayes mencapai 85,90%.

Penelitian dengan judul "Prediksi Penyakit Hipertensi Menggunakan Machine Learning Dengan Algoritma Regresi Logistik" (Tarimana *dkk.*, 2024). Atribut yang digunakan meliputi usia, tekanan darah diastolik, indeks massa tubuh (BMI), kadar kolesterol, dan riwayat hipertensi menggunakan model Regresi Logistik mendapatkan tingkat akurasi sebesar 91%.

Penelitian dengan judul "Klasifikasi Penyakit Hipertensi Menggunakan Metode Svmgrid Search dan Svm Genetic Algorithm (GA)" yang dilakukan oleh (Awalullaili *dkk.*, 2022). Atribut yang digunakan meliputi jenis kelamin, usia, konsumsi garah berlebih, kolestrol tinggi, konsumsi rokok, dan riwayat hipertensi keluarga menggunakan metode SVM Grid Search dan SVM Genetic Algorithm mendapatkan akurasi sebesar 89,92% pada kernel RBF.

Penelitian dengan judul "Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier" yang dilakukan oleh (Uswatun Khasanah *dkk.*, 2022). Tujuan dari penelitian ini adalah untuk mengetahui hasil klasifikasi pasien ke dalam kategori diabetes 'ya' atau 'tidak' menggunakan model *Naïve Bayes* Classifier. Dataset yang digunakan diambil dari data pasien di Rumah Sakit Dirgahayu Samarinda pada tahun 2018 sampai 2021 dengan jumlah 130 data dibagi menjadi 4 skenario pengujian berbeda, yaitu skenario pertama perbandingan data testing dan data training 60:40, skenario kedua dengan perbandingan data testing dan data training 70:30, skenario ketiga perbandingan data testing dan data training 80:20, skenario keempat perbandingan data testing dan data training 90:10. Hasil penelitian tersebut didapatkan bahwa uji skenario dengan proporsi data testing 40% dan 20% mencapai nilai akurasi terbaik yaitu sebesar 92,31%.

Penelitian dengan judul "Klasifikasi Penyakit Paru-Paru Dengan Menggunakan Metode Naïve Bayes Classifier" yang dilakukan oleh (Haffandi *dkk.*, 2022). Tujuan dari penelitian ini mempermudah melakukan klasifikasi pada penyakit paru-paru. Dataset yang digunakan yaitu data medis pasien di Rumah Sakit Umum Daerah Mayjen H.A. Thalib Kota Sungai Penuh sebanyak 134 data pasien dengan 24 atribut dan 1 kelas. Hasil dari penelitian adalah algoritma Naïve Bayes efektif dalam melakukan klasifikasi penyakit paru-paru dengan nilai akurasi sebesar 97,06%.

Penelitian dengan judul "Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes" yang dilakukan oleh (Prasetya dan Sujatmiko, 2022). Penelitian tersebut bertujuan untuk membuat sistem yang digunakan untuk membantu mendeteksi penyakit diabetes. Pada penelitian tersebut, dilakukan perbandingan dua algoritma yaitu algoritma Naïve Bayes dan algoritma K-Nearest Neighbor. Hasil dari penelitian tersebut menyimpulkan bahwa metode Naïve Bayes mempunyai nilai akurasi tertinggi dibandingkan dengan metode K-Nearest Neighbor ditunjukkan dengan hasil akurasi sebesar 95%, sedangkan akurasi sebesar 93% dicapai dengan metode K-Nearest Neighbor.

Hasil penelitian sebelumnya menyimpulkan bahwa metode *Naïve Bayes* cukup efektif dalam melakukan klasifikasi terhadap beberapa jenis penyakit. Berdasarkan temuan tersebut, dalam penelitian ini bertujuan untuk membuat sebuah sistem yang dapat melakukan klasifikasi tingkat risiko penyakit hipertensi ke dalam kategori normal atau hipertensi menggunakan metode *Naïve Bayes* dengan dataset yang diperluas.

2.2 Dasar Teori

2.2.1 Klasifikasi

Klasifikasi merupakan metode pencarian model yang menggunakan analisis kumpulan data latih untuk menunjukkan dan membedakan kelas label suatu data. (Andriani *dkk.*, 2023). Klasifikasi memainkan peran penting dengan menentukan tingkat akurasi data yang diperoleh dengan memanfaatkan berbagai metode yang tersedia. (Akmal *dkk.*, 2023)

2.2.2 Data Mining

Metode data mining digunakan untuk mengekstraksi informasi tertentu dari sebuah *database*. Data mining dapat diterapkan melalui berbagai teknik seperti, *artificial intelligence*, *machine learning,* dan metode statistika yang bermanfaat dalam mengetahui informasi bentuk data dalam skala besar. (Yoliadi, 2023)

Data mining merupakan tahapan penting dalam proses KDD (*Knowledge Discovery Database* yang berperan dalam mengekstraksi pola bermakna dari kumpulan data besar yang tersimpan dalam basis data (Zai, 2022). KDD terdiri atas beberapa langkah sistematis seperti pengumpulan data, praproses untuk membersihkan data mentah, transformasi data, dan terakhir evaluasi. (Kartika *dkk.*, 2022)

1. *Data Selection*

   Tahap ini merupakan langkah awal sebelum proses penambangan data dilakukan. Data yang relevan dipilih dari basis data dan disimpan secara terpisah untuk keperluan analisis lebih lanjut.

2. *Pre-Processing*

   Proses ini melibatkan kegiatan pembersihan data seperti mengidentifkasi dan memperbaiki ketidakkonsistenan, menghapus data yang duplikat, memperbaiki kesalahan serta menambahkan informasi yang diperlukan guna meningkatkan kualitas data.

3. *Transformation*

   Pada tahap ini, data yang telah dipilih akan diubah ke dalam bentuk yang sesuai untuk dianalisis biasanya melalui proses normalisasi agar data berada dalam skala yang seragam dan siap digunakan dalam model.

4. Data Mining

   Ini merupakan tahapan inti dari proses KDD, di mana dilakukan pencarian pola atau pengetahuan yang tersembunyi dalam kata dengan menerapkan metode atau algoritma tertentu. Keberhasilan proses ini sangat tergantung pada metode yang digunakan.

5. *Evaluation*

   Tahapan ini bertujuan untuk menilai dan menyajikan hasil penambangan data. Evaluasi dilakukan untuk mengetahui seberapa baik kinerja model dan untuk memastikan bahwa pola yang ditentukan benar-benar bermanfaat dan relevan bagi pengguna akhir.

2.2.3 Naïve Bayes

Algoritma Naïve Bayes merupakan salah satu teknik dalam metode *machine learning* yang digunakan dalam proses klasifikasi data. Menurut (Siska *dkk.*, 2023), perhitungan metode Naïve Bayes didasarkan pada teorema bayes yang digunakan untuk memprediksi probabilitas dari suatu kejadian. Model Naïve Bayes terdiri dari beberapa langkah yang dimulai dengan pelatihan data dan diakhiri dengan pengujian data untuk menghasilkan keputusan yang akurat. Terdapat 3 tipe jenis model Naïve Bayes, diantaranya : (Damayanti *dkk.*, 2024)

1. Gaussian Naïve Bayes : digunakan pada data dengan atribut kontinyu yang memiliki distribusi gaussian atau normal.
2. Multinomial Naïve Bayes : ideal digunakan untuk data diskrit seperti hitungan kata dalam teks.
3. Bernoulli Naïve Bayes : digunakan untuk data biner, yaitu data yang memiliki nilai 0 atau 1.

Perhitungan metode Naïve Bayes menggunakan persamaan yang tercantum dalam rumus (1) :

$$P\ (A|E) = \frac{P\ (E|A)\ .\ P(A)}{P(A)} \quad (1)$$

Penjelasan dari keterangan rumus diatas adalah sebagai berikut :

- $P(A|E)$: Nilai peluang terjadinya A apabila E diketahui
- $P(E|A)$: Nilai peluang terjadinya E apabla A diketahui
- $P(A)$: Nilai peluang kejadian A
- $P(E)$: Probabilitas kejadian E

Selanjutnya, jika terdapat atribut yang memiliki nilai kontinu, perhitungan nilai posterior dilakukan dengan cara yang berbeda (Wie dan Siddik, 2022). Atribut yang memiliki nilai kontinu diasumsikan mengikuti distribusi Gaussian dengan menggunakan *mean* dan *standar deviasi* yang ditunjukkan pada persamaan (2) :

$$P\,(A) = \frac{1}{\sqrt{2\pi\sigma}}\, e - \frac{(A-\mu)^2}{2\sigma^2} \quad (2)$$

Keterangan dari persamaan rumus di atas adalah sebagai berikut :

$P(A)$ = peluang nilai A dalam distibusi normal

e = bilangan euler

μ = nilai rata-rata setiap atribut

σ = nilai standar deviasi setiap atribut

π = konstanta pi

2.2.4 Hipertensi

Hipertensi merupakan gangguan penyakit kronis yang ditandai oleh meningkatnya tekanan darah di arteri yang dapat berkembang secara perlahan tanpa disadari. Kondisi ini diukur dengan mempertimbangkan dua parameter, yaitu sistolik yang muncul ketika jantung berkontraksi dan diastolik yang terjadi saat jantung berelaksasi antara dua denyut. Menurut panduan dari American Heart Association (AHA), seseorang diklasifikasikan sebagai pengidap penyakit hipertensi apabila tekanan darah sistoliknya tercatat di angka ≥130 mmHg atau tekanan darah diastoliknya tercatat di angka ≥80 mmHg(Tarimana *dkk.*, 2024).

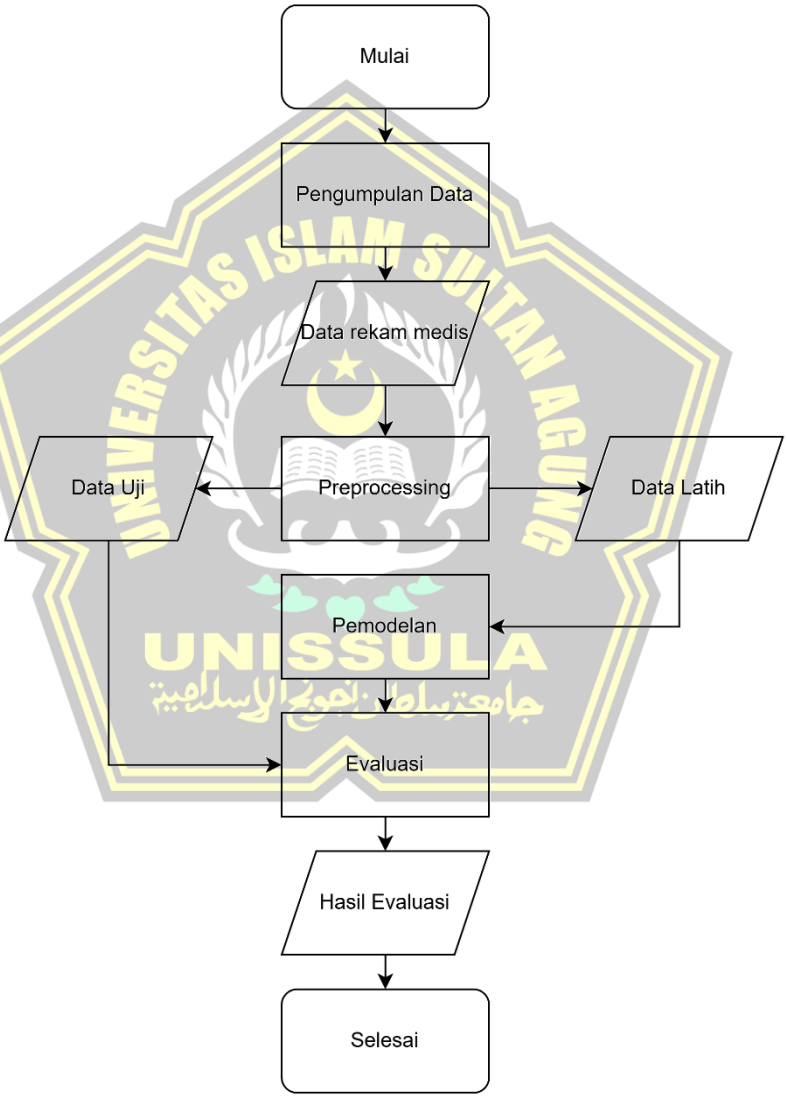
Penyebab hipertensi secara umum diklasifikasikan menjadi dua tipe utama, yaitu hipertensi primer dan hipertensi sekunder. Hipertensi primer merupakan jenis paling umum terjadi dan tidak memiliki penyebab yang pasti. Kondisi ini biasanya muncul secara bertahap dan berakitan dengan erat gaya hidup yang kurang sehat seperti kurang aktivitas fisik, dan obesitas. Sejumlah faktor yang dapat meningkatkan risiko seseorang mengalami hipertensi antaralain adalah usia, jenis kelamin, riwayat keluarga, obesitas, pola hidup, dan stres yang tinggi.

Hipertensi jika tidak segera ditangani, hipertensi dapat menyebabkan komplikasi yang cukup serius. Komplikasi yang umum terjadi mencakup penyakit kardiovaskular seperti stroke dan gagal jantung, kerusakan ginjal, gangguan penglihatan, serta aneurisma.

BAB III
METODE PENELITIAN

3.1 Metode Penelitian

Pada penelitian ini, menggunakan algoritma *Naïve Bayes* untuk proses klasifikasi tingkat risiko penyakit hipertensi. Penelitian ini memerlukan beberapa langkah yang harus dilakukan sebagai berikut :



Gambar 3. 1 Alur diagram penelitian (Putra *dkk.*, 2024)

3.1.1 Studi Literatur

Peneliti mencari berbagai sumber informasi mengenai teori yang mendasari serta berkaitan langsung dengan fokus penelitian ini. Teori yang dipelajari meliputi klasifikasi, data mining, Naïve Bayes, dan hipertensi yang bersumber dari artikel ilmiah, jurnal, buku, tugas akhir, serta situs *website* yang tersedia di situs internet.

3.1.2 Pengumpulan Data

Tahap awal dalam perancangan model klasifikasi adalah pengumpulan data. Pada penelitian ini, data yang digunakan merupakan data sekunder yang diperoleh dari situs web Kaggle dengan judul "Hypertension Risk Model Main". Dataset tersebut diunggah oleh MD Raihan Khan pada tahun 2024 dan terdiri dari 4.240 entri data. Setiap entri memiliki 12 atribut prediktor yang digunakan untuk memodelkan risiko hipertensi serta 1 atribut kelas sebagai label. Untuk keperluan pemrosesan lebih lanjut, dataset disimpan di Google Drive dengan nama file hypertension-risk-model.csv yang dapat diakses melalui tautan berikut : https://tinyurl.com/Hypertension-risk

	male	age	currentSmoker	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
0	1	39	0	0.0	1.0	1	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	0	0.0	0.0	1	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1	20.0	0.0	1	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	1	30.0	0.0	1	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	1	23.0	0.0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4235	0	48	1	20.0	NaN	0	248.0	131.0	72.0	22.00	84.0	86.0	0
4236	0	44	1	15.0	0.0	0	210.0	126.5	87.0	19.16	86.0	NaN	0
4237	0	52	0	0.0	0.0	1	269.0	133.5	83.0	21.47	80.0	107.0	0
4238	1	40	0	0.0	1.0	1	185.0	141.0	98.0	25.60	67.0	72.0	1
4239	0	39	1	30.0	1.0	0	196.0	133.0	86.0	20.91	85.0	80.0	0

Gambar 3. 2 Kumpulan data penyakit hipertensi

Tabel 3. 1 Ringkasan dari atribut dataset penyakit hipertensi

No.	Atribut	Tipe Data	Keterangan
1	*Male*	Kategorikal	Jenis Kelamin (1: Laki-Laki / 0: Perempuan).
2	*Age*	Numerik	Umur yang dinyatakan dalam tahun.
3	*CurrentSmoker*	Kategorikal	Perokok aktif (1: ya / 0: tidak)
4	*cigsPerDay*	Numerik	Jumlah batang rokok yang dihisap per hari
5	*BPMeds*	Kategorikal	Menyatakan penggunaan obat darah tinggi (1: ya / 0: tidak)
6	*Diabetes*	Katgeorikal	riwayat diabetes (1: ya / 0: tidak)
7	*totChol*	Numerik	Total kadar kolesterol diukur dalam satuan mg/dL
8	*sysBP*	Numerik	Tekan darah sistolik diukur dalam satuan mmHg
9	*diaBP*	Numerik	Tekanan darah diastolik diukur dalam satuan mmHg
10	*BMI*	Numerik	Menyatakan indeks massa tubuh diukur dalam satuan kg/m2
11	*heartRate*	Numerik	Menyatakan detak jantung

12	*Glucose*	Numerik	Menyatakan kadar glukosa dalam darah
13	*risk*	Kategorikal	Klasifikasi penyakit hipertensi (1: resiko tinggi / 0: resiko rendah)

3.1.3 Pra-Pemrosesan Data

Pemrosesan data dilakukan dengan tujuan meningkatkan kualitas data yang akan diteliti. Dalam penelitian ini, tahapan pemrosesan data yang akan dilakukan mencakup *missing value* dan *split* data. Pada proses *missing value* dilakukan dengan mengisi data yang memiliki nilai null dengan perhitungan *mean* untuk atribut numerik dan modus untuk atribut kategorikal.

Proses selanjutnya dilakukan *handling outlier* pada fitur numerik dataset. Penanganan nilai outlier dilakukan dengan menggunakan metode *Interquartile Range* (IQR) dengan menghitung selisih antara quartil ketiga (Q3) dan quartil pertama (Q1) dalam distribusi data. Persamaan rumus untuk menghitung nilai IQR adalah sebagai berikut :

$$[Q1 - 1.5\ x\ IQR, Q3 + 1.5\ x\ IQR]$$

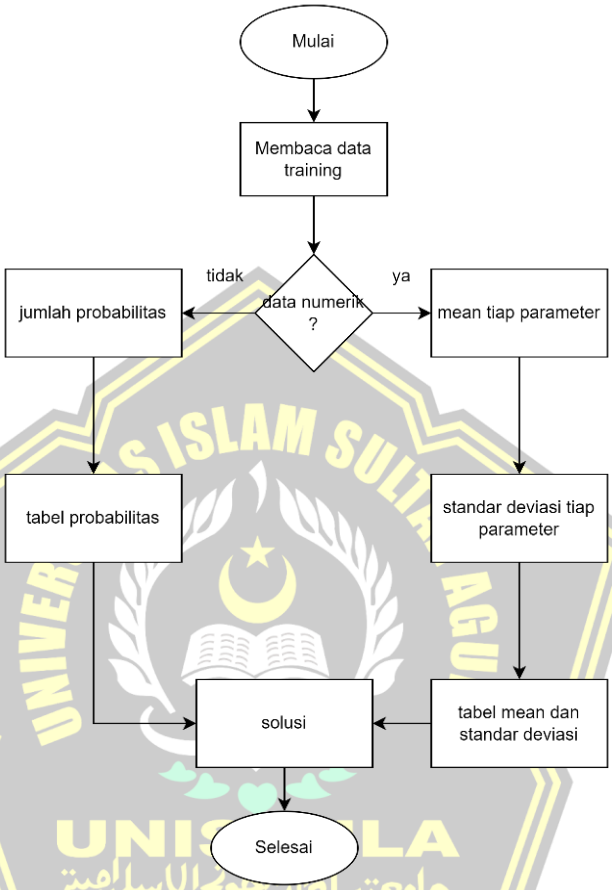
Selanjutnya, dilakukan *split data* yaitu membagi antara data training dan data testing dengan perbandingan 80% untuk data training dan 20% untuk data testing yang dapat dilihat pada tabel 3.2.Tujuan dari *data training* adalah untuk membantu model mengenali pola dalam data sedangkan data testing berfungsi untuk memastikan bahwa model yang telah dilatih memiliki kemampuan untuk memprediksi label yang belum dipelajari oleh model dengan baik.

Tabel 3. 2 Rasio *split* dataset

Keterangan	Rasio	Jumlah Data
Data training	80%	3392
Data testing	20%	848
Jumlah	**100%**	**4240**

3.1.4 Pelatihan Model Naïve Bayes

Pelatihan Metode yang digunakan dalam penelitian Klasifikasi Risiko Penyakit Hipertensi Menggunakan Metode Naïve Bayes dijelaskan pada flowchart gambar 3.3.



Gambar 3. 3 Alur diagram model Naive Bayes (Surorejo *dkk.*, 2022)

Alur dari pelatihan model klasifikasi risiko penyakit hipertensi menggunakan metode Naïve Bayes, yaitu :

1. Membaca data training
2. Kemudian mengecek apakah atribut dalam bentuk numerik atau tidak
3. Jika atribut bukan dalam bentuk numerik atau dalam bentuk kategorikal, maka atribut akan menghitung jumlah probabilitas risiko rendah dan probabilitas risiko tinggi. Kemudian akan mencetak hasil dari tabel probabilitas.

4. Selanjutnya, jika terdapat atribut yang memiliki nilai kontinu, perhitungan nilai posterior dilakukan dengan cara yang berbeda (Wie dan Siddik, 2022). Atribut yang memiliki nilai kontinu diasumsikan mengikuti distribusi Gaussian dengan menggunakan *mean* dan *standar deviasi* yang ditunjukkan pada persamaan (2) :

$$P\,(A) = \frac{1}{\sqrt{2\pi\sigma}}\, e - \frac{(A-\mu)^2}{2\sigma^2} \quad (3)$$

Keterangan dari persamaan rumus di atas adalah sebagai berikut :

$P(A)$ = peluang nilai A dalam distibusi normal

e = bilangan euler

μ = nilai rata-rata setiap atribut

σ = nilai standar deviasi setiap atribut

π = konstanta pi

Adapun formula untuk menghitung rata-rata atau *mean* ditunjukkan pada persamaan (3) :

$$\mu = \frac{A1 + A2 + A3 + \cdots + An}{n} \quad (3)$$

Keterangan dari persamaan rumus diatas adalah sebagai berikut :

μ = nilai rata-rata dari seluruh sampel

Ai = elemen ke -i dalam himpunan data

n = jumlah total sampel yang digunakan dalam perhitungan

Sementara itu, formula perhitungan standar deviasi dijelaskan secara rinci dalam persamaan yang disajikan pada rumus (4) :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(Ai-\mu)^2}{n-1}} \quad (4)$$

Keterangan dari persamaan rumus diatas adalah sebagai berikut :

σ = standar deviasi

Ai = nilai data ke -i pada kumpulan sampel

μ = rata-rata hitung dari seluruh data yang diamati

n = jumlah total sampel yang terlibat dalam analisis

5. Tahapan terakhir adalah menghitung nilai akumulasi peluang dari setiap kelas menggunaan persamaan (5) sebagai berikut : (Pridiptama *dkk.*, 2024)

$$P(A\,|X_1, \dots, X_n) = P(A).\prod_{i=1}^{n} P(X_i\,|A) \quad (5)$$

Keterangan dari persamaan rumus diatas adalah sebagai berikut :

$P(A\,|X_1, \dots, X_n)$ = probabilitas kelas A setelah melihat fitur Xi

$P(A)$ = probabilitas awal dari kelas A

$P(X_i|A)$ = probabilitas fitur Xi muncul jika kelasnya A

6. Mencari nilai confidence dari tiap kelas menggunakan persamaan (6) sebagai berikut : (Haffandi *dkk.*, 2022)

$$Confidence\ (c) = \frac{P(c)}{\sum P(c)} * 100 \quad (6)$$

Keterangan dari persamaan rumus diatas adalah sebagai berikut :

$Confidence(c)$ = nilai keyakinan model masuk ke kelas tertentu

$P(c)$ = probabilitas posterior kelas c

$\sum P(c)$ = jumlah total probabilitas posterior dari semua kelas

3.1.5 Pengujian Skenario

Dalam proses pengujian model klasifikasi, digunakan beberapa skenario pengujian dengan proporsi pembagian data latih dan data uji yang berbeda. Tujuan dari pengujian skenario tersebut adalah untuk mengevaluasi stabilitas dan kinerja model Naïve Bayes dalam kondisi data yang bervariasi serta memastikan bahwa model dapat bekerja secara konsisten. Beberapa skenario pembagian data yang digunakan dalam pengujian ini antara lain :

Tabel 3. 3 Skenario proporsi data latih dan data uji

	Proporsi Data (%)	Data Latih	Data Uji	Jumlah
Skenario 1	60%:40%	2.544	1.696	4.240
Skenario 2	70%:30%	2.968	1.272	
Skenario 3	80%:20%	3.392	848	
Skenario 4	90%:10%	3.816	424	

Setiap skenario dilakukan dengan cara membagi dataset secara acak berdasarkan proporsi yang ditentukan kemudian model dilatih menggunakan data latih dan diuji menggunakan data uji. Hasil pengujian kemudian dievaluasi berdasarkan metrik performa seperti akurasi, presisi, recall dan f1-score.

Dengan menggunakan skenario tersebut dapat diketahui bagaimana performa model dipengaruhi oleh jumlah data yang tersedia untuk pelatihan dan pengujian. Skenario juga berguna untuk menentukan pembagian data terbaik yang dapat menghasilkan performa model paling optimal.

3.1.6 Evaluasi Model

Evaluasi model merupakan proses yang dilakukan untuk menentukan kombinasi model yang paling efektif dalam memprediksi tingkat risiko penyakit hipertensi dengan tingkat akurasi tinggi. Salah satu metrik yang digunakan dalam proses evaluasi ini adalah akurasi yaitu ukuran yang menunjukkan seberapa benar model dalam melakukan prediksi secara keseluruhan. Pada model Naïve Bayes, evaluasis dilakukan menggunakan *confusion matrix*. *Confusion matrix* sendiri merupakan tabel yang menampilkan kinerja model berdasarkan perbandingan antara hasil prediksi dan nilai sebenarnya. (Fuad *dkk.*, 2023). Untuk menunjukkan hasil evaluasi, akurasi dan confusion matrix dihitung menggunakan rumus sebagai berikut :

Tabel 3. 4 Confusion matrix dari evaluasi model

		Kelas Prediksi	
		Positif	Negatif
Kelas Sebenarnya	Positif	*True Positif* (TP)	*False Negatif* (FN)
	Negatif	*False Positif* (FP)	*True Negatif* (TN)

Keterangan dari penjelasan tabel diatas adalah sebagai berikut:

- *True Positif* (TP) : jumlah kasus ketika model memprediksi kelas positif dan hasil aktualnya juga positif.
- *True Negatif* (TN) : jumlah kasus model ketika model memprediksi kelas negatif dan hasil aktualnya juga negatif.
- *False Positif* (FP) : merujuk pada jumlah kasus ketika model memprediksi kelas positif, padahal data sebenarnya temasuk dalam kelas negatif.
- *False Negatif* (FN) : adanya jumlah kasus ketika model memprediksi kelas negatif, namun data sebenarnya merupakan kelas positif.

Untuk menunjukkan hasil akurasi dari *confusion matrix* menggunakan persamaan rumus (5), (6), (7), dan (8) sebagai berikut :

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$Presisi = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1\ Score = \frac{2\ \times Precision\ \times Recall}{Precision + Recall} \quad (8)$$

A. Akurasi

Akurasi merupakan ukuran seberapa tepat model dapat memprediksi label kelas dari data uji. Indikator ini menunjukkan seberapa banyak prediksi yang benar dibandingkan dengan keseluruhan jumlah data.

B. Presisi

Presisi adalah rasio antara jumlah prediksi benar untuk kelas positif dibandingkan dengan total prediksi yang dikategorikan sebagai positif oleh model.

C. Recall

Recall mengukur kemampuan model dalam menemukan seluruh kasus aktual positif. Ini dihitung dengan membandingkan jumlah prediksi benar untuk kelas positif terhadap jumlah keseluruhan data aktual yang memang termasuk dalam kelas positif.

D. F1-score

F1-score adalah nilai rata-rata perbandingan dari presisi dan recall yang memberikan gambaran yang lebih seimbang jika terdapat ketidakseimbangan antar kelas.

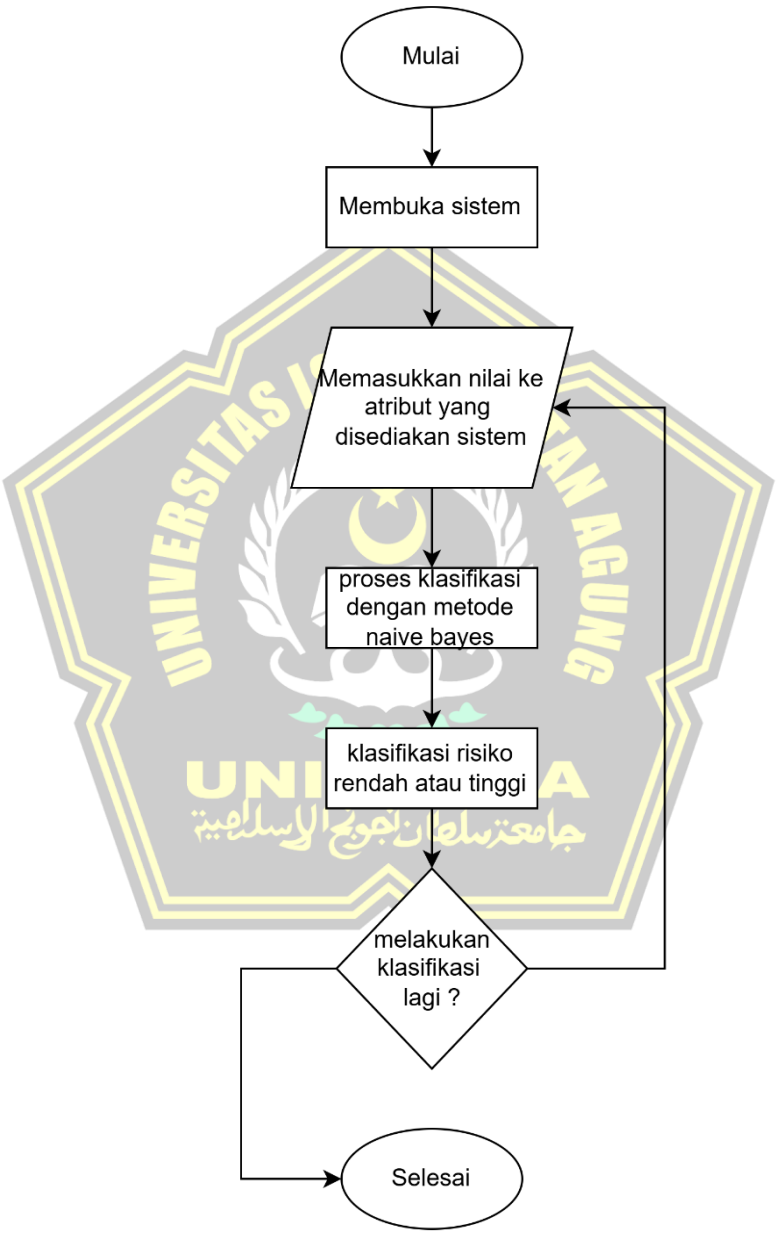
3.1.7 Deploy Model

Penelitian ini menggunakan perangkat lunak *streamlit* sebagai media untuk melakukan proses *deployment* pada model. *Streamlit* merupakan sebuah kerangka kerja yang berbasis pada pemrograman bahasa python yang dirancang untuk visualisasi data dan pengembangan antarmuka pengguna secara interaktif dan efisien (Hastomo *dkk.*, 2022). Platform ini menyediakan berbagai fitur seperti *input teks, input number, button, select box,* dan lain sebagainya. *Software* streamlit memiliki beberapa keunggulan yang diperoleh antara antara lain : (Syafiih, 2023)

1. Bersifat responsif dan dinamis
2. Merupakan perangkat lunak *open source*
3. Memiliki antarmuka yang sederhana
4. Mudah untuk dipelajari oleh pengguna
5. Mendukung integrasi visualisasi data *machine learning*

Berdasarkan keunggulan-keunggulan tersebut, *streamlit* dipilih dalam penelitian ini untuk proses *deployment* model karena kemampuannya dalam menyederhanakan pengujian model serta menampilkan hasil analisis secara interaktif.

3.2 Perancangan Sistem



Gambar 3. 4 Alur diagram dari sistem yang dibuat

Gambar 3.4 merupakan gambar alur kerja dari sistem klasifikasi yang dikembangkan. Proses dimulai ketika pengguna mengakses halaman web melalui tautan streamlit. Selanjutnya, pengguna diminta untuk mengisi nilai dari sejumlah atribut yang berhubungan dengan faktor risiko hipertensi seperti jenis kelamin, usia, status perokok aktif, jumlah konsumsi rokok per hari, penggunaan obat tekanan darah rendah, riwayat diabetes, kadar kolesterol total, tekanan darah sistolik dan diastolik, detak jantung, dan kadar glukosa. Setelah data di*input*kan, pengguna menekan tombol tes prediksi untuk memulai proses klasifikasi. Sistem kemudian memproses data tersebut menggunakan model Naïve Bayes yang telah dilatih sebelumnya dan menampilkan hasil prediksi apakah pengguna termasuk dalam kategori risiko hipertensi rendah atau tinggi. Proses alur sistem dianggap selesai setelah seluruh tahapan tersebut berhasil dijalankan.

3.3 Perancangan Kebutuhan

Dalam membangun model Naïve Bayes pada klasifikasi risiko penyakit hipertensi menggunakan beberapa *library* dari python diantaranya adalah :

1. Numpy

   Numpy merupakan salah satu pustaka(*library*) penting dalam bahasa pemrograman python yang digunakan untuk mendukung perhitungan numerik. *Library* ini tidak hanya menyediakan struktur data *array* multidimensi yang efisien, tetapi juga dilengkapi dengan berbagai fungsi matematis yang memudahkan pengguna dalam melakukan operasi perhitungan terhadap data dalam bentuk *array*.

2. Pandas

   Pandas adalah salah satu *library* dalam bahasa pemrograman python yang umum digunakan untuk keperluan manipulasi dan analisis data. *Library* ini menyediakan struktur data yang efisien dan mudah dipahami sehingga sangat membantu dalam pengolahan data. Salah satu struktur utamanya adalah DataFrame yang memiliki bentuk serupa dengan *spreadsheet* pada Microsfot Excel. Melalui pandas, pengguna

dapat melakukan berbagai operasi analisis data seperti pemilahan, penyaringan, pengelompokan, hingga pembuatan visualisasi data secara praktis.

3. Scikit-learn

   Scikit-learn yang djuga dikenal dengan nama sklearn, merupakan salah satu library python yang banyak digunakan dalam bidang pembelajaran mesin. *Library* ini menyediakan berbagai algoritma *machine learning* yang dirancang agar mudah digunakan namun tetap efektif dalam penerapannya. Selain itu, scikit-learn juga menawarkan berbagai fitur penting seperti fungsi untuk pra-pemrosesan data, validasi model serta evaluasi performa dari model yang dibangun.

4. Seaborn

   Seaborn merupakan salah satu *library* yang banyak digunakan untuk visualisasi data, dikenal karena kekuatan dan kemampuannya dalam membuat visualisasi yang menarik. *Library* ini dibangun dengan dasar matplotlib yang berfungsi sebagai alat utama untuk menghasilkan berbagai jenis plot dan visualisasi data. Dengan matplotlib, pengguna dapat membuat visualisasi data secara fleksibel dan efektif, mendukung berbagai kebutuhan analisis data.

5. Matplotlib

   Matplotlib merupakan salah satu *library* python yang paling banyak digunakan dan memiliki kekuatan besar dalam bidang visualisasi data. *Library* ini dirancang untuk menghasilkan berbagai jenis plot serta visualisasi data lainnya. Dengan matplotlib, pengguna dapat membuat visualisasi secara fleksibel dan efektif yang memungkinkan pengolahan data dengan cara yang sangat kuat.

6. *Framework* streamlit

   Streamlit merupakan salah satu *framework open source* dalam bahasa pemrograman python yang berfungsi untuk membangun aplikasi web khususnya dalam bidang *data science* dan *machine learning*. Pada penelitian ini, streamlit dimanfaatkan sebagai alat untuk melakukan *deployment* terhadap aplikasi yang dikembangkan menggunakan bahasa python sehingga aplikasi dapat diakses dan digunakan secara interaktif melalui antarmuka web.

7. Pickle

   Pickle merupakan pustaka bawaan dalam bahasa pemrograman python yang digunakan untuk melakukan proses serealisasi model deserialisasi terhadap objek python. Dalam konteks penelitian ini, pickle dimanfaatkan untuk menyimpan objek-objek seperti model machine learning ke dalam bentuk file serta memuat kembali objek tersebut saat diperlukan Serealisasi adalah proses mengubah objek python menjadi format *byte-stream* agar dapat disimpan.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil dan Analisis Penelitian

4.1.1 Pra-Pemrosesan Data

	0
male	0
age	0
currentSmoker	0
cigsPerDay	29
BPMeds	53
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
Risk	0

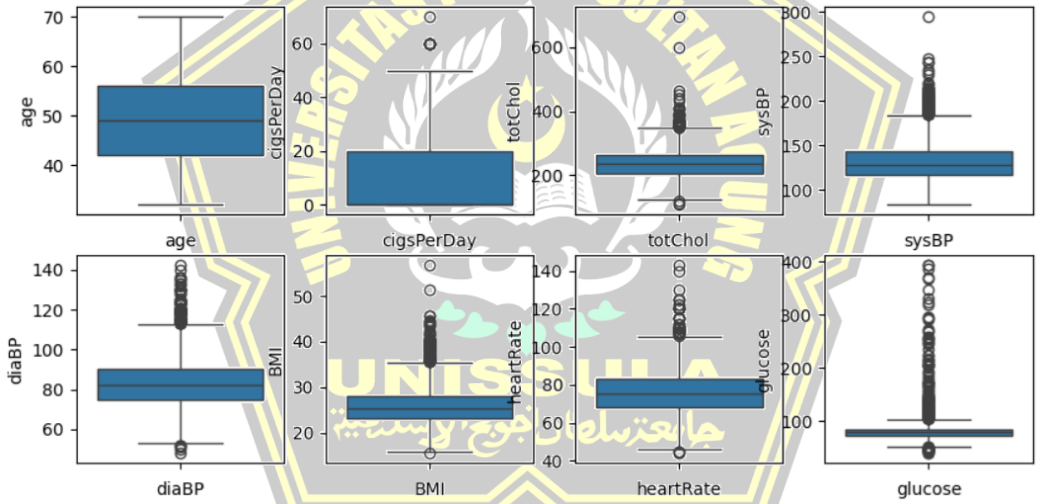
Gambar 4. 1 Mengecek *missing value*

Pada gambar 4.1 merupakan tahapan untuk mencari *missing value* atau nilai yang hilang pada tiap atribut. Dari gambar diatas terdapat atribut yang memiliki nilai null diantaranya atribut cigsPerDay terdapat 29 data, atribut BPMeds terdapat 53 data, atribut totChol terdapat 50 data, atribut BMI terdapat 19 data, atribut heartRate terdapat 1 data, dan terakhir atribut glucose terdapat 388 data.

```
df["cigsPerDay"]=df["cigsPerDay"].fillna(df["cigsPerDay"].mean())

df["totChol"]=df["totChol"].fillna(df["totChol"].mean())
df["BMI"]=df["BMI"].fillna(df["BMI"].mean())
df["heartRate"]=df["heartRate"].fillna(df["heartRate"].mean())
df["glucose"]=df["glucose"].fillna(df["glucose"].mean())
df['BPMeds'].fillna(df['BPMeds'].mode()[0], inplace=True)

df.isna().sum()
```

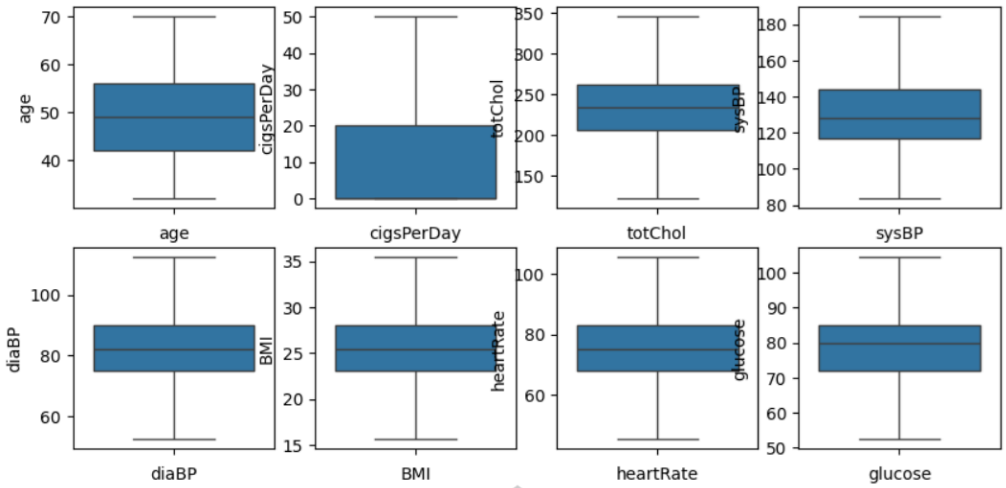
Gambar 4. 2 Imputasi *missing value* dengan mean dan modus

Pada gambar 4.2 dilakukan pra-pemrosesan data untuk menangani nilai yang hilang dalam sebuah dataset. Nilai yang hilang di kolom-kolom tertentu diisi dengan metode yang sesuai, yakni *mean* dan modus. Mean digunakan karena kolom tersebut bersifat numerik dan rata-rata bisa mewakili nilai tengah populasi. Sedangkan, nilai modus digunakan karena kolom tersebut bersifat kategorikal.



Gambar 4. 3 Identifikasi nilai *outlier*

Pada gambar 4.3 mempresentasikan distribusi dari nilai 8 variabel numerik dalam sebuah dataset. Dari gambar tersebut dilakukan proses identifikasi untuk mendeteksi nilai pencilan pada setiap atribut.Terdapat beberapa attribut yang memiliki nilai *outlier* seperti totChol, sysBP, diaBP, BMI, heartRate dan atribut glucose yang memiliki nilai ekstrem hampir 1000.



Gambar 4. 4 Penanganan pada nilai *outlier*

Pada gambar 4.4 menampilkan boxplot dari delapan variabel numerik setelah dilakukan penanganan nilai pencilan dengan mempertahankan ukuran dataset yang memiliki jumlah sampel terbatas atau tidak seimbang antar kelas dengan menggunakan metode IQR tujuannya adalah mengganti nilai *outlier* yang lebih kecil dari batas bawah dan mengganti nilai yang lebih besar dari batas atas dengan nilai batas atas.

4.1.2 Implementasi Metode Naïve Bayes

A. Membaca Data Latih

Tahap pertama dari implementasi metode naïve bayes adalah mencari pola untuk menemukan solusi. Adapun pola yang akan dicari menggunakan data latih sebanyak 3.392 data yang ditunjukkan pada gambar 4.5.

	male	age	currentSmoker	cigsPerDay	BPMeds	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Risk
0	0	53.0	1	20.0	0.0	0	221.0	131.0	89.0	24.09	90.0	95.0	0
1	0	64.0	1	6.0	1.0	0	239.0	143.0	84.0	20.06	55.0	73.0	1
2	0	38.0	0	0.0	0.0	0	185.0	100.0	72.0	22.15	85.0	83.0	0
3	0	49.0	0	0.0	1.0	1	270.0	126.5	67.5	26.56	70.0	77.0	0
4	1	56.0	1	20.0	1.0	1	186.0	116.0	67.0	24.62	70.0	83.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3387	0	36.0	1	5.0	1.0	1	222.0	147.0	94.0	26.79	76.0	71.0	1
3388	0	57.0	1	15.0	1.0	0	250.0	125.0	74.0	21.08	80.0	72.0	0
3389	0	60.0	0	0.0	0.0	1	298.0	133.0	89.0	25.09	83.0	81.0	1
3390	1	39.0	1	10.0	0.0	0	215.0	102.0	64.5	24.50	68.0	62.0	0
3391	0	35.0	0	0.0	1.0	0	248.0	107.0	73.0	20.64	90.0	80.0	0

3392 rows × 13 columns

Gambar 4. 5 Membaca data latih

B. Menghitung Kriteria dan Probabilitas Kelas

Tabel 4. 1 Menghitung probabilitas kelas

Kelas	Jumlah Kejadian	Probabilitas
Risiko Rendah	2.328	2.328/3392 : 0.69
Risiko Tinggi	1.064	1.064/3392 : 0.31
Jumlah	3392	

Pada tabel kriteria tiap kelas terdapat 3392 data pasien dengan status risiko rendah dan status risiko tinggi. Adapun untuk kelas risiko rendah memiliki probabilitas 2.328 pasien sedangkan untuk kelas risiko tinggi memiliki probabilitas 1.064 pasien. Selanjutnya tabel kriteria dan probabilitas kelas bisa dilihat pada tabel 4.1.

C. Menghitung Kriteria dan Probabilitas Gender

Tabel 4. 2 Menghitung probabilitas atribut jenis kelamin

gender	Jumlah kejadian		probabilitas	
	Risiko Rendah	Risiko Tinggi	Risiko Rendah	Risiko Tinggi
Perempuan	1.330	619	1330/2.328 : 0.57	619/1.064 : 0.58
Laki-Laki	998	445	998/2.328 : 0.43	445/1.064 : 0.42
Jumlah	2.328	1.064		

Pada tabel distribusi kriteria tiap kelas, terdapat total 3.392 data pasien yang dianalisis menggunakan atribut gender atau jenis kelamin. Atribut ini terbagi menjadi dua kategori, yakni “perempuan” menggunakan obat dan “laki-laki” menggunakan obat. Adapun untuk kategori perempuan, sebanyakk 1.330 pasien tergolong dalam kelompok dengan risiko rendah, sedangkan 619 pasien termasuk dalam kelompok dengan risiko tinggi. Sementara itu, pada kategori pasien laki-laki, terdapat 998 pasien dengan risiko rendah dan 445 pasien dengan risiko tinggi.

D. Menghitung Kriteria Probabilitas currentSmoker

Tabel 4. 3 Menghitung probabilitas atribut status perokok

current Smoker	Jumlah kejadian		probabilitas	
	Risiko Rendah	Risiko Tinggi	Risiko Redah	Risiko Tinggi
Tidak	1.064	611	1.064/2.328 : 0.46	611/1.064 : 0.57
Ya	1.264	453	1264/2.328 : 0.54	453/1.064 : 0.43
Jumlah	2.328	1.064		

Pada tabel distribusi kriteria tiap kelas, terdapat total 3.392 data pasien yang dianalisis menggunakan atribut currentSmoker atau perokok aktif. Atribut ini terbagi menjadi dua kategori, yakni "tidak" dan "ya". Adapun untuk kategori tidak, sebanyakk 1.064 pasien tergolong dalam kelompok dengan risiko rendah, sedangkan 661 pasien termasuk dalam kelompok dengan risiko tinggi. Sementara itu, pada kategori pasien yang perokok aktif terdapat 1.264 pasien dengan risiko rendah dan 453 pasien dengan risiko tinggi.

E. Menghitung Kriteria dan Probabilitas BPMeds

Tabel 4. 4 Menghitung probabilitas atribut penggunaan obat tekanan darah rendah

BPMeds	Jumlah kejadian		probabilitas	
	Risiko Rendah	Risiko Tinggi	Risiko Rendah	Risiko Tinggi
Tidak	1.214	551	1.214/2.328 : 0.52	511/1.064 : 0.52
Ya	1.114	513	1.114/2.328 : 0.48	513/1.064 : 0.48
Jumlah	2.328	1.064		

Pada tabel distribusi kriteria tiap kelas, terdapat total 3.392 data pasien yang dianalisis menggunakan atribut BPMeds atau penggunaan obat tekanan darah rendah. Atribut ini terbagi menjadi dua kategori, yakni "tidak" menggunakan obat dan "ya" menggunakan obat. Adapun untuk kategori tidak, sebanyakk 1.1214 pasien tergolong dalam kelompok

dengan risiko rendah, sedangkan 551 pasien termasuk dalam kelompok dengan risiko tinggi. Sementara itu, pada kategori pasien yang menggunakan obat tekanan darah rendah terdapat 1.114 pasien dengan risiko rendah dan 513 pasien dengan risiko tinggi.

F. Menghitung Kriteria dan Probabilitas diabetes

Tabel 4. 5 Menghitung probabilitas atirbut riwayat diabetes

diabetes	Jumlah kejadian		probabilitas	
	Risiko Rendah	Risiko Tinggi	Risiko Rendah	Risiko Tinggi
Tidak	1.136	513	1.136/2.328 : 0.48	513/1.064 : 0.48
Ya	1.192	551	1.192/2.328 : 0.52	551/1.064 : 0.52
Jumlah	2.328	1.064		

Pada tabel distribusi kriteria tiap kelas, terdapat total 3.392 data pasien yang dianalisis menggunakan atribut riwayat diabetes. Atribut ini terbagi menjadi dua kategori, yakni “tidak” dan “ya”. Adapun untuk kategori tidak, sebanyakk 1.136 pasien tergolong dalam kelompok dengan risiko rendah, sedangkan 513 pasien termasuk dalam kelompok dengan risiko tinggi. Sementara itu, pada kategori pasien yang diabetes terdapat 1.192 pasien dengan risiko rendah dan 551 pasien dengan risiko tinggi.

G. Menghitung Standar Deviasi dan Mean tiap atribut

Pada penelitian ini, untuk setiap atribut yang digunakan dalam model dilakukan perhitungan rata-rata dan standar deviasi agar dapat memperoleh gambaran umum mengenai sebaran data dan tingkat variasinya. Perhitungan rata-rata menggunakan persamaan (3) sedangkan standar devasi menggunakan persamaan (4). Hasil dapat dilihat pada tabel 4.6 dan 4.7

Tabel 4. 6 Menghitung rata-rata dari setiap atribut

Mean								
	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
Risiko Rendah	47.767182	9.674877	232.341059	121.914948	78.007517	24.959818	74.753651	79.270249
Risiko Tinggi	53.552632	7.994400	246.458735	154.108553	93.355733	27.432010	78.635225	81.268508

Tabel 4. 7 Menghitung standar deviasi tiap atribut

Standar Deviasi								
	age	cigsPerDay	totChol	sysBP	diaBP	BMI	heartRate	glucose
Risiko Rendah	8.103696	11.679775	40.992106	12.916439	8.353560	3.468136	11.096831	11.083053
Risiko Tinggi	8.157726	11.883966	42.542100	17.548742	10.215348	3.993102	12.119861	12.028455

H. Menghitung Nilai Uji

Tahapan menghitung nilai uji digunakan untuk menentukan apakah terdapat perbedaan yang signifikan antara kelompok data atau apakah data memenuhi asumsi tertentu dalam pengujian statistik. Perhitungan dilakukan menggunakan rumus dan metode yang telah ditentukan sebelumnya antara lain probabilitas , *mean* , dan *standar deviasi* antar atribut.

Tabel 4. 8 Sampel data uji

gender	**Age**	**currentSmoker**	**cigsPerDay**	**BPMeds**	**diabetes**
1	45	1	1	1	0
totChol	**sysBP**	**diaBP**	**BMI**	**heartRate**	**glucose**
227	140	84	28.74	69	74

1. Menghitung kelas prior

   P(risiko rendah) = 2.328/3.392 = 0.6863

   P(risiko tinggi) = 1.064/3.392 = 0.3137

2. Menghitung probabilitas setiap atribut

- Jenis kelamin

Berdasarkan data uji, variabel jenis kelamin dikategorikan sebagai data betipe kategorikal. Untuk menentukan nilai probabilitas dari kategori "laki-laki" digunakan rumus yang tercantum dalam persamaan (1). Diketahui bahwa terdapat 1.443 data pasien yang termasuk dalam kategori jenis kelamin "laki-laki", dari jumlah tersebut 998 pasien tergolong dalam kelas "risiko rendah" sedangkan 445 pasien berada dalam kelas "risiko tinggi". Oleh karena itu, perhitungan nilai peluang kategori "laki-laki" terhadap kelas "risiko rendah" dilakukan menggunakan pendekatan probabilitas sesuai dengan rumus sebagai berikut :

P(Gender)= laki-laki | "Risiko rendah" = 998/2.328 = 0.4287

Nilai peluang atribut jenis kelamin dengan kategori "laki-laki" terhadap kelas "risiko tinggi" melalui perhitungan sebagai berikut :

P(Gender)= laki-laki | "Risiko tinggi" = 445/1.064 = 0.4182

- Umur

Berdasarkan data uji pertama, diketahui bahwa usia 45 tahun termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel usia dapat dilihat pada tabel 4.9.

Tabel 4. 9 Parameter variabel umur tiap kelas

age	**kelas risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	47.767182	53.552632
Standar deviasi	8.103696	8.157726

$$P(age=45|\text{"Risiko rendah"}) = \frac{1}{\sqrt{2\pi}\,.\,8.103696}\; e - \frac{(45-47.767182)^2}{2.\,8.103696^2}$$

$$= 0.0464$$

$$P(age=45|\text{"Risiko tinggi"}) = \frac{1}{\sqrt{2\pi.\ 8.157726}}\ e - \frac{(45-53.552632)^2}{2.\ 8.157726^2}$$

$$= 0.0282$$

- Perokok aktif

Berdasarkan data uji, variabel perokok aktif dikategorikan sebagai data betipe kategorikal. Untuk menentukan nilai probabilitas dari kategori "ya" digunakan rumus yang tercantum dalam persamaan (1). Diketahui bahwa terdapat 1.717 data pasien yang termasuk dalam kategori perokok aktif "ya", dari jumlah tersebut 1.264 pasien tergolong dalam kelas "risiko rendah" sedangkan 453 pasien berada dalam kelas "risiko tinggi". Oleh karena itu, perhitungan nilai peluang kategori "ya" terhadap kelas "risiko rendah" dilakukan menggunakan pendekatan probabilitas sesuai dengan rumus sebagai berikut :

P(currentSmoker)= ya | "Risiko rendah" = 1.264/2.328 = 0.543

Nilai peluang atribut perokok aktif dengan kategori "ya" terhadap kelas "risiko tinggi" melalui perhitungan sebagai berikut :

P(currentSmoker)= ya | "Risiko tinggi" = 453/1.064 = 0.4258

- Batang rokok

Berdasarkan data uji pertama, diketahui bahwa batang rokok 1 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel batang rokok dapat dilihat pada tabel 4.10.

Tabel 4. 10 Parameter variabel batang rokok tiap kelas

cigsPerDay	**Status risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	9.674877	7.994400
Standar deviasi	11.679775	11.883966

P(cigsPerDay)= 1| “Risiko rendah” = $\frac{1}{\sqrt{2\pi.11.679775}} e - \frac{(1-9.674877)^2}{2.\ 11.679775^2}$

= 0.0259

P(cigsPerDay)= 1 | “Risiko tinggi” = $\frac{1}{\sqrt{2\pi.\ 11.883966}} e - \frac{(1-7.994400)^2}{2.\ 11.883966^2}$

= 0.0282

- Obat tekanan darah rendah

Berdasarkan data uji, variabel obat tekanan darah rendah dikategorikan sebagai data betipe kategorikal. Untuk menentukan nilai probabilitas dari kategori “ya” digunakan rumus yang tercantum dalam persamaan (1). Diketahui bahwa terdapat 1.627 data pasien yang termasuk dalam kategori obat tekanan darah rendah “ya”, dari jumlah tersebut 1.114 pasien tergolong dalam kelas “risiko rendah” sedangkan 513 pasien berada dalam kelas “risiko tinggi”. Oleh karena itu, perhitungan nilai peluang kategori “ya” terhadap kelas “risiko rendah” dilakukan menggunakan pendekatan probabilitas sesuai dengan rumus sebagai berikut :

P(BPMeds)= ya | “Risiko rendah” = 1.114/2.328 = 0.4785

Nilai peluang atribut penggunaan obat tekanan darah rendah dengan kategori “ya” terhadap kelas “risiko tinggi” melalui perhitungan sebagai berikut :

P(BPMeds)= ya | “Risiko tinggi” = 513/1.064 = 0.4821

- Diabetes

Berdasarkan data uji, variabel diabetes dikategorikan sebagai data betipe kategorikal. Untuk menentukan nilai probabilitas dari kategori “tidak” digunakan rumus yang tercantum dalam persamaan (1). Diketahui bahwa terdapat 1.649 data pasien yang termasuk dalam kategori diabetes “tidak”, dari jumlah tersebut 1.136 pasien tergolong dalam kelas “risiko rendah” sedangkan 513 pasien

berada dalam kelas “risiko tinggi”. Oleh karena itu, perhitungan nilai peluang kategori “ya” terhadap kelas “risiko rendah” dilakukan menggunakan pendekatan probabilitas sesuai dengan rumus sebagai berikut :

P(diabetes)= tidak | “Risiko rendah” = 1.136/2.328 = 0.488

Nilai peluang atribut diabetes dengan kategori “ya” terhadap kelas “risiko tinggi” melalui perhitungan sebagai berikut :

P(diabetes)= tidak | “Risiko tinggi” = 513/1.064 = 0.4821

- Total kolestrol

Berdasarkan data uji pertama, diketahui bahwa total kolestrol 227 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel kolesterol dapat dilihat pada tabel 4.11.

Tabel 4. 11 Parameter variabel total kolesterol tiap kelas

totChol	**Status risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	232.341059	246.458735
Standar deviasi	40.992106	42.542100

P(totChol)= 227 | “Risiko rendah” = $\frac{1}{\sqrt{2\pi.\ 40.992106}}\ e -$

$\frac{(227- 232.341059)^2}{2.\ 40.992106^2}$

= 0.0096

P(totChol)= 227 | “Risiko tinggi” = $\frac{1}{\sqrt{2\pi.\ 42.542100}}\ e -$

$\frac{(227-246.458735)^2}{2.\ 42.542100^2}$

= 0.0084

- Tekanan darah sistolik

Berdasarkan data uji pertama, diketahui bahwa tekanan darah sistolik 140 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel tekanan darah sistolik dapat dilihat pada tabel 4.12.

Tabel 4. 12 Parameter variabel tekanan darah sistolik tiap kelas

sysBP	**Status risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	121.914948	154.108553
Standar deviasi	12.916439	17.548742

P(sysBP)= 140 | "Risiko rendah" = $\frac{1}{\sqrt{2\pi}.\ 12.916439}\ e -$

$\frac{(140-121.914948)^2}{2.\ 12.916439^2}$

= 0.0116

P(sysBP)= 140 | "Risiko tinggi" = $\frac{1}{\sqrt{2\pi}.\ 17.548742}\ e -$

$\frac{(140-154.108553)^2}{2.\ 17.548742^2}$

= 0.0165

- Tekanan darah diastolik

Berdasarkan data uji pertama, diketahui bahwa tekanan darah diastolik 84 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel tekanan darah diastolik dapat dilihat pada tabel 4.13.

Tabel 4. 13 Parameter variabel tekanan darah diastolik tiap kelas

diaBP	Status risiko hipertensi	
	Risiko rendah	Risiko tinggi
Rata-rata	78.007517	93.355733
Standar deviasi	8.353560	10.215348

P(diaBP)= 84 | "Risiko rendah" = $\frac{1}{\sqrt{2\pi.\ 8.353560}}\ e - \frac{(84-78.007517)^2}{2.\ 8.353560^2}$

= 0.0369

P(diaBP)= 84 | "Risiko tinggi" = $\frac{1}{\sqrt{2\pi.\ 10.215348}}\ e - \frac{(84-93.355733)^2}{2.\ 10.215348^2}$

= 0.0257

- BMI

Berdasarkan data uji pertama, diketahui bahwa BMI 28.74 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel BMI dapat dilihat pada tabel 4.14.

Tabel 4. 14 Parameter variabel BMI tiap kelas

BMI	Status risiko hipertensi	
	Risiko rendah	Risiko tinggi
Rata-rata	24.959818	27.432010
Standar deviasi	3.468136	3.993102

P(BMI)= 28.74 | "Risiko rendah" = $\frac{1}{\sqrt{2\pi.\ 3.468136}}\ e - \frac{(28.74-24.959818)^2}{2.\ 3.468136^2}$

= 0.0635

P(BMI)= 28.74 | "Risiko tinggi" = $\frac{1}{\sqrt{2\pi.\ 3.993102}}\ e - \frac{(28.74-27.432010)^2}{2.\ 3.993102^2}$

= 0.0947

- Detak jantung

Berdasarkan data uji pertama, diketahui bahwa detak jantung 69 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel detak jantung dapat dilihat pada tabel 4.15.

Tabel 4. 15 Parameter variabel detak jantung tiap kelas

heartRate	**Status risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	74.753651	78.635225
Standar deviasi	11.096831	12.119861

P(heartRate)= 69 | "Risiko rendah" = $\frac{1}{\sqrt{2\pi.\ 11.096831}}\ e - \frac{(69-74.753651)^2}{2.\ 11.096831^2}$

= 0.0314

P(heartRate)= 69 | "Risiko tinggi" = $\frac{1}{\sqrt{2\pi.\ 12.119861}}\ e - \frac{(69-78.635225)^2}{2.\ 12.119861^2}$

= 0.0240

- Glukosa

Berdasarkan data uji pertama, diketahui bahwa kadar glukosa 74 termasuk dalam kategori data numerikal. Oleh karena itu, untuk menentukan nilai peluang dari data tersebut, digunakan rumus yang dijelaskan pada persamaan (3). Diketahui bahwa rata-rata dan standar deviasi variabel glukosa dapat dilihat pada tabel 4.16.

Tabel 4. 16 Parameter variabel glukosa tiap kelas

glucose	**Status risiko hipertensi**	
	Risiko rendah	**Risiko tinggi**
Rata-rata	79.270249	81.268508
Standar deviasi	11.083053	12.028455

P(glucose)= 74 | "Risiko rendah" = $\frac{1}{\sqrt{2\pi.\ 11.083053}}\ e - \frac{(74-79.270249)^2}{2.\ 11.083053^2}$

= 0.0321

P(glucose)= 74 | "Risiko tinggi" = $\frac{1}{\sqrt{2\pi.\ 12.028455}}\ e - \frac{(74-81.268508)^2}{2.\ 12.028455^2}$

= 0.0276

3. Menghitung prior

Setelah seluruh nilai probabilitas dari masing-masing variabel terhadap setiap kelas diperoleh, tahap berikutnya adalah menghitung nilai gabungan probabilitas. Proses ini dilakukan dengan cara mengalikan nilai peluang prior dengan hasil perkalian dari seluruh probabilitas dalam satu kelas. Perhitungan ini mengacu pada formula yang telah dijelaskan pada persamaan (5).

- Kelas risiko rendah

  = 0.6863 x 0.4287 x 0.0464 x 0.543 x 0.0259 x 0.4785 x 0.488 x 0.0096 x 0.0116 x 0.0369 x 0.0635 x 0.0314 x 0.0321

  = $1.1791005\ x\ 10^{-14}$

- Kelas risiko tinggi

  = 0.3137 x 0.4182 x 0.0282 x 0.4258 x 0.0282 x 0.4821 x 0.4821 x 0.0084 x 0.0165 x 0.0257 x 0.0947 x 0.0240 x 0.0276

  = $2.3069774\ x\ 10^{-15}$

Berdasarkan hasil perhitungan yang telah dilakukan, diketahui bahwa kelas dengan nilai probabilitas tertinggi adalah kelas dengan status 0, yaitu sebesar $1.1791005\ x\ 10^{-14}$. Oleh karena itu, data uji pertama diklasifikasikan ke dalam kelas tersebut yang menunjukkan bahwa pasien berada pada kategori rekam medis dengan status **risiko rendah**.

4. Menghitung confidence

Confidence merupakan ukuran tingkat keyakinan model terhadap hasil prediksi kelas tertentu berdasarkan probabilitas yang dihitung dengan membandingkan probabilitas hasil klasifikasi suatu kelas terhadap total probabilitas dari seluruh kelas menggunakan persamaan (6) :

$$Confidence\ (c) = \frac{P(c)}{\sum P(c)} * 100$$

$$Confidence\ (rendah) = \frac{1.1791005\ x\ 10^{-14}}{1.1791005\ x\ 10^{-14} + 2.3069774\ x\ 10^{-15}} * 100 = 83.74\%$$

$$Confidence\ (tinggi) = \frac{2.3069774\ x\ 10^{-15}}{1.1791005\ x\ 10^{-14} + 2.3069774\ x\ 10^{-15}} * 100 = 16.26\%$$

Berdasarkan hasil perhitungan diatas, dapat diambil kesimpulan bahwa model lebih yakin bahwa data uji termasuk ke dalam kelas risiko rendah karena memiliki tingkat confidence yang lebih tinggi yakni sebesar 83.74%.

4.1.3 Pengujian Skenario dataset

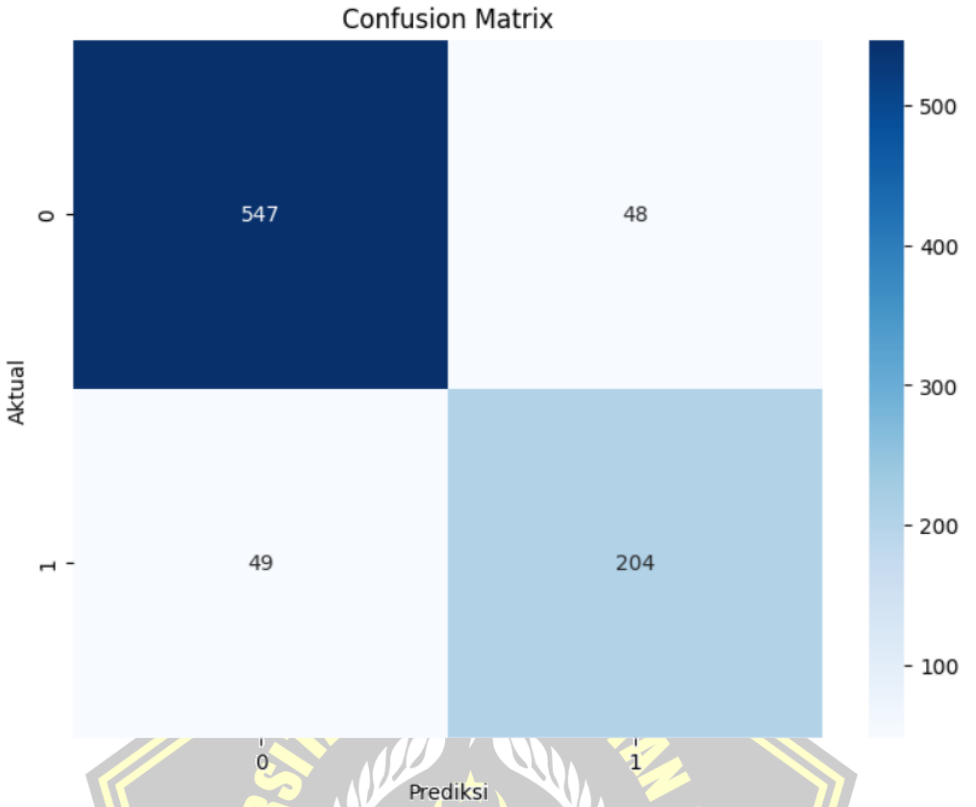
Tabel 4. 17 Evaluasi model pada skenario dataset yang berbeda

	Rasio	Kelas	Presisi	Recall	F1-Score	Akurasi
Skenario 1	60:40	Risiko rendah	0.92	0.91	0.92	0.88 (88%)
		Risiko tinggi	0.81	0.82	0.81	
Skenario 2	70:30	Risiko rendah	0.92	0.92	0.92	0.89 (89%)
		Risiko tinggi	0.81	0.81	0.81	
Skenario 3	80:20	Risiko rendah	0.92	0.92	0.92	0.89 (89%)
		Risiko tinggi	0.81	0.81	0.81	
Skenario 4	90:10	Risiko rendah	0.91	0.91	0.91	0.88 (88%)
		Risiko tinggi	0.79	0.79	0.79	

Berdasarkan tabel 4.17 hasil evaluasi terhadap empat skenario pembagian data latih dan data uji, diketahui bahwa model mencapai akurasi tertinggi pada skenario 2 dan skenario 3 dengan nilai sebesar 89%. Namun, secara keseluruhan skenario 3 dengan rasio 80:20 memberikan hasil yang seimbang dan stabil, baik dalam kelas mayoritas maupun minoritas. Oleh karena itu, skenario ini dapat dipertimbangkan sebagai konfigurasi terbaik untuk digunakan dalam implementasi model klasifikasi risiko penyakit hipertensi.

4.1.4 Evaluasi Model

Setelah konfigurasi terbaik dari masing-masing model berhasil ditentukan, proses evaluasi dilanjutkan dengan menghitung *confusion matrix* secara manual. Langkah ini dilakukan untuk memverifikasi kesesuaian hasil yang diberikan otomatis oleh sistem. Melalui perhitungan manual tersebut, berbagai metrik evaluasi seperti akurasi, presisi, recall, dan f1-Score dapat diterapkan secara lebih cermat. Keakuratan model dalam melakukan prediksi kemudian dinilai dengan cara membandingkan jumlah prediksi yang tepat baik untuk kelas positif maupun negatif terhadap keseluruhan data yang digunakan dalam pengujian.



Gambar 4. 6 Tabel confusion matrix dari split data 80:20

Gambar 4.16 menyajikan grafik *confusion matrix* yang diperoleh dari hasil klasifikasi terhadap risiko penyakit hipertensi. Evaluasi dilakukan dengan menggunakan pembagian data latih dan data uji sebesar 80:20. Untuk memperjelas analisis, hasil tersebut disajikan dalam bentuk tabel di bawah. Tabel ini menggunakan notasi standar dalam evaluasi klasifikasi meliputi TP (*True Positive*), FP (*False Positive*), FN (*False Negative*), dan TN (*True Negative*) guna mempermudah perhitungan dan interpretasi terhadap kinerja model.

Tabel 4. 18 TP, FP, FN, TN dari model tiap kelas

Kelas	TP	FP	FN	TN
0	547	49	48	204
1	204	48	49	547

*Keterangan tiap kelas:

a. 0 = risiko rendah

b. 1 = risiko tinggi

Percobaan perhitungan *confusion matrix* secara manual dilakukan untuk mengevaluasi kinerja model klasifikasi dalam mendeteteksi risiko penyakit hipertensi. Perhitungan ini bertujuan untuk memastikan keakuratan hasil evaluasi model yang diperoleh otomatis dari sistem. Hasil lengkap dari perhitungan manual disajikan pada tabel 4.19.

Tabel 4. 19 Perhitungan presisi, recall, dan f1-score model tiap kelas

	Perhitungan	Hasil
$Presisi\ (P) = \frac{TP}{TP + FP}$		
P_0	$= \frac{TP_0}{TP_0 + FP_0} = \frac{547}{547 + 49} = \frac{547}{596} = 0.917$	0.92
P_1	$= \frac{TP_1}{TP_1 + FP_1} = \frac{204}{204 + 48} = \frac{204}{252} = 0.809$	0.81
$Recall\ (R) = \frac{TP}{TP + FN}$		
R_0	$= \frac{TP_0}{TP_0 + FN_0} = \frac{547}{547 + 48} = \frac{547}{595} = 0.919$	0.92
R_1	$= \frac{TP_0}{TP_1 + FN_1} = \frac{204}{204 + 49} = \frac{204}{253} = 0.806$	0.81
$F1\ Score\ (F) = \frac{2\ \times Precision\ \times Recall}{Precision + Recall}$		
F_0	$= \frac{2\ \times P_0\ \times R_0}{P_0 + R_0} = \frac{2\ \times 0.92\ \times 0.92}{0.92 + 0.92} = \frac{1.6928}{1.84} = 0.92$	0.92
F_1	$= \frac{2\ \times P_1\ \times R_1}{P_1 + R_1} = \frac{2 \times 0.81 \times 0.81}{0.81 + 0.81} = \frac{1.3122}{1.62} = 0.81$	0.81

Pada tabel 4.19 menyajikan hasil evaluasi terhadap model klasifikasi risiko penyakit hipertensi dengan menggunakan metrik presisi, recall, dan f1-score. Evaluasi ini dilakukan pada dataset yang dibagi dengan rasio 80% data latih dan 20% data uji. Proses perhitungan dilakukan secara manual dengan mempertimbangkan jumlah kelas serta nilai presisi, recall, dan f1-score pada masing-masing kelas. Berdasarkan hasil tersebut, diperoleh rata-

rata nilai presisi sebesar 0.865, recall sebesar 0.865, dan f1-score sebesar 0.865 untuk kedua kelas yang dianalisis.

```
Data dengan prediksi yang benar:
     Actual  Predicted
0         0          0
1         0          0
3         1          1
4         0          0
5         1          1
..      ...        ...
843       0          0
844       0          0
845       1          1
846       0          0
847       0          0

[751 rows x 2 columns]
```

Gambar 4. 7 Menampilkan data prediksi model yang benar

```
Data dengan prediksi yang salah:
     Actual  Predicted
2         1          0
13        0          1
26        1          0
29        0          1
34        1          0
..      ...        ...
801       0          1
812       1          0
813       1          0
816       0          1
841       0          1

[97 rows x 2 columns]
```

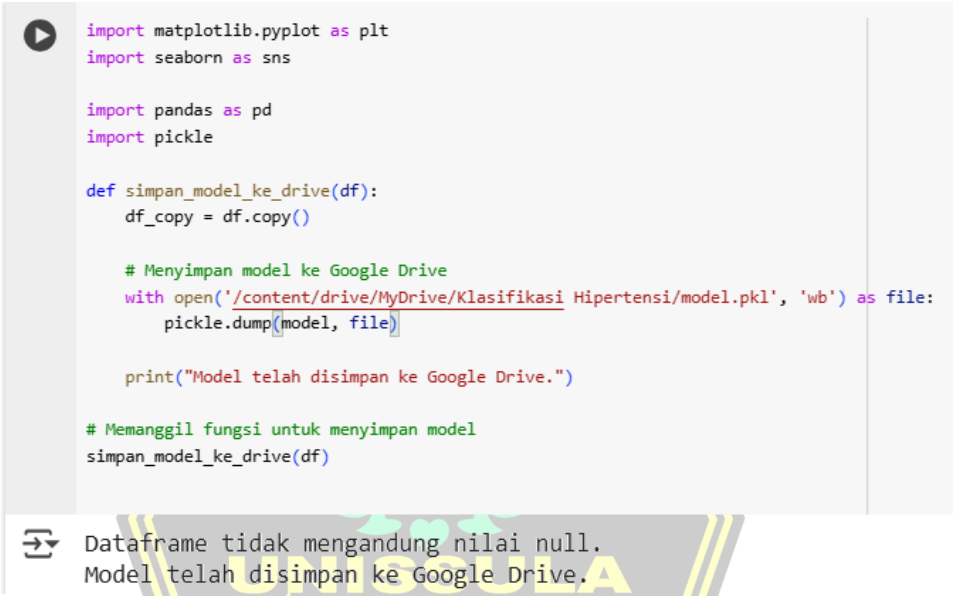
Gambar 4. 8 Menampilkan data prediksi model yang salah

Pada gambar 4.7 dan 4.8 menyajikan dua jenis sampel dari data uji. Sampel pertama menunjukkan hasil klasifikasi yang tepat oleh model, sedangkan sampel kedua memperlihatkan contoh ketika model melakukan kesalahan dalam mengklasifikasikan data. Gambar tersebut digunakan untuk menunjukkan perbedaan antara prediksi yang akurat dengan prediksi yang tidak tepat sebagai bagian dari evaluasi model.

4.2 Deploy Model ke Sistem

4.2.1 Menyimpan Model

Model yang telah selesai dilatih kemudian disimpan melalui proses serialisasi. Serialisasi model merupakan tahapan untuk mengubah model yang telah dilatih menjadi bentuk yang dapat disimpan dalam media penyimpanan. Tujuan dari proses ini adalah agar model tidak perlu dilatih ulang saat ingin digunakan kembali di masa mendatang. Proses ini memudahkan pemanggilan model saat proses prediksi. Berikut merupakan tahapan dan kode program yang digunakan dalam proses serialisasi tersebut:

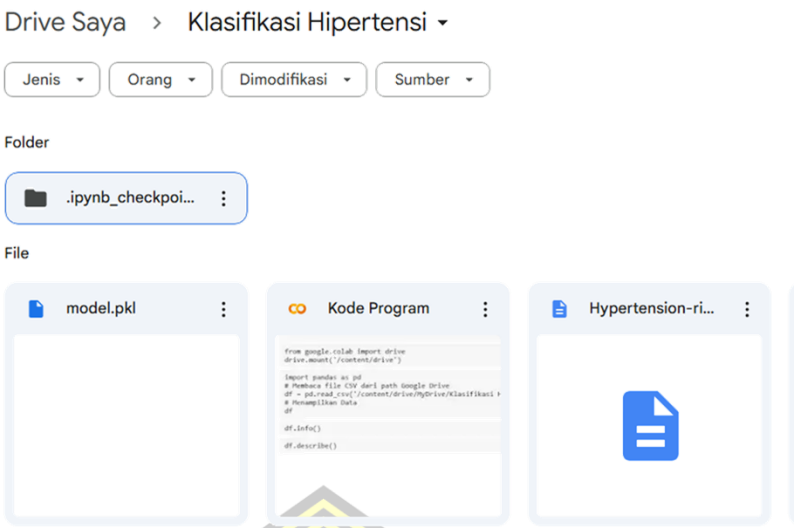


```
import matplotlib.pyplot as plt
import seaborn as sns

import pandas as pd
import pickle

def simpan_model_ke_drive(df):
    df_copy = df.copy()

    # Menyimpan model ke Google Drive
    with open('/content/drive/MyDrive/Klasifikasi Hipertensi/model.pkl', 'wb') as file:
        pickle.dump(model, file)

    print("Model telah disimpan ke Google Drive.")

# Memanggil fungsi untuk menyimpan model
simpan_model_ke_drive(df)
```

```
Dataframe tidak mengandung nilai null.
Model telah disimpan ke Google Drive.
```

Gambar 4. 9 Kode program menyimpan model

Gambar 4.9 memperlihatkan tahapan penyimpanan model yang telah melalui proses pelatihan dan pengujian. Proses penyimpanan ini dilakukan melalui metode serialisasi dengan memanfaatkan *library* pickle pada python. Model yang telah diserialisasi disimpan dalam sebuah berkas dengan nama “model.pkl”.

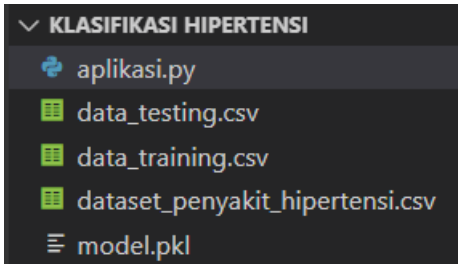


Gambar 4. 10 Menyimpan model ke google drive

Gambar 4.10 memperlihatkan file model yang telah diserialisasi dan disimpan ke dalam folder google drive. File hasil serialisasi ini akan dimanfaatkan sebagai komponen utama dalam membangun sistem klasifikasi untuk mendeteksi risiko penyakit hipertensi.

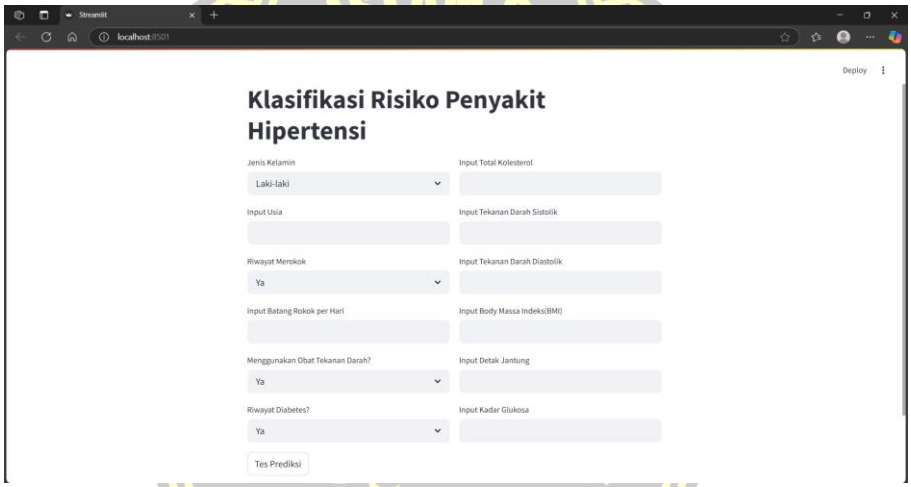
4.2.2 Pembuatan sistem

Pembuatan sistem dilakukan dengan memanfaatkan *framework* streamlit yang dirancang khusus untuk membangun antarmuka aplikasi berbasis web secara sederhana dan cepat. Dalam proses pembuatan sistem, penulis menggunakan *Visual Studio Code* sebagai editor teks utama dengan dukungan bahasa pemrograman python versi 3.13.3. Setelah itu, dilakukan instalasi *framework* Streamlit beserta beberapa library pendukung python yang dibutuhkan selama proses pembuatan model. Langkah selanjutnya adalah melakukan import terhadap file dataset, file pelatihan model, serta file model yang telah disimpan sebelumnya ke dalam folder proyek untuk keperluan pengolahan data dan pengujian model.



Gambar 4. 11 Pembuatan sistem

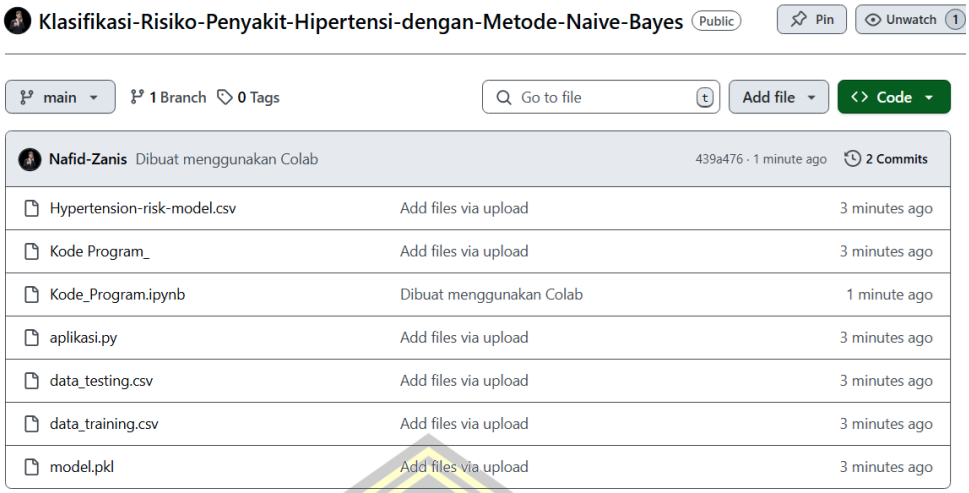
Pada gambar 4.11 menampilkan sebuah file bernama aplikasi.py yang digunakan sebagai media penulisan kode program untuk membangun aplikasi menggunakan *framework* streamlit. File ini memuat seluruh fungsi dan komponen antarmuka yang membentuk sistem secara keseluruhan. Adapaun tampilan dari sistem yang sudah dirancang ditunjukkan pada gambar 4.12



Gambar 4. 12 Tampilan antarmuka pada website streamlit

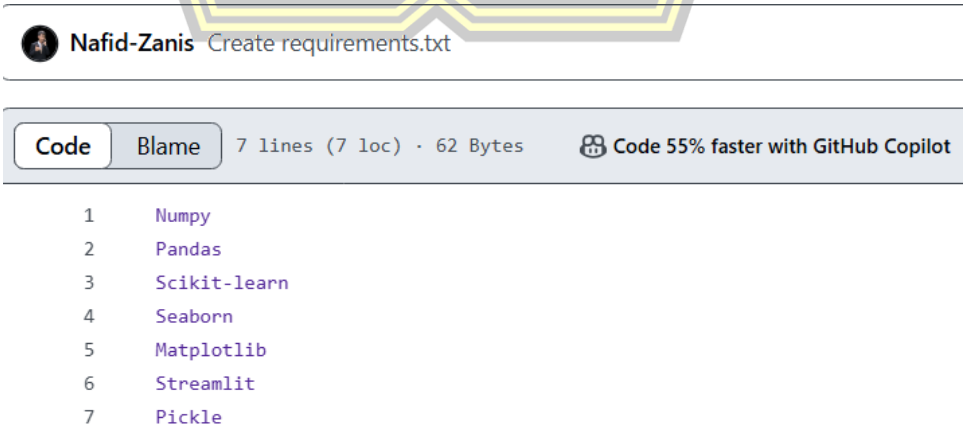
Pada sistem yang dikembangkan, tersedia beberapa form *input* yang harus diisi oleh pengguna sebelum menjalankan proses klasifikasi. Input tersebut mencakup data pribadi dan informasi kesehatan seperti jenis kelamin, usia, status perokok aktif, jumlah batang rokok yang dikonsumsi per hari, penggunaan obat tekanan darah rendah, riwayat diabetes, kadar kolesterol total, tekanan darah sistolik, tekanan darah diastolik, *body massa indeks*, detak jantung, dan kadar glukosa. Setelah seluruh data terisi, pengguna dapat menekan tombol “Tes Prediksi” yang tersedia di bagian bawah untuk melihat hasil klasifikasi yang dihasilkan oleh sistem.

4.2.3 Proses deploy model

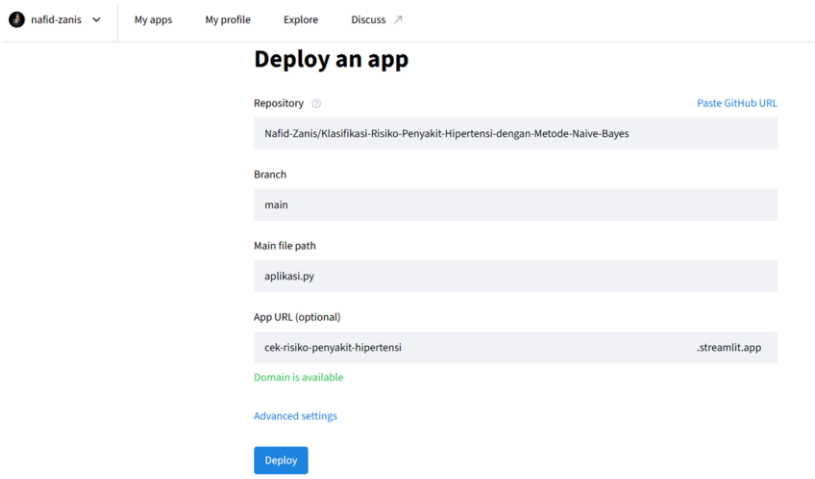


Gambar 4. 13 Repository proyek pada github

Gambar 4.13 memperlihatkan tampilan dari repository tempat proyek ini diunggah. Setelah proyek berhasil dikirim ke platform GitHub, langkah berikutnya adalah membuat file requirements.txt. File ini berfungsi untuk mencantumkan seluruh pustaka python yang digunakan dalam pengembangan sistem. Keberadaan file ini sangat penting karena memberi tahu platform streamlit cloud pustaka apa saja yang perlu diinstal secara otomatis saat proses *deployment*. Pada file tersebut, versi pustaka dapat dicantumkan jika diperlukan. Namun, karena proyek ini dikembangkan menggunakan versi pustaka pyhton terbaru, maka penulisan versi pustaka tidak diwajibkan. Tampilan file requirements.txt dapat dilihat pada gambar

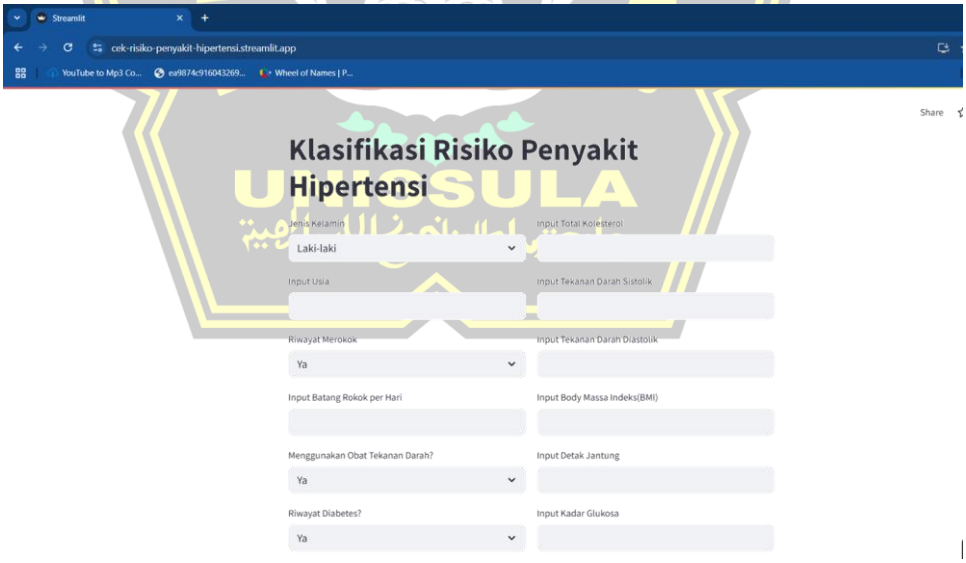


Gambar 4. 14 Pustaka yang dibutuhkan



Gambar 4. 15 Proses deploy model ke streamlit cloud

Gambar 4.15 memperlihatkan tahapan proses *deployment* ke sistem platform streamlit cloud. Pada tahap ini, pengguna perlu mengisi informasi terkait proyek yang sebelumnya dikirim ke repositori github. Selain itu, pengguna juga diberikan opsi untuk mengatur alamat domain sistem sesuai kebutuhan. Dalam penelitian ini, domain yang dipilih dan diguanakan adalah "cek-risiko-penyakit-hipertensi.app"



Gambar 4. 16 Tampilan antarmuka website setelah dilakukan proses deploy

Tampilan pada gambar 4.16 menunjukkan bahwa sistem klasifikasi risiko penyakit hipertensi telah berhasil di-*deploy* menggunakan platform *streamlit cloud*. Aplikasi ini dapat diakses melalui tautan berikut : https://cek-risiko-penyakit-hipertensi.streamlit.app/. Pada antarmuka

website, pengguna diminta untuk mengisi sejumlah informasi terkait riwayat kesehatan mereka. Selanjutnya, pengguna dapat menjalankan proses klasifikasi dengan menekan tombol ‘tes prediksi’. Setelah proses tersebut, sistem akan menampilkan hasil klasifikasi berupa status risiko apakah tergolong dalam kategori rendah atau tinggi.

4.3 Pengujian Sistem

Metode pengujian sistem yang diterapkan dalam penelitian ini adalah *blackbox testing.* Metode ini merupakan pendekatan pengujian yang berfokus pada pemeriksaan fungsionalitas sistem berdasarkan *input* dan *output*, tanpa melihat struktur internal atau sumber kode sistem. Pada pengujian ini digunakan *fuctional testing* yaitu salah satu jenis pengujian dalam metode *blackbox testing*. Fungsional testing bertujuan untuk memastikan bahwa seluruh fitur dan fungsi dalam sistem berjalan sesuai dengan spesifikasi fungsional yang telah dirancang.

Tabel 4. 20 Hasil pengujian menggunakan *blackbox testing*

Komponen yang diuji	Deskripsi Pengujian	Hasil yang diharapkan	Evaluasi
Tautan aplikasi	Memverifikasi apakah tautan aplikasi dapat diakses dengan baik	Berfungsi dengan benar	Normal
Formulir input data	Mengisi form dengan atribut yang ditentukan	Sistem menerima data	Normal
Fungsi tombol submit	Menekan tombol ‘tes prediksi’	Sistem merespons input	Normal
Proses klasifikasi risiko	Meninjau hasil klasifikasi risiko hipertensi setelah data dikirim	Hasil tampilan sesuai	Normal

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan , dapat disimpulkan bahwa metode Naïve Bayes dapat diterapkan secara efektif dalam mengelompokkan risiko penyakit hipertensi. Model ini terbukti mampu memberikan akurasi yang baik dalam melakukan prediksi. Keakuratan tersebut telah diverifikasi melalui serangkain pengujian dengan skenario yang berbeda. Adapun beberapa kesimpulan yang diperoleh dari pelaksanaan penelitian ini disampaikan dalam poin sebagai berikut :

1. Metode Naïve Bayes dapat diimplementasikan dengan baik pada website untuk klasifikasi risiko penyakit hipertensi yaitu risiko rendah dan risiko tinggi hal ini ditunjukkan dengan akurasi klasifikasi sebesar 89%.
2. Skenario pembagian dataset 70:30 dan 80:20 menghasilkan akurasi tertinggi sebesar 89%.
3. Presisi dan F1-Score tertinggi sebesar 0.92 untuk kelas risiko rendah diperoleh pada skenario 60:40, 70:30 dan 80:20.
4. Recall tertinggi untuk kelas risiko tinggi sebesar 0.82 tercapai pada skenario 60:40 menandakan model lebih baik mengenali data minoritas pada skenario tersebut.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan, penulis menyarankan untuk penelitian yang akan datang adalah :

1. Menerapkan teknik penyeimbangan data karena terdapat ketidakseimbangan jumlah data antar kelas untuk menghindari bias pada model.
2. Melakukan perbadingan dengan algoritma yang lain untuk mendapatkan hasil klasifikasi yang lebih optimal.
3. Karena terbatasnya dataset pada dua kelas, penelitian selanjutnya dapat memodifikasi dataset dengan menambahkan jumlah kelas yang berbeda.

DAFTAR PUSTAKA

Akmal, K. *dkk.* (2023) "PERBANDINGAN METODE ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENYAKIT STROKE," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), hal. 470–477.

Andriani, W. *dkk.* (2023) "Prediksi Nilai Emas Menggunakan Algoritma Regresi Linear," *Jurnal Ilmiah Informatika Komputer*, 28(1), hal. 27–35. Tersedia pada: https://doi.org/10.35760/ik.2023.v28i1.8096.

Awalullaili, F.O. *dkk.* (2022) "KLASIFIKASI PENYAKIT HIPERTENSI MENGGUNAKAN METODE SVM GRID SEARCHDAN SVM GENETIC ALGORITHM(GA)," *JURNAL GAUSSIAN*, 11(4), hal. 488–498. Tersedia pada: https://doi.org/10.14710/j.gauss.11.4.488-498.

Damayanti, A. *dkk.* (2024) "Prediksi Angka Kemiskinan Desa Kemang Bejalu Menggunakan MetodeNaive Bayes," *Rekayasa Teknik Informatika dan Informasi*, 4(3), hal. 196–201. Tersedia pada: https://djournals.com/resolusi.

Fuad, M. *dkk.* (2023) "Implementasi Klasifikasi Naive Bayes Dalam Memprediksi Lama Studi Mahasiswa," *Prosiding SISFOTEK*, 7(1), hal. 209–312. Tersedia pada: https://www.kaggle.com/datasets.

Haffandi, M.Y. *dkk.* (2022) "Klasifikasi Penyakit Paru-Paru Dengan Menggunakan Metode Naïve Bayes Classifier," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, 5(2), hal. 176. Tersedia pada: https://doi.org/10.37600/tekinkom.v5i2.649.

Hastomo, W. *dkk.* (2022) "Metode Pembelajaran Mesin untuk Memprediksi Emisi Manure Management," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 11(2), hal. 131–139. Tersedia pada: https://doi.org/10.22146/jnteti.v11i2.2586.

Kartika, Y. *dkk.* (2022) "Implementasi Algoritma Naïve Bayes Untuk Prediksi Persediaan Barang Rotan," *KOPERTIP : Jurnal Ilmiah Manajemen Informatika dan Komputer*, 4(1), hal. 28–34. Tersedia pada:

https://doi.org/10.32485/kopertip.v4i1.112.

Kharits, A.K. *dkk.* (2023) "PERBANDINGAN PREDIKSI PENYAKIT HIPERTENSI MENGGUNAKAN METODE RANDOM FOREST DAN NAÏVE BAYES," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1), hal. 498–504.

Prasetya, W.D. dan Sujatmiko, B. (2022) "Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes," *Journal of Informatics and Computer Science (JINACS)*, 3(04), hal. 515–525. Tersedia pada: https://doi.org/10.26740/jinacs.v3n04.p515-525.

Pridiptama, R.P. *dkk.* (2024) "Perbandingan Algoritma Support Vector Machine dan Nave Bayes pada Klasi kasi Penyakit Tekanan Darah Tinggi (Studi Kasus: Klinik Polresta Samarinda)," *Basis*, 3(1), hal. 1–16. Tersedia pada: http://jurnal.fmipa.unmul.ac.id/index.php/Basis.

Purwono, P. *dkk.* (2022) "Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network," *INSECT (Informatics and Security) : Jurnal Teknik Informatika*, 7(2), hal. 82–90.

Putra, I.M.A.A.D. *dkk.* (2024) "Perbandingan Algoritma Naive Bayes Berbasis Feature Selection Gain Ratio dengan Naive Bayes Kovensional dalam Prediksi Komplikasi Hipertensi," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, 6(1), hal. 37–49. Tersedia pada: https://doi.org/10.35746/jtim.v6i1.488.

Riany, A.F. dan Testiana, G. (2023) "PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT JANTUNG KORONER MENGGUNAKAN ALGORITMA NAIVE BAYES," *PROCEEDING Multi Data Palembang Student Conference*, 2(1), hal. 297–305. Tersedia pada: https://doi.org/10.35957/mdp-sc.v2i1.4388.

Rinanda, P.D. *dkk.* (2022) "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 2(2), hal.

68–75. Tersedia pada: https://doi.org/10.57152/malcom.v2i2.432.

Setiandari L.O, E. (2022) "Hubungan Pengetahuan, Pekerjaan dan Genetik (riwayat hipertensi dalam keluarga) Terhadap Perilaku Pencegahan Penyakit Hipertensi," *Media Publikasi Promosi Kesehatan Indonesia*, 5(4), hal. 457–462.

Siska, S. *dkk.* (2023) "Implementasi Metode Naive Bayes pada Prediksi Penyakit Seliak," *KOPERTIP Jurnal Ilmiah Manajemen Informatika dan Komputer*, 7(1), hal. 8–13. Tersedia pada: https://doi.org/10.32485/kopertip.v7i1.325.

Surorejo, S. *dkk.* (2022) "Penerapan Metode Naïve Bayes Pada Sistem Pakar Untuk Diagnosa Penyakit Hipertensi," *Indonesian Journal of Informatic and Research*, 3(1), hal. 8–17.

Syafiih, M. (2023) "Klasifikasi Kategori Berdasarkan Tingkat Ketergantungan Siswa Terhadap Penggunaan Smartphone Di SMK Negeri 1 Suboh Situbondi," *JEECOM Journal of Electrical Engineering and Computer*, 5(2), hal. 329–338. Tersedia pada: https://doi.org/10.33650/jeecom.v5i2.6833.

Tarimana, A.A. *dkk.* (2024) "PREDIKSI PENYAKIT HIPERTENSI MENGGUNAKAN MACHINE LEARNING DENGAN ALGORITMA REGRESI LOGISTIK," *JATI(Jurnal Mahasiswa Teknik Informatika)*, 8(6), hal. 12062–12068.

Uswatun Khasanah, L. *dkk.* (2022) "Klasifikasi Penyakit Diabetes Melitus Menggunakan Naive Bayes Classifier," *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, 7(1), hal. 59–66. Tersedia pada: https://doi.org/10.32528/justindo.v7i1.4949.

Wie, J.V. dan Siddik, M. (2022) "PENERAPAN METODE NAÏVE BAYES DALAM MENGKLASIFIKASI TINGKAT OBESITAS PADA PRIA," *JOISIE Journal Of Information System And Informatics Engineering*, 6(2), hal. 69–77.

Yoliadi, D.N. (2023) "Data Mining Dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-Means," *Insearch: Information System Research Journal*, 3(01). Tersedia pada:

https://doi.org/10.15548/isrj.v3i01.5829.

Yusup, R.M. dan Rijanto, E. (2024) “Analisis Komparatif Model Pembelajaran Mesin Untuk Memprediksi Hipertensi Ke Dalam Empat Kelas Berdasarkan JNC 8 Comparative Analysis of Machine Learning Models for Predicting Hypertension into Four Classes,” *Jurnal Tata Kelola dan Kerangka Kerja Teknologi Informasi*, 10, hal. 92–102.

Zai, C. (2022) “Implementasi Data Mining Sebagai Pengolahan Data,” *Jurnal Portal Data*, 2(3), hal. 1–12. Tersedia pada: http://portaldata.org/index.php/portaldata/article/view/107.