ANALISIS PENGARUH SELEKSI FITUR ANOVA TERHADAP PERFORMA MODEL KLASIFIKASI GAUSSIAN NAÏVE BAYES PADA DATASET PIMA INDIANS DIABETES

SKRIPSI

Oleh:
FAHRIZA KURNIAWAN
NIM. 18650080

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

ANALISIS PENGARUH SELEKSI FITUR ANOVA TERHADAP PERFORMA MODEL KLASIFIKASI GAUSSIAN NAÏVE BAYES PADA DATASET PIMA INDIANS DIABETES

SKRIPSI

Oleh :

FAHRIZA KURNIAWAN
NIM. 18650080

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

HALAMAN PERSETUJUAN

ANALISIS PENGARUH SELEKSI FITUR ANOVA TERHADAP PERFORMA MODEL KLASIFIKASI GAUSSIAN NAÏVE BAYES PADA DATASET PIMA INDIANS DIABETES

SKRIPSI

Oleh :
FAHRIZA KURNIAWAN
NIM. 18650080

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 1 Desember 2023

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Dr. M. Imamudin Lc, MA
NIP. 19680519 200312 1 001	NIP. 19740602 200901 1 010

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

ANALISIS PENGARUH SELEKSI FITUR ANOVA TERHADAP PERFORMA MODEL KLASIFIKASI GAUSSIAN NAÏVE BAYES PADA DATASET PIMA INDIANS DIABETES

SKRIPSI

Oleh :
FAHRIZA KURNIAWAN
NIM. 18650080

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 7 Desember 2023

Susunan Dewan Penguji

Ketua Penguji	: Dr. Totok Chamidy, M.Kom NIP. 19691222 200604 1 001	( )
Anggota Penguji I	: Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. M. Imamudin Lc, MA NIP. 19740602 200901 1 010	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Fahriza Kurniawan
NIM	: 18650080
Fakultas / Jurusan	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Analisis Pengaruh Seleksi Fitur ANOVA Terhadap Performa Model Klasifikasi Gaussian Naïve Bayes Pada Dataset Pima Indians Diabetes.

.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 14 Desember 2023
Yang membuat pernyataan,

Fahriza Kurniawan
NIM.18650080

HALAMAN MOTTO

... NEVER STOP TRYING ...

HALAMAN PERSEMBAHAN

Dalam penulisan skripsi ini, penulis ingin mengucapkan rasa terima kasih dan penghargaan kepada individu dan kelompok yang telah memberikan dukungan dan kontribusi selama proses penelitian dan penulisan skripsi ini. Tanpa bantuan mereka, penyelesaian skripsi ini tidak mungkin terwujud.

Pertama-tama, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada keluarga tercinta. Terima kasih kepada orangtua tercinta, Suharno dan Pancaningtyas, yang selalu memberikan dukungan moral, cinta, dan semangat selama proses penulisan skripsi ini. Terima kasih atas doa, dorongan, dan pengertian yang telah diberikan oleh kalian berdua. Tanpa kehadiran kalian, penulis tidak akan mampu menyelesaikan skripsi ini dengan baik.

Penulis juga ingin mengucapkan terima kasih kepada saudara-saudari dan kerabat-kerabat dekat. Terima kasih atas dukungan, semangat, dan motivasi yang kalian berikan. Persaudaraan kita telah memberikan kehangatan dan keceriaan dalam setiap langkah penulisan skripsi ini. Terima kasih atas kebersamaan kita dan semua momen indah yang telah dilalui bersama.

Tidak lupa, penulis juga ingin mengucapkan terima kasih kepada Bapak Prof. Dr. Suhartono, M.Kom dan Bapak Dr. M. Imamudin, Lc. MA. selaku pembimbing skripsi. Terima kasih atas bimbingan, arahan, serta masukan yang berharga dari keduanya dalam menyelesaikan skripsi ini. Bantuan dan kesabaran yang diberikan sangatlah berarti bagi penulis dalam mengatasi berbagai kendala yang muncul selama penelitian ini berlangsung.

Dalam penulisan skripsi ini, penulis ingin mengucapkan rasa terima kasih dan penghargaan kepada teman-teman seperjuangan yang telah memberikan dukungan dan kontribusi selama proses penelitian dan penulisan skripsi ini. Tanpa bantuan mereka, penyelesaian skripsi ini tidak mungkin terwujud.

Akhir kata, penulis ingin menyampaikan rasa terima kasih kepada semua pihak yang tidak dapat disebutkan satu per satu namun telah memberikan dukungan dan motivasi dalam berbagai bentuk selama penulisan skripsi ini. Semoga segala bantuan dan doa yang telah diberikan oleh keluarga tercinta dapat diberikan balasan yang setimpal oleh Allah SWT.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat, hidayah, serta karunia-Nya yang telah melimpah dalam perjalanan penulisan skripsi ini. Skripsi dengan judul "Analisis Pengaruh Seleksi Fitur ANOVA Terhadap Performa Model Klasifikasi Gaussian Naïve Bayes Pada Dataset Pima Indians Diabetes" merupakan bagian dari upaya penulis dalam menyelesaikan pendidikan program studi Teknik Informatika di Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Tanpa dukungan doa, izin, dan dorongan dari berbagai pihak serta tekad penulis sendiri, Skripsi ini tidak akan dapat terealisasi. Oleh karena itu, penulis ingin menyampaikan apresiasi yang sangat besar kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan M.MT., IPM selaku Ketua Prodi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhartono, M.Kom dan Dr. M. Imamudin, Lc. MA. selaku pembimbing skripsi. Terima kasih atas bimbingan, arahan, serta masukan yang berharga dari keduanya dalam menyelesaikan skripsi ini.
5. Okta Qomaruddin Aziz, M.Kom dan Dr. Totok Chamidy, M.Kom. selaku penguji skripsi. Terima kasih telah meluangkan waktu dan tenaga untuk

membaca, menelaah, serta memberikan masukan dan evaluasi terhadap skripsi ini.

6. Keluarga tercinta, terutama orangtua dan saudara-saudari, atas dukungan, cinta, dan doa yang tak henti-hentinya diberikan. Ridho, semangat, serta dorongan yang berasal dari keluarga adalah sumber inspirasi utama penulis dalam menyelesaikan skripsi ini.

Terakhir, penulis ingin menyampaikan rasa terima kasih kepada semua pihak yang tidak dapat disebutkan satu per satu namun telah memberikan dukungan dan motivasi dalam berbagai bentuk selama penulisan skripsi ini.

Semoga skripsi ini dapat memberikan manfaat dan kontribusi yang positif bagi perkembangan ilmu pengetahuan di bidang yang relevan. Penulis menyadari bahwa skripsi ini tidak luput dari kekurangan dan keterbatasan. Oleh karena itu, penulis sangat mengharapkan kritik, saran, dan masukan yang membangun untuk pengembangan penelitian ini di masa yang akan datang.

Akhir kata, semoga Allah SWT senantiasa memberikan rahmat, hidayah, dan keberkahan-Nya kepada kita semua. Amin.

Malang, 5 Desember 2023

Fahriza Kurniawan

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Kurniawan, Fahriza. 2023. **Analisis Pengaruh Seleksi Fitur ANOVA Terhadap Performa Model Klasifikasi Gaussian Naïve Bayes Pada Dataset Pima Indians Diabetes.** Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. M. Imamudin Lc, MA.

.

Kata kunci: Gaussian Naïve Bayes, Diabetes Mellitus, Seleksi Fitur, ANOVA.

Penelitian ini bertujuan untuk menganalisis pengaruh seleksi fitur ANOVA terhadap performa model klasifikasi Gaussian Naïve Bayes pada dataset Pima Indians Diabetes. Diabetes merupakan masalah kesehatan global yang mempengaruhi jutaan orang di seluruh dunia. Dalam upaya untuk mengidentifikasi faktor risiko yang berkontribusi terhadap diabetes mellitus, analisis data dan pemodelan klasifikasi dapat membantu dalam memprediksi kemungkinan seseorang mengalami kondisi ini. Hasil penelitian menunjukkan bahwa seleksi fitur ANOVA memiliki pengaruh signifikan terhadap performa model klasifikasi Gaussian Naïve Bayes pada dataset Pima Indians Diabetes. Dengan memilih fitur-fitur yang paling informatif, model klasifikasi dapat meningkatkan kemampuan prediksi diabetes dengan meningkatkan akurasi dan metrik evaluasi lainnya. Dengan akurasi sebesar 86,0%, model yang hanya menggunakan lima fitur terbaik ini mampu mengklasifikasikan data dengan tingkat akurasi yang lebih tinggi dibandingkan model lainnya.

ABSTRACT

Kurniawan, Fahriza. 2023. ***Analysis of the Influence of ANOVA Feature Selection on the Performance of Gaussian Naïve Bayes Classification Model on the Pima Indians Diabetes Dataset.*** Theses. Department of Informatics Engineering Faculty of Science and Technology Maulana Malik Ibrahim State Islamic University of Malang. Advisor : (I) Prof. Dr. Suhartono, M.Kom (II) Dr. M. Imamudin Lc., MA.

The purpose of this research is to analyze the influence of ANOVA feature selection on the performance of the Gaussian Naïve Bayes classification model on the Pima Indians Diabetes dataset. Diabetes is a global health problem that affects millions of people worldwide. In an effort to identify risk factors that contribute to diabetes mellitus, data analysis and classification modeling can help predict the likelihood of someone experiencing this condition. The research results indicate that ANOVA feature selection has a significant influence on the performance of the Gaussian Naïve Bayes classification model on the Pima Indians Diabetes dataset. By selecting the most informative features, the classification model can improve the ability to predict diabetes by increasing accuracy and other evaluation metrics. With an accuracy of 86.0%, the model that uses only the top five features is able to classify data with a higher accuracy rate compared to other models.

Keywords : Gaussian Naïve Bayes, Diabetes Mellitus, Feature Selection, ANOVA.

الملخص

كورنياوان، فخريزا. ٢٠٢٣. تحليل تأثير اختيار ميزة ANOVA على أداء نموذج تصنيف البايزي الساذج (Naïve Bayes Gaussian) في مجموعة بيانات Pima Indians Diabetes. البحث الجامعي. قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية، مالانج. المشرف الأول: أ. د. سوهارتونو، الماجستير. المشرف الثاني: د. محمد إمام الدين، الماجستير.

الكلمات الرئيسية: بايزي ساذج، داء السكري، اختيار الميزة، ANOVA.

يهدف هذا البحث إلى تحليل تأثير اختيار ميزة ANOVA على أداء نموذج تصنيف البايزي الساذج (Gaussian Naïve Bayes) في مجموعة بيانات Pima Indians Diabetes. مرض السكري هو مشكلة صحية عالمية تؤثر على الملايين من الناس في جميع أنحاء العالم. في محاولة لتحديد عوامل الخطر التي تساهم في مرض السكري، يمكن أن يساعد تحليل البيانات ونمذجة التصنيف في التنبؤ باحتمالية إصابة الشخص بهذه الحالة. أظهرت النتائج أن اختيار ميزة ANOVA كان له تأثير كبير على أداء نموذج تصنيف البايزي الساذج في مجموعة بيانات Pima Indians Diabetes. من خلال اختيار الميزات الأكثر إفادة، يمكن لنماذج التصنيف تحسين القدرات التنبئية لمرض السكري من خلال تحسين الدقة ومقاييس التقييم الأخرى. بدقة ٨٦,٠%، فإن النموذج الذي يستخدم أفضل خمس ميزات فقط قادر على تصنيف البيانات بمستوى أعلى من الدقة من الطرز الأخرى

BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus merupakan salah satu penyakit kronis yang menjadi masalah serius dalam dunia kesehatan global (H. Nugroho et al., 2023). Penyakit ini terjadi ketika tubuh tidak dapat memproduksi atau menggunakan insulin dengan baik, sehingga mengakibatkan peningkatan kadar glukosa (gula) dalam darah. Tingginya kadar glukosa dalam darah dapat menyebabkan berbagai komplikasi serius, termasuk kerusakan pada organ tubuh seperti mata, ginjal, saraf, dan jantung.

Dalam upaya pencegahan dan pengelolaan diabetes yang lebih baik pengembangan model klasifikasi yang akurat untuk mendiagnosis diabetes pada individu menjadi sangat penting. Salah satu algoritma klasifikasi yang umum digunakan adalah Gaussian Naïve Bayes (GNB). GNB adalah metode klasifikasi probabilitas yang berdasarkan pada asumsi bahwa fitur-fitur yang ada dalam data adalah independen secara kondisional.

Namun, tidak semua fitur dalam dataset memberikan kontribusi yang signifikan terhadap performa model klasifikasi. Terkadang, beberapa fitur dapat menjadi redundan atau memiliki pengaruh yang kurang signifikan terhadap klasifikasi yang akurat. Oleh karena itu, seleksi fitur menjadi langkah yang penting dalam membangun model klasifikasi yang efektif.

Salah satu metode seleksi fitur yang umum digunakan adalah metode Anova Feature Selection. Anova Feature Selection digunakan untuk menguji apakah ada perbedaan signifikan antara rata-rata kelompok-kelompok yang berbeda dalam suatu variabel. Dalam konteks ini, Anova Feature Selection dapat digunakan untuk mengevaluasi apakah terdapat perbedaan signifikan antara kelompok individu yang menderita diabetes dan kelompok individu yang tidak menderita diabetes dalam setiap fitur yang ada dalam dataset Pima Indians Diabetes.

Dalam ajaran agama Islam, terdapat sebuah hadis yang menyampaikan:

بالحرام تداووا ولا فتداووا وَاءً َد دَاءٍ لِكُلِّ وَجَعَلَ وَالدَّوَاءَ الدَّاءَ أَنْزَلَ تعالى الله إن

"Sesungguhnya Allah menurunkan penyakit dan obatnya dan menjadikan bagi setiap penyakit ada obatnya. Maka berobatlah kalian, dan jangan kalian berobat dengan yang haram". (HR. Abu Dawud dari Abu Darda).

Syaikh Muhammad bin Shalih Al-Utsaimin Rahimahullah menjelaskan bahwa pada dasarnya berobat itu wajib hukumnya. Meninggalkannya berarti membiarkan diri kita dalam keadaan bahaya. Artinya sakit itu di sisi lain bisa bermakna cobaan, ujian dari Allah SWT (R. Tantawi, 2019).

Pesan dalam hadis ini sejalan dengan tujuan dari penelitian yang sedang dilakukan. Penelitian ini bertujuan untuk menemukan model klasifikasi terbaik dalam mendeteksi penyakit diabetes berdasarkan dataset yang tersedia. Dengan menjalankan penelitian ini, diharapkan dapat "menemukan obat" atau model klasifikasi terbaik untuk penyakit diabetes, sejalan dengan pesan hadis yang menyatakan bahwa setiap penyakit memiliki obatnya. Dan juga hasil dari

penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam proses diagnosis penyakit diabetes di masa depan.

Dengan demikian, penelitian ini bertujuan untuk memberikan kontribusi penting dalam upaya pencegahan dan pengelolaan diabetes yang lebih baik, sehingga dapat mengurangi beban penyakit dan meningkatkan kualitas hidup para penderitanya. Dalam pandangan agama Islam, upaya ini juga diharapkan sebagai bagian dari usaha dalam menuntut ilmu dan mengaplikasikannya demi kemaslahatan umat manusia.

1.2 Rumusan Masalah

Bagaimana pengaruh seleksi fitur ANOVA pada pemrosesan data awal terhadap performa model klasifikasi menggunakan metode Gaussian Naïve Bayes?

1.3 Batasan Masalah

1. Sumber data untuk penelitian ini adalah dataset PIMA Indians. Dataset ini mencakup informasi kesehatan dari wanita PIMA Indians, termasuk atribut-atribut seperti usia, kandungan glukosa plasma, tekanan darah, ketebalan lipatan kulit, dan lain-lain.
2. Data untuk penelitian ini akan diambil dari sumber dataset yang telah tersedia secara publik. Tidak ada proses pengumpulan data lanjutan yang dilakukan oleh peneliti.

3. Algoritma Gaussian Naïve Bayes, KNN Imputation, dan seleksi fitur ANOVA akan diimplementasikan menggunakan bahasa pemrograman Python dengan bantuan pustaka scikit-learn.

1.4 Tujuan Penelitian

Untuk mengetahui bagaimana pengaruh seleksi fitur ANOVA pada pemrosesan data awal terhadap performa model klasifikasi menggunakan metode Gaussian Naïve Bayes.

1.5 Manfaat Penelitian

Dengan melakukan penelitian ini, diharapkan dapat memberikan manfaat dan keuntungan di kemudian hari. Beberapa contoh manfaat yang diharapkan termasuk hal-hal berikut.

1. Kontribusi positif bagi dunia kesehatan dalam upaya pencegahan dan pengelolaan diabetes.
2. Memberikan panduan bagi praktisi medis dalam mengambil keputusan berdasarkan hasil prediksi yang lebih dapat diandalkan.
3. Membuka peluang pengembangan teknologi medis yang lebih maju berdasarkan hasil penelitian ini.

BAB II
STUDI PUSTAKA

2.1 Penelitian Terkait

Pada penelitiannya (V. Chang et al., 2022) hasil untuk subset data 3-faktor dan 5-faktor yang menggunakan seleksi fitur menunjukkan bahwa model klasifikasi Naïve Bayes memiliki performa lebih baik dalam hal akurasi dibandingkan dengan random forest dan decision tree J48. Model Naïve Bayes pada subset data 3-faktor memiliki akurasi yang sama baiknya dengan model random forest pada dataset lengkap, yaitu 79,13% dibandingkan dengan 79,57%, yang merupakan akurasi tertinggi dalam eksperimen ini. Ini mengindikasikan bahwa model Naïve Bayes bekerja baik dengan pemilihan fitur yang lebih disesuaikan, tetapi kurang efektif dengan fitur yang banyak.

Pada penelitian yang dilakukan (S. Shakeela et al., 2021) seleksi fitur ANOVA efisien untuk menemukan informasi fitur yang esensial dari dataset. Tujuan penelitiannya adalah untuk mengevaluasi metode yang dikembangkan untuk mendeteksi serangan jahat (malicious attacks) dalam jaringan komunikasi data. Desain penelitiannya bekerja dengan baik jika fitur-fitur yang relevan dengan kelas serangan yang disebutkan tersedia dalam skenario pemantauan waktu nyata. Fokus utama penelitiannya adalah mempelajari fitur-fitur data. Pada penelitiannya tidak selalu perlu mempertimbangkan seluruh set fitur. Penggunaan seluruh set fitur dapat meningkatkan peluang deteksi palsu dan memerlukan lebih banyak waktu untuk memproses hasil.

Hasil penelitian dari (M. Alassaf & A. M. Qamar, 2022) menunjukkan bahwa Support Vector Machine (SVM) dan Naive Bayes (NB) memiliki performa sangat baik ketika digabungkan dengan seleksi fitur ANOVA, melebihi hasil eksperimen dasar. Tema penelitiannya adalah Analisis Sentimen (SA), sebagai sistem klasifikasi teks, bertujuan untuk mengidentifikasi pendapat penulis melalui teks. penelitian tersebut mengumpulkan tweet berbahasa Arab yang ditulis tentang sebuah universitas tertentu dan mengkaji potensi penggunaan seleksi fitur ANOVA dalam SA. Percobaan dalam tahap FS dibagi menjadi dua metode. Yang pertama berdasarkan peringkat fitur berdasarkan persentase tertentu dari nilai F tertinggi dalam ANOVA satu arah. Demikian pula, metode kedua bekerja dengan memilih fitur berdasarkan p-value.

2.2 Diabetes Mellitus

Diabetes Mellitus, yang umumnya disebut diabetes, adalah suatu kelompok penyakit metabolik yang ditandai oleh tingginya kadar glukosa (gula) dalam darah, yang disebut juga dengan hiperglikemia (A. F. Ilmi & D. M. Utari, 2020). Glukosa merupakan sumber utama energi bagi tubuh dan diatur oleh hormon insulin yang diproduksi oleh pankreas. Namun, pada penderita diabetes, sistem pengaturan glukosa ini mengalami gangguan, menyebabkan kadar glukosa dalam darah tetap tinggi, baik karena kurangnya produksi insulin, kerja insulin yang tidak efektif, atau kombinasi keduanya.

2.3 Machine Learning

Machine Learning (ML) adalah sub bidang dari kecerdasan buatan (AI) yang berfokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk belajar dan meningkatkan kinerjanya dalam tugas tertentu tanpa harus diprogram secara eksplisit (I. D. Id, 2021). Dalam pemrograman tradisional, para pengembang menulis aturan dan instruksi eksplisit yang harus diikuti oleh komputer. Namun, dalam machine learning, komputer menggunakan data untuk mempelajari pola dan hubungan, sehingga dapat membuat prediksi, mengenali pola, atau memecahkan masalah tanpa harus diprogram secara eksplisit untuk setiap kasus spesifik. Ide inti di balik machine learning adalah mengembangkan algoritma yang dapat belajar dari data dan membuat prediksi atau keputusan berdasarkan data tersebut.

Machine learning telah mendapatkan adopsi yang luas dan telah merevolusi berbagai industri, termasuk kesehatan, keuangan, transportasi, pemasaran, dan lain-lain. Pengembangan di bidang ini terus aktif, dengan algoritma dan teknik baru terus dikembangkan untuk meningkatkan kinerja dan mengatasi berbagai tantangan.

2.4 Naïve Bayes

Metode Naïve Bayes adalah salah satu metode yang popular digunakan untuk keperluan data mining karena kemudahan penggunaannya serta waktu pemrosesannya yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan tingkat efektifitas yang tinggi (R. Harjadinata, 2022).

Pendekatan Naïve Bayes didasarkan pada teorema Bayes yang memungkinkan kita untuk menghitung probabilitas suatu peristiwa berdasarkan informasi yang ada. Dalam konteks klasifikasi, Naïve Bayes memprediksi label atau kelas yang paling mungkin untuk suatu data berdasarkan fitur-fitur yang diamati.

Namun, aspek "naive" dari Naïve Bayes adalah menganggap setiap fitur atau atribut sebagai independen satu sama lain, meskipun dalam dunia nyata beberapa fitur mungkin saling terkait. Ini adalah asumsi yang sering tidak realistis, tetapi dalam implementasinya, metode ini dapat memberikan hasil yang baik dalam banyak kasus praktis.

2.5 Feature Selection

Seleksi Fitur (FS) adalah metode paling umum menggunakan pendekatan reduksi dimensi. FS digunakan untuk membersihkan data yang memiliki banyak noise, berlebihan atau tidak relevan. Sebagian besar hasil fari FS adalah performa yang meningkat (B. Venkatesh & J. Anuradha, 2019). Seleksi fitur merujuk pada langkah memperoleh subset dari himpunan fitur asli sesuai dengan suatu kriteria seleksi fitur tertentu, yang menentukan fitur-fitur yang relevan dari dataset. Proses ini memiliki peran dalam mengurangi skala pemrosesan data dengan menghilangkan fitur-fitur yang redundan dan tidak relevan. Teknik seleksi fitur dapat digunakan sebagai langkah pra-pemrosesan untuk algoritma pembelajaran, dan hasil seleksi fitur yang baik dapat meningkatkan akurasi pembelajaran, mengurangi waktu pembelajaran, serta menyederhanakan hasil pembelajaran. Penting untuk dicatat bahwa seleksi fitur dan ekstraksi fitur merupakan dua

metode pengurangan dimensionalitas. Berbeda dengan seleksi fitur, ekstraksi fitur umumnya melibatkan transformasi data asli menjadi fitur-fitur dengan kemampuan pengenalan pola yang kuat, di mana data asli dapat dianggap sebagai fitur dengan kemampuan pengenalan pola yang kurang optimal. (J. Cai et al., 2018).

2.6 ANOVA Feature Selection

Seleksi fitur ANOVA (Analysis of Variance) merupakan metode untuk menentukan apakah terdapat perbedaan signifikan antara rata-rata kelompok-kelompok yang berbeda dalam suatu dataset (A. D. N. Mazlan et al,. 2020). Dalam konteks seleksi fitur, ANOVA digunakan untuk mengidentifikasi fitur-fitur yang memiliki pengaruh signifikan terhadap perbedaan antara kelompok-kelompok tersebut.

Secara umum, seleksi fitur ANOVA bekerja dengan menghitung nilai statistik ANOVA untuk setiap fitur dalam dataset. Statistik ini memberikan informasi tentang sejauh mana variasi atau perbedaan antara rata-rata kelompok-kelompok yang berbeda. Jika nilai statistik ANOVA tinggi, itu menandakan bahwa terdapat perbedaan yang signifikan antara kelompok-kelompok tersebut untuk fitur tersebut.

Fitur-fitur yang memiliki dampak signifikan terhadap perbedaan antar kelompok kemudian dapat dipilih untuk digunakan dalam analisis lebih lanjut atau dalam pembangunan model. Seleksi fitur ANOVA berguna terutama ketika Anda ingin mengidentifikasi fitur-fitur yang paling informatif untuk membedakan kelompok-kelompok dalam data Anda.

2.7 Evaluation Metrics

Metrik evaluasi memainkan peran penting dalam mencapai pengklasifikasian optimal selama proses klasifikasi (M. Hossin, & M. N. Sulaiman, 2015). Metrik evaluasi mengacu pada kriteria atau ukuran yang digunakan untuk menilai kinerja suatu sistem, model, proses, atau entitas lainnya. Metrik ini memberikan cara kuantitatif atau kualitatif untuk menilai seberapa baik suatu sistem atau model dalam mencapai tujuan yang dimaksud. Metrik evaluasi banyak digunakan dalam berbagai bidang, termasuk pembelajaran mesin (machine learning), analisis data, pengembangan perangkat lunak, dan bisnis, untuk mengukur dan meningkatkan efektivitas proses dan hasil.

Dalam konteks pembelajaran mesin, metrik evaluasi sangat penting untuk menilai kinerja model pada tugas-tugas tertentu. Tugas-tugas yang berbeda (misalnya, klasifikasi, regresi, pengelompokan) mungkin memerlukan metrik yang berbeda pula. Beberapa metrik evaluasi umum dalam pembelajaran mesin (machine learing) adalah akurasi, presisi, recall, spesifisitas dan f1-score.

2.7.1 Akurasi

Dalam konteks evaluasi model, "akurasi" (accuracy) adalah metrik yang digunakan untuk mengukur sejauh mana model dapat mengklasifikasikan contoh secara benar. Akurasi menggambarkan persentase keseluruhan contoh yang diklasifikasikan dengan benar oleh model.

Secara matematis, akurasi dihitung dengan membagi jumlah contoh yang diklasifikasikan dengan benar oleh model (true positive + true negative) dengan jumlah total contoh dalam dataset. Hasilnya adalah angka antara 0 hingga 1, di

mana nilai 1 menunjukkan akurasi yang sempurna (model mengklasifikasikan semua contoh dengan benar) dan nilai 0 menunjukkan akurasi yang sangat buruk (model tidak mengklasifikasikan contoh dengan benar).

Akurasi sering digunakan sebagai metrik evaluasi awal untuk mengukur performa model secara keseluruhan. Metrik ini memberikan gambaran tentang tingkat kecocokan model dalam mengklasifikasikan contoh secara benar, termasuk contoh positif dan negatif.

Namun, perlu diingat bahwa akurasi dapat memberikan informasi yang bias jika dataset memiliki ketidakseimbangan antara kelas positif dan negatif. Misalnya, jika terdapat jumlah contoh positif yang jauh lebih sedikit daripada contoh negatif, model yang cenderung mengklasifikasikan semua contoh sebagai negatif dapat menghasilkan akurasi yang tinggi secara bias.

Dalam beberapa kasus, metrik evaluasi lain seperti presisi (precision), recall, atau F1 score perlu dipertimbangkan bersama dengan akurasi untuk mendapatkan gambaran yang lebih lengkap tentang performa model. Metrik-metrik ini memberikan informasi lebih spesifik tentang kemampuan model dalam mengenali contoh positif dan negatif, serta membantu mengatasi bias yang mungkin terjadi akibat ketidakseimbangan kelas.

Dalam kesimpulannya, akurasi adalah metrik evaluasi yang mengukur sejauh mana model dapat mengklasifikasikan contoh dengan benar secara keseluruhan. Namun, akurasi perlu dianalisis bersama dengan metrik lainnya dan dapat menjadi bias dalam kasus ketidakseimbangan kelas.

2.7.2 Presisi

Dalam konteks evaluasi performa model klasifikasi, "presisi" (precision) mengacu pada metrik yang mengukur sejauh mana contoh yang diklasifikasikan sebagai positif oleh model benar-benar relevan. Presisi memberikan gambaran tentang jumlah contoh positif yang berhasil diidentifikasi dengan benar oleh model dibandingkan dengan total jumlah contoh yang diklasifikasikan sebagai positif oleh model.

Secara matematis, presisi dihitung dengan membagi jumlah contoh positif yang benar diidentifikasi oleh model (true positive) dengan jumlah total contoh yang diklasifikasikan sebagai positif oleh model (true positive + false positive). Hasilnya adalah angka antara 0 hingga 1, di mana nilai 1 menunjukkan presisi yang sempurna, sementara nilai 0 menunjukkan presisi yang sangat buruk.

Berdasarkan definisi di atas, presisi berfokus pada kemampuan model untuk menghindari kesalahan "false positive," yaitu ketika model salah mengklasifikasikan contoh negatif sebagai positif. Oleh karena itu, presisi yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan contoh yang relevan sebagai positif.

Namun, perlu diingat bahwa presisi harus dianalisis bersama dengan metrik lain, seperti "recall" dan "F1 score," untuk mendapatkan gambaran yang lebih lengkap tentang performa keseluruhan model. Recall mengukur kemampuan model dalam mengidentifikasi semua contoh positif yang sebenarnya, sementara F1 score menggabungkan informasi tentang presisi dan recall dalam satu angka.

Dalam kesimpulannya, presisi adalah metrik evaluasi yang mengukur sejauh mana contoh yang diklasifikasikan sebagai positif oleh model benar-benar relevan. Nilai presisi yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan contoh yang relevan sebagai positif, sementara nilai presisi yang rendah menunjukkan kecenderungan model menghasilkan kesalahan "false positive".

2.7.3 Recall

Dalam konteks evaluasi performa model klasifikasi, "recall" mengacu pada metrik yang digunakan untuk mengukur sejauh mana model mampu mengidentifikasi semua contoh positif yang sebenarnya. Lebih khususnya, recall memberikan gambaran tentang jumlah contoh positif yang berhasil diidentifikasi oleh model dibandingkan dengan total jumlah contoh positif yang sebenarnya ada dalam dataset.

Secara matematis, recall dihitung dengan membagi jumlah contoh positif yang benar diidentifikasi oleh model (true positive) dengan jumlah total contoh positif yang sebenarnya ada dalam dataset. Hasilnya adalah angka antara 0 hingga 1, di mana nilai 1 menunjukkan recall yang sempurna, sementara nilai 0 menunjukkan recall yang sangat buruk.

Berdasarkan definisi di atas, recall berfokus pada kemampuan model untuk menghindari kesalahan "false negative," yaitu ketika model gagal mengidentifikasi contoh positif yang sebenarnya. Oleh karena itu, recall yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali contoh positif yang ada.

Namun, perlu diingat bahwa recall harus dianalisis bersama dengan metrik lain, seperti "precision" (presisi), untuk mendapatkan gambaran yang lebih lengkap tentang performa keseluruhan model. Precision mengukur sejauh mana contoh yang diklasifikasikan sebagai positif oleh model benar-benar relevan. Kedua metrik ini sering digunakan bersama-sama dalam evaluasi model klasifikasi, terutama ketika dataset memiliki ketidakseimbangan kelas, yaitu jumlah contoh positif dan negatif yang tidak seimbang.

Dalam kesimpulannya, recall adalah metrik evaluasi yang mengukur kemampuan model untuk mengidentifikasi contoh positif yang sebenarnya. Nilai recall yang tinggi menunjukkan bahwa model mampu mengenali sebagian besar contoh positif yang ada, sementara nilai recall yang rendah menunjukkan kecenderungan model menghasilkan kesalahan "false negative."

2.7.4 Spesifisitas

Dalam konteks evaluasi performa model klasifikasi, "spesifisitas" mengacu pada metrik yang digunakan untuk mengukur sejauh mana model mampu mengidentifikasi semua contoh negatif yang sebenarnya. Lebih spesifiknya, spesifisitas memberikan gambaran tentang jumlah contoh negatif yang berhasil diidentifikasi dengan benar oleh model dibandingkan dengan total jumlah contoh negatif yang sebenarnya ada dalam dataset.

Secara matematis, spesifisitas dihitung dengan membagi jumlah contoh negatif yang benar diidentifikasi oleh model (true negative) dengan jumlah total contoh negatif yang sebenarnya ada dalam dataset. Hasilnya adalah angka antara 0

hingga 1, di mana nilai 1 menunjukkan spesifisitas yang sempurna, sementara nilai 0 menunjukkan spesifisitas yang sangat buruk.

Berdasarkan definisi di atas, spesifisitas berfokus pada kemampuan model untuk menghindari kesalahan "false positive," yaitu ketika model salah mengidentifikasi contoh negatif sebagai positif. Oleh karena itu, spesifisitas yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali contoh negatif yang ada.

Seperti halnya recall, spesifisitas juga harus dianalisis bersama dengan metrik lain, seperti precision, untuk mendapatkan gambaran yang lebih lengkap tentang performa keseluruhan model. Precision mengukur sejauh mana contoh yang diklasifikasikan sebagai positif oleh model benar-benar relevan. Kedua metrik ini sering digunakan bersama-sama dalam evaluasi model klasifikasi.

Dalam kesimpulannya, spesifisitas adalah metrik evaluasi yang mengukur kemampuan model untuk mengidentifikasi contoh negatif yang sebenarnya. Nilai spesifisitas yang tinggi menunjukkan bahwa model mampu mengenali sebagian besar contoh negatif yang ada, sementara nilai spesifisitas yang rendah menunjukkan kecenderungan model menghasilkan kesalahan "false positive".

2.7.5 F1-Score

F1-score adalah metrik evaluasi yang digunakan untuk menggabungkan informasi tentang presisi (precision) dan recall dari model klasifikasi ke dalam satu angka yang merepresentasikan performa keseluruhan model. Metrik ini berguna ketika kita ingin memperhatikan keseimbangan antara presisi dan recall dalam pengambilan keputusan.

F1-score akan memiliki nilai antara 0 hingga 1, di mana nilai 1 menunjukkan performa yang sempurna dan nilai 0 menunjukkan performa yang buruk.

F1-score berguna ketika kita memiliki ketidakseimbangan antara kelas positif dan negatif dalam dataset. Misalnya, jika kita memiliki dataset di mana jumlah contoh positif jauh lebih sedikit daripada contoh negatif, maka model yang hanya mencapai tingkat presisi yang tinggi dengan mengklasifikasikan semua contoh sebagai negatif akan memiliki nilai presisi yang tinggi tetapi recall yang rendah. Sebaliknya, model yang mengklasifikasikan banyak contoh sebagai positif akan memiliki recall yang tinggi tetapi presisi yang rendah. Dalam kasus ini, F1 score memberikan gambaran yang lebih seimbang tentang performa keseluruhan model.

Dalam kesimpulannya, F1-score adalah metrik evaluasi yang menggabungkan informasi tentang presisi dan recall dalam satu angka untuk merepresentasikan performa keseluruhan model klasifikasi. Metrik ini berguna ketika kita ingin memperhatikan keseimbangan antara presisi dan recall dalam pengambilan keputusan, terutama dalam kasus ketidakseimbangan antara kelas positif dan negatif dalam dataset.

BAB III
DESAIN DAN IMPLEMENTASI

3.1 Desain Penelitian

Pada bab ini, akan dijelaskan secara rinci mengenai metodologi yang digunakan dalam penelitian ini untuk melakukan klasifikasi penyakit diabetes menggunakan metode Naïve Bayes. Bab ini mencakup langkah-langkah yang diambil untuk mengumpulkan data, preprocessing, seleksi fitur, melakukan pelatihan model, evaluasi model, serta analisis hasil.

3.2 Pengumpulan Data

Proses pengumpulan data merupakan langkah awal dalam menjalankan penelitian ini. Data yang digunakan dalam penelitian ini berasal dari dataset Wanita Pima Indian Diabetes. Dataset Pima Indian diperoleh dari National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dan tersedia secara bebas melalui platform *Kaggle* dan dapat di unduh dari sumber https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Dataset ini memiliki catatan medis dari sejumlah pasien wanita Pima Indian yang meliputi berbagai variabel yang berkaitan dengan faktor risiko diabetes. Data yang terkandung dalam dataset ini telah digunakan dalam penelitian sebelumnya dan dianggap sebagai sumber yang kredibel dalam menganalisis hubungan antara faktor-faktor tersebut dengan perkembangan penyakit diabetes. Contoh dataset Pima Indian adalah sebagai berikut.

Tabel 3.1 Dataset Pima Indians Diabetes

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35		33,6	0,627	50	1
1	85	66	29		26,6	0,351	31	0
8	183	64			23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
5	116	74			25,6	0,201	30	0
3	78	50	32	88	31	0,248	26	1
10	115				35,3	0,134	29	0
2	197	70	45	543	30,5	0,158	53	1
8	125	96				0,232	54	1
4	110	92			37,6	0,191	30	0
10	168	74			38	0,537	34	1
10	139	80			27,1	1,441	57	0
1	189	60	23	846	30,1	0,398	59	1
5	166	72	19	175	25,8	0,587	51	1
7	100				30	0,484	32	1
0	118	84	47	230	45,8	0,551	31	1
7	107	74			29,6	0,254	31	1
1	103	30	38	83	43,3	0,183	33	0
1	115	70	30	96	34,6	0,529	32	1
3	126	88	41	235	39,3	0,704	27	0
8	99	84			35,4	0,388	50	0
7	196	90			39,8	0,451	41	1
9	119	80	35		29	0,263	29	1
11	143	94	33	146	36,6	0,254	51	1
10	125	70	26	115	31,1	0,205	41	1
7	147	76			39,4	0,257	43	1
1	97	66	15	140	23,2	0,487	22	0
13	145	82	19	110	22,2	0,245	57	0
5	117	92			34,1	0,337	38	0

Dataset Pima Indian Diabetes mencakup beberapa atribut seperti usia, jumlah kehamilan, kadar glukosa darah, tekanan darah, BMI (Body Mass Index), riwayat keluarga diabetes, dan atribut lainnya. Dataset memiliki 9 kolom dan 768 baris (500 non-penderita diabetes dan 268 penderita diabetes). Variabel hasil atau

target berupa angka (0 atau 1), di mana 0 menunjukkan tes negatif untuk diabetes, dan 1 menyiratkan tes positif. Berikut adalah penjelasan tentang semua atribut/kolom dalam dataset Pima Indian Diabetes:

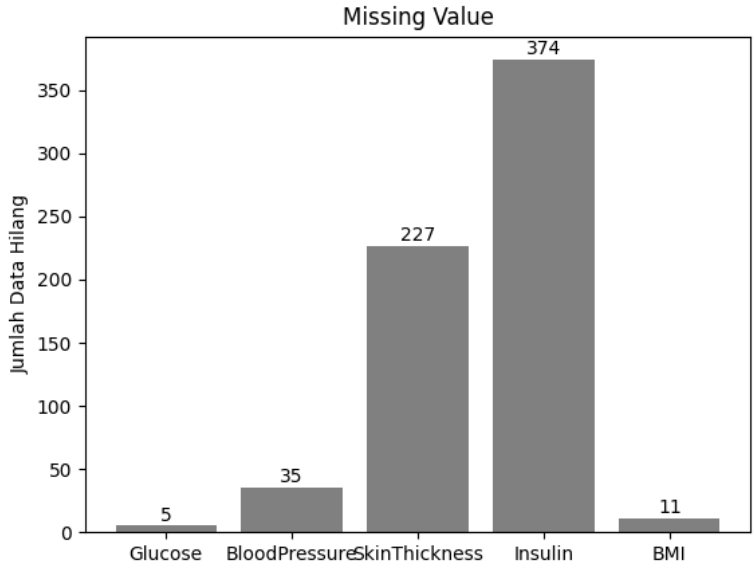
1. Pregnancies (Kehamilan): Jumlah kali wanita tersebut hamil.
2. Glucose (Glukosa): Kadar glukosa dalam plasma puasa pasien, yang merupakan indikator penting dalam diagnosis diabetes.
3. Blood Pressure (Tekanan Darah): Tekanan darah sistolik pasien dalam mm Hg.
4. Skin Thickness (Tebal Kulit): Ketebalan lipatan kulit pada daerah trisep, yang dapat memberikan informasi tentang lemak tubuh dan faktor risiko diabetes.
5. Insulin: Tingkat insulin dalam serum pasien, yang berhubungan dengan fungsi pankreas dan metabolisme glukosa.
6. BMI (Body Mass Index): Indeks massa tubuh, dihitung dari berat badan (kg) dibagi kuadrat tinggi badan (m^2). Indeks ini digunakan untuk menilai status gizi dan risiko obesitas.
7. Diabetes Pedigree Function (Fungsi Pedigri Diabetes): Skor yang mencerminkan sejauh mana ada riwayat diabetes dalam keluarga pasien.
8. Age (Usia): Usia pasien dalam tahun.

Outcome: Variabel target, menunjukkan apakah pasien menderita diabetes (1) atau tidak (0).

3.3 Preprocessing Data

Sebelum data digunakan untuk melatih model Naïve Bayes, data perlu melewati proses preprocessing. Proses preprocessing mencakup langkah-langkah seperti mengatasi missing values, mengubah tipe data jika diperlukan, serta pemilihan atribut yang relevan untuk digunakan dalam pelatihan model (Pebdika et al., 2023). Preprocessing data bertujuan untuk memastikan kualitas data yang baik dan menghindari bias dalam hasil prediksi.

Salah satu permasalahan yang relevan dalam konteks kualitas data adalah keberadaan data yang hilang (missing value). Data yang tidak lengkap bisa berasal dari berbagai sumber seperti catatan tentang kematian pasien, rusaknya peralatan, penolakan responden untuk menjawab pertanyaan tertentu, dan lain sebagainya. Setelah memeriksa setiap fitur pada dataset Pima Indian diketahui secara keseluruhan 51% kasus memiliki nilai yang hilang dari satu atau lebih atribut. Jumlah nilai yang hilang dari berbagai atribut diberikan pada Gambar 3.1.



Gambar 3.1 Jumlah Missing Value

Setelah mengetahui jumlah data yang hilang pada setiap fitur/atribut, dilakukan pengecekan pada setiap baris untuk mengetahui berapa jumlah fitur yang hilang pada setiap kasus/sampelnya. Ditemukan bahwa terdapat 7 kasus dengan empat nilai atribut yang hilang, 28 kasus dengan tiga nilai atribut yang hilang, 199 kasus dengan dua nilai atribut yang hilang, 142 kasus dengan satu nilai atribut hilang dan 392 kasus tidak ada yang hilang. Oleh karena itu kasus yang memiliki jumlah atribut hilang lebih dari satu di hapus semuanya dan sekarang tersisa 534 data yang berguna untuk analisis lebih lanjut (178 kasus positif dan 369 kasus negatif). Berikut adalah contoh dataset setelah dilakukan penghapusan baris data yang memiliki lebih dari satu atribut/fitur yang hilang.

Tabel 3.2 Data Cleaned

Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome**
6	148	72	35		33,6	0,627	50	1
1	85	66	29		26,6	0,351	31	0
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
3	78	50	32	88	31	0,248	26	1
2	197	70	45	543	30,5	0,158	53	1
1	189	60	23	846	30,1	0,398	59	1
5	166	72	19	175	25,8	0,587	51	1
0	118	84	47	230	45,8	0,551	31	1
1	103	30	38	83	43,3	0,183	33	0
1	115	70	30	96	34,6	0,529	32	1
3	126	88	41	235	39,3	0,704	27	0
9	119	80	35		29	0,263	29	1
11	143	94	33	146	36,6	0,254	51	1
10	125	70	26	115	31,1	0,205	41	1
6	148	72	35		33,6	0,627	50	1
1	85	66	29		26,6	0,351	31	0

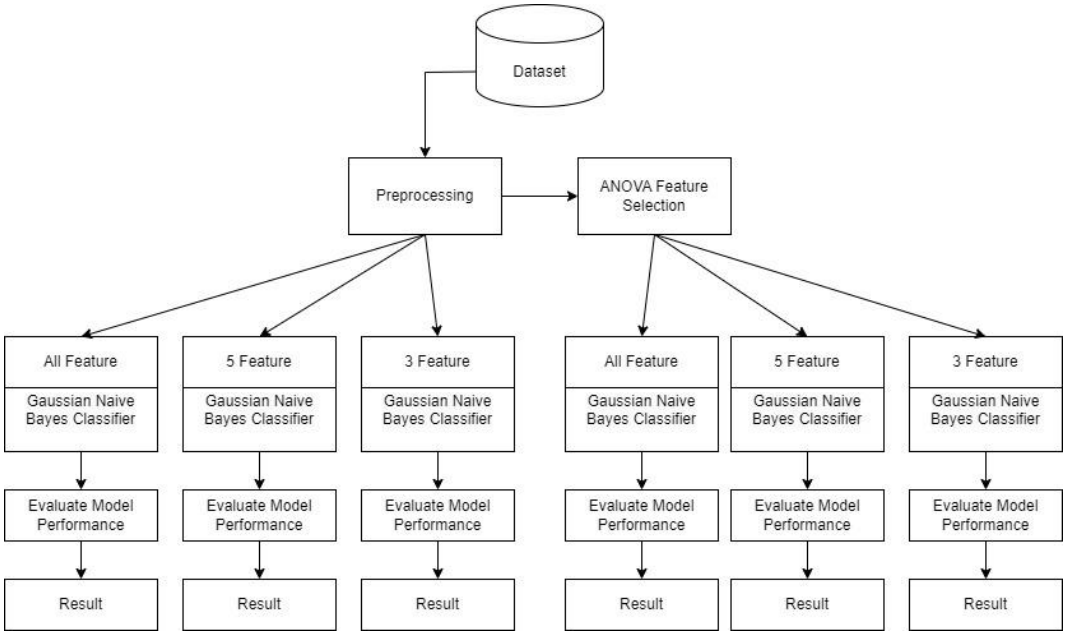
Untuk mengisi nilai yang hilang, digunakan metode K-Nearest Neighbors Imputation. Metode K-Nearest Neighbors (KNN) Imputation merupakan salah satu teknik dalam pengolahan data yang digunakan untuk mengatasi masalah data yang hilang atau missing data. Metode ini didasarkan pada konsep bahwa data yang hilang dapat diestimasi dengan mempertimbangkan nilai-nilai dari tetangga terdekat dalam ruang fitur.

3.4 Skenario Pengujian

Untuk menentukan jumlah fitur yang optimal guna meningkatkan performa klasifikasi, akan diimplementasikan enam model pengujian terdiri dari model A sampai model F. Pada tahap awal atau Model A, kami akan menggunakan semua fitur yang ada dalam dataset hasil implementasi seleksi fitur ANOVA. Pada tahap berikutnya atau Model B, kami akan menggunakan semua fitur secara acak tanpa melakukan seleksi fitur menggunakan metode ANOVA.

Setelah itu Model C melibatkan pemilihan lima fitur terbaik berdasarkan proses seleksi fitur ANOVA. Tahap selanjutnya Model D menggunakan lima fitur acak tanpa melalui proses seleksi fitur. Lalu Model E melibatkan pemilihan tiga fitur terbaik berdasarkan proses seleksi fitur ANOVA, dan yang terakhir Model F menggunakan tiga fitur acak tanpa melalui proses seleksi fitur.

Pendekatan ini bertujuan untuk memberikan pemahaman yang lebih mendalam tentang bagaimana performa klasifikasi dipengaruhi oleh jumlah fitur yang digunakan. Gambar 3.2. menunjukkan desain scenario pengujian yang akan diterapkan pada penelitian ini.



Gambar 3.2 Skenario Pengujian

Berikut adalah rincian model pengujian yang akan dibangun berdasarkan skenario uji yang telah dibuat.

Model	**Jumlah Fitur**	**Deskripsi**
A	Semua Fitur	Seleksi fitur ANOVA pada seluruh fitur dataset
B	Semua Fitur	Tanpa seleksi fitur, menggunakan semua fitur secara acak
C	Lima Fitur	Seleksi lima fitur terbaik berdasarkan ANOVA
D	Lima Fitur	Tanpa seleksi fitur, menggunakan lima fitur secara acak
E	Tiga Fitur	Seleksi tiga fitur terbaik berdasarkan ANOVA
F	Tiga Fitur	Tanpa seleksi fitur, menggunakan tiga fitur secara acak

3.5 Perhitungan Manual KNN Imputation

Berikut adalah prosedur atau langkah-langkah untuk melakukan imputasi data menggunakan metode K-Nearest Neighbors (KNN):

Tabel 3.3 Contoh imputasi missing value dengan KNN Imputation

No	Insulin	Blood Pressure	BMI	Glucose
1	207	74	26.6	189
2	175	NaN	25.6	166
3	235	96	33.6	197
4	190	60	28.1	177

Di dalam tabel tersebut, ditemukan nilai yang hilang pada baris kedua pada fitur Tekanan Darah (Blood Pressure). Untuk melakukan imputasi dengan menggunakan metode KNN pada nilai yang hilang ini, langkah-langkah penerapannya adalah sebagai berikut.

1. Menentukan nilai k (misal k = 2).
2. Menghitung jarak euclidean pada suatu baris yang memiliki nilai yang hilang ke semua baris lainnya dengan rumus jarak euclidean:

$$(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

$$d(2,1) = \sqrt{(207 - 175)^2 + (26{,}6 - 25{,}6)^2 + (189 - 166)^2} = 39{,}42$$

$$d(2,3) = \sqrt{(235 - 175)^2 + (33{,}6 - 25{,}6)^2 + (197 - 166)^2} = 68{,}01$$

$$d(2,4) = \sqrt{(190 - 175)^2 + (18{,}1 - 25{,}6)^2 + (177 - 166)^2} = 20{,}05$$

3. Mengurutkan jarak terkecil hingga terbesar. Dalam perhitungan tersebut diketahui jarak baris 1 dan 4 adalah yang terdekat.
4. Menghitung rata-rata nilai fitur yang hilang pada 2 tetangga terdekat.

$$\bar{x} = \frac{74+60}{2} = 67$$

Setelah mengetahui nilai fitur yang hilang selanjutnya tinggal melakukan imputasi terhadap nilai yang hilang dengan hasil rata-rata 2 tetangga terdekat tersebut yaitu 67.

3.6 Implementasi KNN Imputation

Untuk melakukan proses imputasi data menggunakan metode K-Nearest Neighbors (KNN), dalam penelitian ini memanfaatkan fasilitas yang disediakan oleh kelas KNNImputer yang terdapat dalam library scikit-learn. Scikit-learn adalah suatu library atau pustaka yang populer dalam pengembangan dan penerapan algoritma pembelajaran mesin (machine learning) di lingkungan Python.

Metode K-Nearest Neighbors (KNN) yang diaplikasikan dalam konteks imputasi data memanfaatkan pendekatan berbasis tetangga terdekat untuk mengisi nilai-nilai yang hilang atau tidak lengkap dalam suatu dataset. Dengan menggunakan KNNImputer dari scikit-learn, kita dapat mengakses implementasi algoritma KNN yang telah dioptimalkan untuk tugas imputasi data, memungkinkan kita untuk meningkatkan kualitas dan kelengkapan dataset yang sedang diolah.

Langkah-langkah yang dilakukan dalam penggunaan KNNImputer melibatkan impor library scikit-learn, mempersiapkan dataset yang akan diimputasi, dan kemudian menerapkan fungsi-fungsi yang terdapat pada kelas KNNImpute yaitu fungsi fit_transform() untuk melakukan imputasi dengan mempertimbangkan informasi dari tetangga terdekat dalam ruang fitur yang

relevan. Berikut adalah hasil dataset yang telah diimputasi menggunakan metode Metode K-Nearest Neighbors (KNN) menggunakan library scikit-learn.

Tabel 3.4 Data Imputed

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
3	113	50	10	85	29,5	0,626	25	0
4	132	86	31	258,6	28	0,419	63	0
1	90	62	12	43	27,2	0,58	24	0
1	149	68	29	127	29,3	0,349	42	1
1	109	58	18	116	28,5	0,219	22	0
3	84	68	30	106	31,9	0,591	25	0
2	56	56	28	45	24,2	0,332	22	0
2	197	70	45	543	30,5	0,158	53	1
1	91	64	24	66,6	29,2	0,192	21	0
0	86	68	32	109,4	35,8	0,238	25	0
0	124	56	13	105	21,8	0,452	21	0
2	128	78	37	182	43,3	1,224	31	1
0	102	86	17	105	29,3	0,695	27	0
7	150	66	42	342	34,7	0,718	42	0
8	155	62	26	495	34	0,543	46	1
5	97	76	27	145,2	35,6	0,378	52	1
3	96	78	39	129	37,3	0,238	40	0
5	126	78	27	22	29,6	0,439	40	0
1	97	70	40	119,4	38,1	0,218	30	0
7	109	80	31	134,8	35,9	1,127	43	1

3.7 Perhitungan Manual ANOVA Feature Selection

Pada bab ini, akan dijelaskan secara rinci mengenai metodologi yang digunakan dalam penelitian ini untuk melakukan klasifikasi penyakit diabetes menggunakan metode Naïve Bayes. Bab ini mencakup langkah-langkah yang diambil untuk mengumpulkan data, preprocessing, seleksi fitur, melakukan pelatihan model, evaluasi model, serta analisis hasil.

Pemilihan Fitur (Feature Selection) adalah langkah penting dalam analisis data dan pembelajaran mesin (machine learing) di mana tujuannya adalah untuk mengidentifikasi dan memilih subset fitur yang paling informatif dan relevan dari dataset. Dalam konteks penelitian klasifikasi penyakit diabetes, pemilihan fitur bertujuan untuk mengurangi dimensi data dan meningkatkan efisiensi model.

Salah satu metode feature selection yang umum digunakan adalah Analysis of Variance (ANOVA). Tujuan ANOVA dalam konteks seleksi fitur adalah untuk mengidentifikasi fitur-fitur yang memiliki pengaruh yang signifikan terhadap variabel target atau output. Metode ANOVA dipilih karena dataset yang digunakan pada penelitian ini terdiri dari fitur berupa data kontinu dan target berupa data kategorik (J. Brownlee, 2019). Berikut adalah contoh penerapan seleksi fitur menggunakan metode ANOVA.

Tabel 3.5 Contoh penerapan ANOVA untuk seleksi fitur.

Blood Pressure	**Glucose**	**Age**	**Outcome**
62	92	46	0
106	68	47	0
76	89	23	0
70	137	22	0
68	187	41	1
90	180	35	1
68	90	27	1
90	196	41	1

Metode seleksi fitur ANOVA menggunakan nilai statistik uji F untuk menilai korelasi atau pengaruh atribut terhadap target. Semakin besar nilai F, semakin tinggi korelasi atau pengaruh atribut terhadap target.

Misal kita ingin menghitung statistik uji F untuk fitur umur (Age) terhadap variabel target berdasarkan tabel 2.2. Langkah-langkahnya adalah sebagai berikut.

1. Memisahkan fitur Age yang memiliki kelas target 0 dan yang memiliki kelas target 1.

Age(1)	**Age(0)**
41	46
35	47
27	23
41	22

2. Menghitung rata-rata dari setiap kolom.

$$\bar{x}\ Age(1) = \frac{41+35+27+41}{4} = 36$$

$$\bar{x}\ Age(0)\ = \frac{46+47+23+22}{4} = 34{,}5$$

3. Menghitung nilai grand mean.

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k}\left(\sum_{j-1}^{n_i} x_{ij}\right)}{N}$$

$$\bar{\bar{x}} = \frac{41+35+27+41+46+47+23+22}{8} = 35{,}25$$

4. Menghitung nilai MST.

$$MST = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{\bar{x}})^2}{k-1} = \frac{4(36-35{,}25)^2+\ 4(34{,}5-35{,}25)^2}{2-1} = 4{,}5$$

5. Menghitung nilai MSE

$$MSE = \frac{\sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2\right)}{n-k}$$

$$MSE = \frac{(41-36)^2+(35-36)^2+(27-36)^2+(41-36)^2+(46-34{,}5)^2+(47-34{,}5)^2+(23-34{,}5)^2+(22-34{,}5)^2}{8-2}$$

$$MSE = 118{,}167$$

6. Menghitung nilai F.

   $F = \frac{MST}{MSE} = 0{,}038081$

7. Setelah itu kita menghitung nilai F dari setiap fitur yang ada dan merangking mulai dari nilai F terbesar ke terkecil. Semakin besar nilai F semakin tinggi korelasi atau pengaruh atribut terhadap target.
   a. Glucose, F score: 5,44925
   b. Age, F score: 0,038081
   c. Blood Pressure, F score: 0,00188

Karena fitur kadar glukosa (Glucose) memiliki nilai F paling tinggi, maka dapat disimpulkan bahwa kadar glukosa memiliki pengaruh yang paling signifikan terhadap variabel target.

3.8 Implementasi ANOVA Feature Selection

Pada tahap ini, seleksi fitur menggunakan metode Analisis Varians (ANOVA) dilakukan untuk menentukan fitur-fitur yang paling signifikan dalam mempengaruhi variabel target dalam model klasifikasi. Hasil dari metode ANOVA memberikan nilai F-Score untuk setiap fitur yang diuji. Semakin besar nilai F maka semakin baik korelasi suatu fitur dengan target.

Untuk menerapkan metode seleksi fitur ANOVA, kami memanfaatkan fasilitas yang disediakan oleh library scikit-learn, khususnya menggunakan kelas SelectKBest dan f_classif. Scikit-learn, sebagai library atau pustaka pemelajaran mesin (machine learning) yang komprehensif di lingkungan Python, menyediakan alat atau tool yang kuat untuk melakukan analisis statistik seperti ANOVA dan

mempermudah implementasi seleksi fitur. Dalam penggunaan SelectKBest, kita dapat menentukan jumlah fitur terbaik yang ingin dipertahankan dalam proses seleksi. Fungsi f_classif yang terintegrasi dalam SelectKBest digunakan khusus untuk tugas klasifikasi, memungkinkan kita menilai signifikansi statistik dari setiap fitur terhadap pemisahan kelas dalam dataset. Langkah-langkah dalam menerapkan seleksi fitur dengan metode ANOVA menggunakan SelectKBest dan f_classif melibatkan import library scikit-learn, mempersiapkan dataset, dan kemudian menerapkan metode seleksi fitur pada dataset tersebut menggunakan fungsi fit_transform(). Hasil dari proses ini adalah subset fitur terbaik yang dapat digunakan dalam pembuatan model pembelajaran mesin, meningkatkan akurasi dan interpretabilitas model secara keseluruhan. Berdasarkan hasil perhitungan ANOVA, berikut adalah nilai F-Score untuk setiap fitur yang diuji:

Tabel 3.6 Hasil Implementasi feature selection ANOVA.

Rank	**Feature**	**F Score**
1	Glucose	181.00516536088764
2	Insulin	61.732592423768445
3	Age	59.36318855134598
4	BMI	53.268347684457666
5	SkinThickness	37.40419139547131
6	Pregnancies	36.798510885926476
7	DiabetesPedigreeFunction	28.679532491846235
8	BloodPressure	18.58550924119974

3.9 Metode Gaussian Naïve Bayes

Pada penelitian kali ini, Gaussian Naïve Bayes dipilih sebagai alternatif yang lebih tepat daripada Multinomial Naïve Bayes. Metode Gaussian Naïve Bayes diadopsi karena mengasumsikan bahwa data dalam setiap kelas mengikuti distribusi Gaussian, yang lebih sesuai untuk data kontinu (N. Rezaeian & G. Novikova, 2020). Sebaliknya, Multinomial Naïve Bayes lebih cocok untuk data diskrit, seperti dalam analisis teks. Berikut adalah rumus Gaussian Naive Bayes untuk menghitung probabilitas posterior dari suatu kelas C_k berdasarkan fitur-fitur $x_1, x_2, \dots x_n$:

$$P(C_k|x_1, x_2, \dots x_n) = \frac{P(x_1,x_2,\dots x_n|C_k)\cdot P(C_k)}{P(x_1,x_2,\dots x_n)}$$

Dimana:

- $P(C_k|x_1, x_2, \dots x_n)$ adalah probabilitas posterior dari kelas C_k setelah melihat data fitur.
- $P(x_1, x_2, \dots x_n|C_k)$ adalah likelihood dari data fitur $x_1, x_2, \dots x_n$ bila kelasnya adalah C_k
- $P(C_k)$ adalah prior probability dari kelas C_k
- $P(x_1, x_2, \dots x_n)$ adalah marginal probability dari data fitur. Dalam konteks klasifikasi marginal probability dapat diabaikan

Rumus Gaussian Naive Bayes untuk menghitung likelihood dari data fitur $x_1, x_2, \dots x_n$ bila kelasnya adalah C_k dapat dihitung menggunakan distribusi Gaussian:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\,\sigma_{k,i}^2}}\, e^{-\frac{\left(x_i-\mu_{k,i}\right)^2}{2\cdot\sigma_{k,i}^2}}$$

Dimana :

- $P(x_i|C_k)$ adalah likelihood dari fitur x_1 bila kelasnya adalah C_k
- $\mu_{k,i}$ adalah rata-rata (mean) dari fitur x_1 untuk kelas C_k
- $\sigma_{k,i}$ adalah standar deviasi dari fitur x_1 untuk kelas C_k

3.10 Perhitungan Manual Metode Gaussian Naïve Bayes

Tabel 3.7 Contoh penerapan Gaussian Naive Bayes.

Age	Blood Pressure	Outcome
78	81	1
70	74	1
62	69	1
70	68	1
62	54	0
70	52	0
67	49	0
69	48	0

Dengan dataset diatas misal kita ingin memprediksi seseorang pasien apakah menderita diabetes atau tidak dengan fitur umur (Age) = 60 dan tekanan darah (Blood Pressure) = 71 menggunakan metode gaussian naive bayes. Langkah-langkahnya adalah sebagai berikut.

1. Menghitung probabilitas prior masing masing kelas target.

   $$P(Y = 0) = \frac{N_{Y=0}}{N} = \frac{4}{8} = 0{,}5$$

   $$P(Y = 1) = \frac{N_{Y=1}}{N} = \frac{4}{8} = 0{,}5$$

2. Menghitung probabilitas likelihood untuk setiap fitur pada setiap kelas menggunakan distribusi Gaussian.

$$P(X_1 = 60|Y = 0) = \frac{1}{2\pi \cdot 3{,}08}\, e^{-\frac{(60-67)^2}{2 \cdot 3{,}08^2}} = 0{,}0097$$

$$P(X_2 = 71|Y = 0) = \frac{1}{2\pi \cdot 2{,}38}\, e^{-\frac{(71-50{,}75)^2}{2 \cdot 2{,}38^2}} = 3{,}194$$

$$P(X_1 = 60|Y = 1) = \frac{1}{2\pi \cdot 5{,}65}\, e^{-\frac{(60-70)^2}{2 \cdot 5{,}65^2}} = 0{,}0147$$

$$P(X_2 = 71|Y = 1) = \frac{1}{2\pi \cdot 5{,}14}\, e^{-\frac{(60-70)^2}{2 \cdot 5{,}14^2}} = 0{,}0719$$

3. Menghitung probabilitas posterior dari setiap kelas untuk data yang akan diklasifikasikan.

$$P(Y = 0|X = (60{,}71)) = 0{,}5 \cdot 0{,}0097 \cdot 3{,}194 = 0.015$$

$$P(Y = 1|X = (60{,}71)) = 0{,}5 \cdot 0{,}0147 \cdot 0{,}0719 = 0.00052$$

Karena probabilitas non-diabetes lebih tinggi, maka pasien tersebut diklasifikasikan sebagai non-diabetes.

3.11 Implementasi Metode Gaussian Naïve Bayes

Untuk melakukan implementasi klasifikasi menggunakan metode Gaussian Naive Bayes, Pertama, library scikit-learn (sklearn) diimpor, termasuk modul-modul yang diperlukan seperti train_test_split, GaussianNB, dan confusion_matrix. Kemudian, dataset yang tersedia dipisahkan menjadi dua komponen utama: fitur (X) dan target (y). Fitur-fitur dari dataset disimpan dalam variabel X, sedangkan target atau label kelas disimpan dalam variabel y.

Selanjutnya, dataset dibagi menjadi data latih dan data uji menggunakan fungsi train_test_split(). Data latih akan digunakan untuk melatih model, sedangkan data uji akan digunakan untuk melakukan evaluasi kinerja model yang

telah dilatih. Dalam penelitian ini, data uji sebesar 20% dari keseluruhan dataset, sedangkan data latih sebesar 80%.

Tabel 3.8 Data Train.

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	12	140	85	33	196,8	37,4	0,244	41	0
2	1	168	88	29	196	35	0,905	52	1
3	1	97	70	40	119,4	38,1	0,218	30	0
4	1	99	72	30	18	38,6	0,412	21	0
5	1	116	78	29	180	36,1	0,496	25	0
…	…	…	…	…	…	…	…	…	…
422	1	172	68	49	579	42,4	0,702	28	1
423	1	128	88	39	110	36,5	1,057	37	1
424	0	101	65	28	92,2	24,6	0,237	22	0
425	2	117	90	19	71	25,2	0,313	21	0
426	6	111	64	39	132,4	34,2	0,26	24	0
427	1	81	74	41	57	46,3	1,096	32	0

Tabel 3.9 Data Test

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	0	162	76	56	100	53,2	0,759	25	1
2	0	180	90	26	90	36,5	0,314	35	1
3	0	86	68	32	109,4	35,8	0,238	25	0
4	2	100	68	25	71	38,5	0,324	26	0
5	2	88	58	26	16	28,4	0,766	22	0
6	2	112	68	22	94	34,1	0,315	26	0
…	…	…	…	…	…	…	…	…	…
102	2	122	52	43	158	36,2	0,816	28	0
103	0	95	80	45	92	36,5	0,33	26	0
104	10	122	78	31	205,4	27,6	0,512	45	0
105	5	111	72	28	127	23,9	0,407	27	0
106	2	114	68	22	85,2	28,7	0,092	25	0
107	0	189	104	25	180,6	34,3	0,435	41	1

Fungsi train_test_split() akan secara acak membagi dataset menjadi dua subset yang disebut data pelatihan dan data uji. Subset data pelatihan digunakan

untuk melatih model klasifikasi, sedangkan subset data uji digunakan untuk menguji kinerja model yang telah dilatih.

Saat memanggil fungsi train_test_split(X, y, test_size, random_state), di mana X adalah matriks fitur yang berisi sampel-sampel dataset dan y adalah vektor yang berisi label atau kelas yang sesuai dengan setiap sampel, fungsi ini akan menghasilkan empat keluaran, yaitu X_train, X_test, y_train, dan y_test. X_train dan y_train akan berisi subset data pelatihan yang digunakan untuk melatih model, Sedangkan X_test dan y_test akan berisi subset data uji yang digunakan untuk menguji model.

Argumen test_size digunakan untuk mengatur proporsi data yang akan dialokasikan sebagai data uji. Misalnya, jika kita mengatur test_size menjadi 0.2, maka 20% dari data akan diambil sebagai data uji dan sisanya akan menjadi data pelatihan. Proporsi ini dapat disesuaikan sesuai dengan kebutuhan dan ukuran dataset yang digunakan.

Setelah berhasil membagi dataset menjadi data latih dan data uji selanjutnya langkah yang perlu dilakukan adalah menginisialisasi objek dari kelas gaussianNB. Setelah objek gaussianNB diinisialisasi, kita dapat menggunakan metode-metode yang tersedia pada objek tersebut untuk melatih dan menggunakan model Gaussian Naive Bayes.

Fungsi utama yang digunakan adalah fit(X,y). Metode ini digunakan untuk melatih model dengan menggunakan data pelatihan. Argumen X adalah matriks fitur yang berisi sampel-sampel pelatihan, sedangkan y adalah vektor yang berisi label atau kelas yang sesuai dengan setiap sampel. Selama proses pelatihan, model

Gaussian Naive Bayes akan mempelajari distribusi Gaussian dari setiap fitur dalam dataset. Model akan mengestimasi parameter-parameter yang diperlukan, seperti nilai rata-rata dan simpangan baku, untuk setiap kelas. Estimasi ini dilakukan berdasarkan pada data pelatihan yang diberikan. Fungsi fit(X, y) akan mengubah objek model sehingga mencerminkan model Gaussian Naive Bayes yang telah dilatih.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Pengujian

Dalam sub bab 3.11, dijelaskan bahwa semua model telah melewati proses atau tahap pelatihan menggunaan fungsi fit() pada objek model. Fungsi fit() ini bertanggung jawab untuk melatih model dengan mengeksplorasi pola dan relasi antara fitur-fitur dan label kelas dari data latih. Selama proses pelatihan, model secara efektif "mempelajari" informasi dari data latih untuk dapat membuat prediksi yang akurat di kemudian hari.

Setelah model selesai dilatih, langkah selanjutnya adalah melakukan prediksi kelas pada data uji yang telah ditentukan. Untuk melakukan ini, kita menggunakan fungsi predict(). Fungsi ini menerima argumen X, yang merupakan matriks fitur dari data yang ingin diprediksi. Dalam kasus model Gaussian Naive Bayes, model menggunakan distribusi Gaussian yang telah dipelajari selama proses pelatihan untuk menghitung probabilitas kelas yang mungkin untuk setiap sampel dalam X.

Dalam proses prediksi, model kemudian memilih kelas dengan probabilitas tertinggi sebagai prediksi untuk setiap sampel. Dengan kata lain, model memilih kelas yang memiliki probabilitas paling tinggi berdasarkan informasi yang telah dipelajari selama pelatihan. Dengan menggunakan distribusi Gaussian yang telah dipelajari, model dapat menghitung probabilitas kelas yang paling mungkin untuk setiap sampel dalam data uji.

Ketika fungsi predict(X) dijalankan, ia mengembalikan vektor hasil prediksi yang berisi prediksi kelas untuk setiap sampel dalam X. Dengan demikian, kita dapat melihat hasil prediksi dari model untuk setiap data uji yang telah ditentukan. Dengan menggunakan model yang telah dilatih dan fungsi predict(), kita dapat melakukan prediksi kelas dengan cepat dan efisien berdasarkan informasi yang telah dipelajari selama proses pelatihan. Berikut adalah hasil dari setiap tahap pengujian.

4.1.1 Hasil Pengujian Model A

Pengujian model A berdasarkan skenario pengujian yaitu digunakan semua fitur dengan melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikasi dua puluh baris pertama dari fungsi predict() untuk Model A. Untuk data lengkapnya bisa dilihat pada halaman lampiran.

Tabel 3.10 Model A : Hasil Klasifikasi Semua Fitur berdasarkan FS

Glucose	**Insulin**	**Age**	**BMI**	**Skin Thickness**	**Pregnancies**	**Diabetes Pedigree Function**	**Blood Pressure**	**Actual**	**Predicted**
162	100	25	53,2	56	0	0,759	76	1	1
180	90	35	36,5	26	0	0,314	90	1	1
86	109,4	25	35,8	32	0	0,238	68	0	0
100	71	26	38,5	25	2	0,324	68	0	0
88	16	22	28,4	26	2	0,766	58	0	0
112	94	26	34,1	22	2	0,315	68	0	0
84	115	28	36,9	23	1	0,471	64	0	0
99	86	24	25,6	19	3	0,154	54	0	0
127	155	28	34,5	11	4	0,598	88	0	0
129	125	43	38,5	49	7	0,439	68	1	1
139	160	25	31,6	35	5	0,361	80	1	0
122	200	26	35,9	27	2	0,483	76	0	0
174	194	36	32,9	22	3	0,593	58	1	1
61	53,2	46	34,4	28	3	0,243	82	0	0
108	278	22	25,3	10	2	0,881	62	0	0

4.1.2 Hasil Pengujian Model B

Pengujian model B berdasarkan skenario pengujian yaitu digunakan semua fitur tanpa melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikasi dua puluh baris pertama dari fungsi predict() untuk Model B. Untuk data lengkapnya bisa dilihat pada halaman lampiran.

Tabel 3.11 Model B : Hasil Klasifikasi Semua Fitur Acak

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Actual	Predicted
0	162	76	56	100	53,2	0,759	25	1	1
0	180	90	26	90	36,5	0,314	35	1	1
0	86	68	32	109,4	35,8	0,238	25	0	0
2	100	68	25	71	38,5	0,324	26	0	0
2	88	58	26	16	28,4	0,766	22	0	0
2	112	68	22	94	34,1	0,315	26	0	0
1	84	64	23	115	36,9	0,471	28	0	0
3	99	54	19	86	25,6	0,154	24	0	0
4	127	88	11	155	34,5	0,598	28	0	0
7	129	68	49	125	38,5	0,439	43	1	1
5	139	80	35	160	31,6	0,361	25	1	0
2	122	76	27	200	35,9	0,483	26	0	0
3	174	58	22	194	32,9	0,593	36	1	1
3	61	82	28	53,2	34,4	0,243	46	0	0
2	108	62	10	278	25,3	0,881	22	0	0
2	87	58	16	52	32,7	0,166	25	0	0
0	104	64	23	116	27,8	0,454	23	0	0
1	86	66	52	65	41,3	0,917	29	0	0
3	80	82	31	70	34,2	1,292	27	1	0

4.1.3 Hasil Pengujian Model C

Pengujian model C berdasarkan skenario pengujian yaitu digunakan lima fitur terbaik melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikasi lima belas baris pertama dari fungsi predict() untuk Model C. Untuk data lengkapnya bisa dilihat pada halaman lampiran

Tabel 3.12 Model C : Hasil Klasifikasi Lima Fitur Terbaik Hasil ANOVA FS

Glucose	SkinThickness	Insulin	BMI	Age	Actual	Predicted
162	56	100	53,2	25	1	1
180	26	90	36,5	35	1	1
86	32	109,4	35,8	25	0	0
100	25	71	38,5	26	0	0
88	26	16	28,4	22	0	0
112	22	94	34,1	26	0	0
84	23	115	36,9	28	0	0
99	19	86	25,6	24	0	0
127	11	155	34,5	28	0	0
129	49	125	38,5	43	1	1
139	35	160	31,6	25	1	0
122	27	200	35,9	26	0	0
174	22	194	32,9	36	1	1

4.1.4 Hasil Pengujian Model D

Pengujian model D berdasarkan skenario pengujian yaitu digunakan lima fitur acak tanpa melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikas lima belas baris pertama dari fungsi predict() untuk Model D. Untuk data lengkapnya bisa dilihat pada halaman lampiran.

Tabel 3.13 Model D : Hasil Klasifikasi Lima Fitur Acak

Pregnancies	Blood Pressure	Skin Thickness	BMI	Diabetes Pedigree Function	Actual	Predicted
0	76	56	53,2	0,759	1	1
0	90	26	36,5	0,314	1	0
0	68	32	35,8	0,238	0	0
2	68	25	38,5	0,324	0	0
2	58	26	28,4	0,766	0	0
2	68	22	34,1	0,315	0	0
1	64	23	36,9	0,471	0	0
3	54	19	25,6	0,154	0	0
4	88	11	34,5	0,598	0	0
7	68	49	38,5	0,439	1	1
5	80	35	31,6	0,361	1	0

4.1.5 Hasil Pengujian Model E

Pengujian model E berdasarkan skenario pengujian yaitu digunakan tiga fitur terbaik dengan melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikasi dua puluh baris pertama dari fungsi predict() untuk Model E. Untuk data lengkapnya bisa dilihat pada halaman lampiran.

Tabel 3.14 Model E : Hasil Klasifikasi Tiga Fitur Terbaik Hasil ANOVA FS

Glucose	Insulin	Age	Actual	Predicted
162	100	25	1	1
180	90	35	1	1
86	109,4	25	0	0
100	71	26	0	0
88	16	22	0	0
112	94	26	0	0
84	115	28	0	0
99	86	24	0	0
127	155	28	0	0
129	125	43	1	0
139	160	25	1	0
122	200	26	0	0
174	194	36	1	1
61	53,2	46	0	0
108	278	22	0	0
87	52	25	0	0
104	116	23	0	0
86	65	29	0	0
80	70	27	1	0
112	132,6	25	1	0

4.1.6 Hasil Pengujian Model F

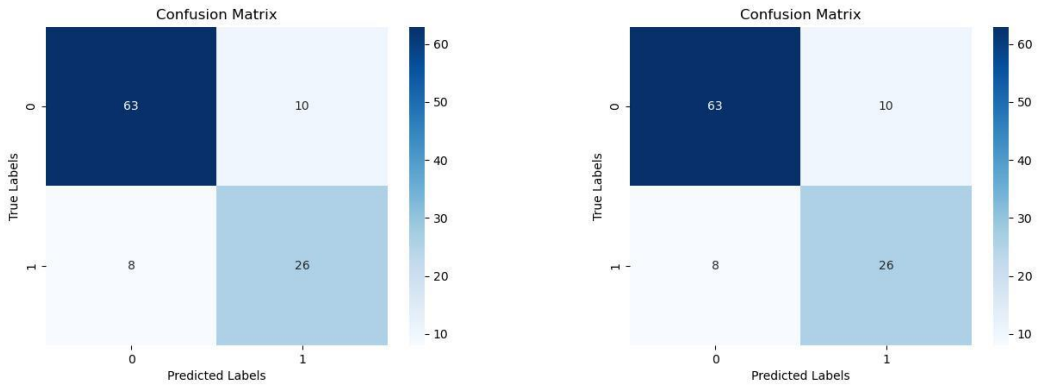
Pengujian model F berdasarkan skenario pengujian yaitu digunakan tiga fitur acak tanpa melalui proses seleksi fitur metode ANOVA. Berikut adalah hasil klasifikasi dua puluh baris pertama dari fungsi predict() untuk Model F. Untuk data lengkapnya bisa dilihat pada halaman lampiran.

Tabel 3.15 Model F : Hasil Klasifikasi Tiga Fitur Terbaik Hasil ANOVA FS

Pregnancies	BloodPressure	DiabetesPedigreeFunction	Actual	Predicted
0	76	0,759	1	0
0	90	0,314	1	0
0	68	0,238	0	0
2	68	0,324	0	0
2	58	0,766	0	0
2	68	0,315	0	0
1	64	0,471	0	0
3	54	0,154	0	0
4	88	0,598	0	0
7	68	0,439	1	0

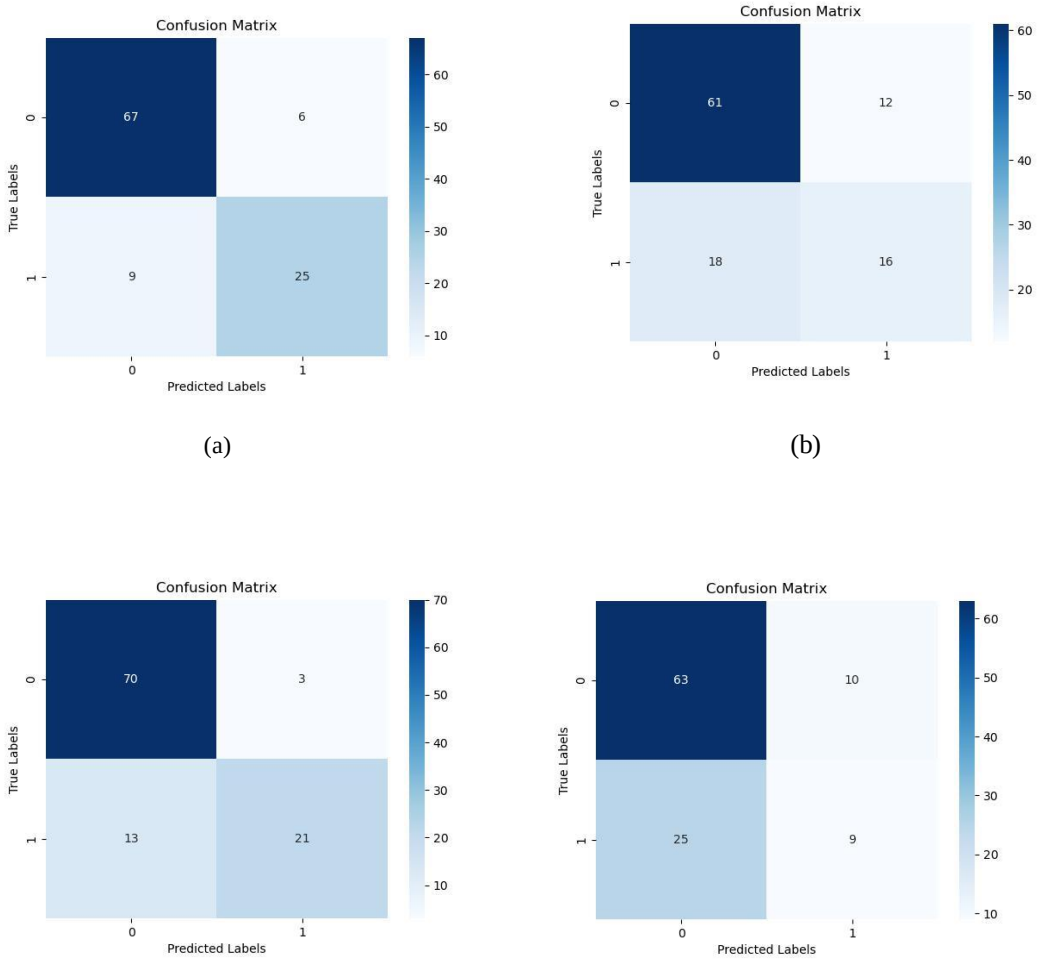
4.2 Confusion Matrix

Setelah memperoleh hasil prediksi dari seluruh model pengujian, langkah selanjutnya adalah membuat confusion matrix melalui penggunaan fungsi confusion_matrix(). Dengan confusion matrix dapat diambil informasi matriks evaluasi mencakup akurasi, presisi, recall, serta metrik evaluasi lainnya yang memiliki peran signifikan dalam mengevaluasi kinerja model klasifikasi. Berikut adalah hasil confussion matrix pada setiap model pengujian.



(a) (b)

Gambar 4.1 Confusion Matrix (a) skenario A, (b) skenario B



Gambar 4.1 Confusion Matrix (a) skenario C, (b) skenario D, (c) skenario E, (d) skenario F

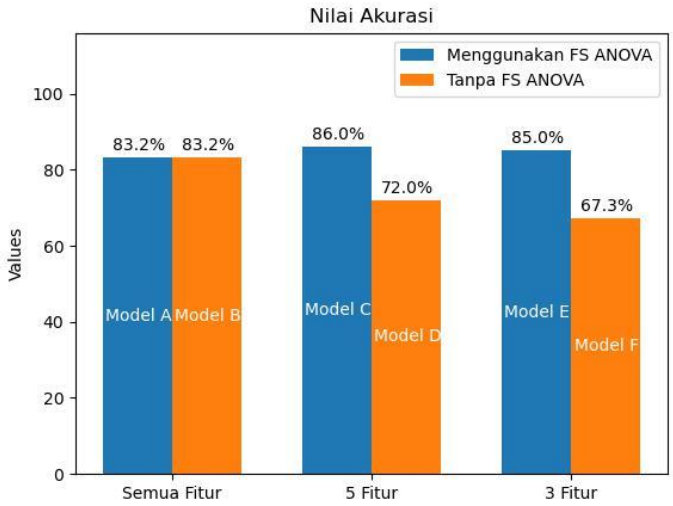
4.3 Evaluation Metrics Analysis

Setelah memahami confusion matrix dari setiap model pengujian, dapat menghitung berbagai matriks evaluasi yang memberikan wawasan lebih mendalam tentang kinerja model. Beberapa matriks evaluasi penting meliputi akurasi, presisi, recall, spesifisitas dan F1-score. Dengan mempertimbangkan berbagai matriks evaluasi ini, kita dapat menggambarkan dengan lebih

komprehensif sejauh mana model tersebut berhasil atau gagal dalam melakukan prediksi dan klasifikasi.

4.3.1 Akurasi

Akurasi digunakan untuk mengukur sejauh mana model klasifikasi dapat mengklasifikasikan keseluruhan kasus dengan benar. Akurasi memberikan gambaran umum tentang performa model dan seberapa baik model dapat mengenali penyakit pada dataset yang diamati. akurasi dihitung dengan membagi jumlah contoh yang diklasifikasikan dengan benar oleh model (true positive + true negative) dengan jumlah total contoh dalam dataset. Berikut adalah hasil nilai akurasi untuk setiap model yang diuji.



Gambar 4.7 Nilai Akurasi

Pada Model A menggunakan semua fitur yang tersedia dengan melakukan seleksi fitur metode Anova Feature Selection. Dalam hal akurasi, model ini mencapai nilai 83,2%, yang mengindikasikan kemampuan model untuk

melakukan klasifikasi dengan tingkat akurasi yang baik. Model B juga menggunakan semua fitur yang tersedia, tetapi kali ini tanpa melakukan seleksi fitur menggunakan metode Anova Feature Selection. Meskipun metode seleksi fitur tidak digunakan, akurasi tetap sama dengan Model A, yaitu 83,2%. Hal ini menunjukkan bahwa metode Anova Feature Selection tidak memberikan perbaikan dalam hal akurasi untuk model yang menggunakan semua fitur pada dataset.

Pada Model C melakukan seleksi fitur dengan menggunakan metode Anova Feature Selection dan memilih 5 fitur terbaik. Hasilnya, akurasi meningkat menjadi 86,0%. Ini menunjukkan bahwa memilih fitur terbaik dengan mempertimbangkan metode seleksi fitur dapat meningkatkan kinerja model dalam melakukan klasifikasi. Model D juga menggunakan 5 fitur, tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova Feature Selection. Dalam hal akurasi, model ini mencapai nilai 72,0%, yang lebih rendah dibandingkan dengan model-model sebelumnya. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa mempertimbangkan metode seleksi fitur menggunkaan Anova Feature Selection dapat menurunkan kinerja model.

Pada Model E melakukan seleksi fitur dengan menggunakan metode Anova Feature Selection dan memilih 3 fitur terbaik. Akurasi pada model ini adalah 85,0%, yang lebih rendah dibandingkan dengan Model C, tetapi lebih tinggi daripada Model A dan Model B. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat mempengaruhi kinerja model, namun jumlah fitur yang lebih sedikit tidak selalu menghasilkan performa yang lebih baik. Model F

juga menggunakan 3 fitur tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova Feature Selection. Akurasi pada model ini adalah 67,3%, yang merupakan nilai paling rendah di antara semua model yang telah dibangun. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat sangat merugikan kinerja model.

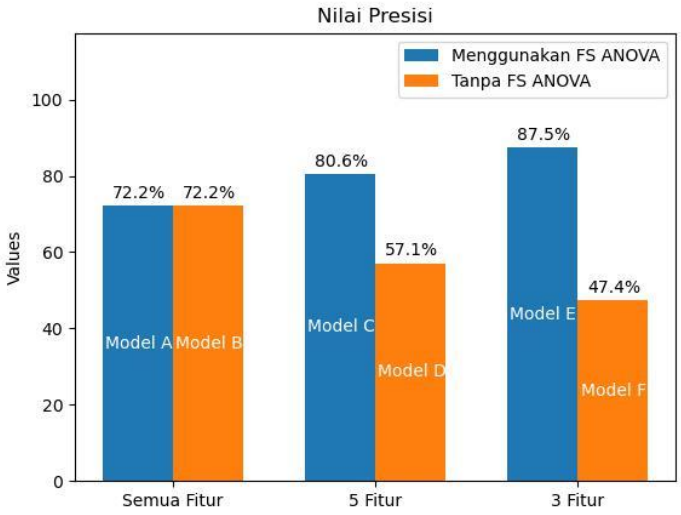
Berdasarkan uraian di atas, dapat disimpulkan bahwa pemilihan fitur yang tepat dapat mempengaruhi kinerja model klasifikasi Gaussian Naive Bayes. Model dengan semua fitur tanpa dan dengan seleksi fitur Anova Feature Selection (Model A dan Model B) memberikan akurasi yang cukup baik. Namun, hasil penelitian untuk matrik akurasi ini juga menunjukkan bahwa untuk Model C yang hanya menggunakan 5 fitur terbaik dengan mempertimbangkan metode Anova Feature Selection memberikan peningkatan akurasi yang baik. Ini menunjukkan bahwa memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan kinerja model.

Namun selain itu, mengurangi jumlah fitur secara acak tanpa menggunakan metode seleksi fitur (seperti pada Model D, dan Model F) dapat merugikan kinerja model.

4.3.2 Presisi

Presisi mengukur sejauh mana model dapat mengidentifikasi kasus positif secara akurat. Dalam klasifikasi penyakit, presisi memberikan informasi tentang seberapa banyak pasien yang diidentifikasi sebagai positif oleh model yang benar-benar menderita penyakit tersebut. Presisi yang tinggi menunjukkan kemampuan model dalam menghindari kesalahan dalam mengklasifikasikan pasien yang

sebenarnya tidak menderita penyakit, sehingga membantu dalam pengambilan keputusan yang lebih akurat dalam diagnosis dan perawatan pasien. Presisi dihitung dengan membagi jumlah contoh positif yang benar diidentifikasi oleh model (true positive) dengan jumlah total contoh yang diklasifikasikan sebagai positif oleh model (true positive + false positive). Berikut adalah hasil nilai presisi untuk setiap model yang diuji.



Gambar 4.8 Nilai Presisi

Pada Model A menggunakan semua fitur yang tersedia dengan melakukan seleksi fitur menggunakan metode Anova. Dalam hal presisi, model ini mencapai nilai 72,2%, yang mengindikasikan kemampuan model untuk mengklasifikasikan dengan tingkat kebenaran yang cukup baik. Model B juga menggunakan semua fitur yang tersedia, tetapi kali ini tanpa melakukan seleksi fitur menggunakan metode Anova. Meskipun metode seleksi fitur tidak digunakan, presisi tetap sama dengan Model A, yaitu 72,2%. Hal ini menunjukkan bahwa Anova Feature

Selection tidak memberikan perbaikan dalam hal presisi untuk dataset yang digunakan

Pada Model C memilih 5 fitur terbaik berdasarkan seleksi fitur dengan menggunakan metode Anova Feature Selection. Hasilnya, presisi meningkat menjadi 80,6%. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat meningkatkan kinerja model dalam melakukan klasifikasi dengan tingkat presisi yang lebih tinggi. Model D juga menggunakan 5 fitur yang dipilih secara acak tetapi tanpa melakukan seleksi fitur menggunakan metode Anova. Dalam hal presisi, model ini mencapai nilai 57,1%, yang lebih rendah dibandingkan dengan model-model sebelumnya (Model A, Model B, dan Model C). Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat merugikan kinerja model dalam melakukan klasifikasi dengan tingkat presisi yang rendah.

Pada Model E melakukan seleksi fitur dengan menggunakan metode Anova Feature Selection dan memilih 3 fitur terbaik. Presisi pada model ini adalah 87,5%, yang lebih tinggi dibandingkan dengan Model A, Model B, Model C. dan Model D. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat mempengaruhi kinerja model, dan memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan tingkat presisi. Model F juga menggunakan 3 fitur yang dipilih secara acak tetapi tanpa melakukan seleksi fitur menggunakan metode Anova. Presisi pada model ini adalah 47,4%, yang merupakan nilai paling rendah di antara semua model yang telah dibangun. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat

sangat merugikan kinerja model dalam melakukan klasifikasi dengan tingkat presisi yang rendah.

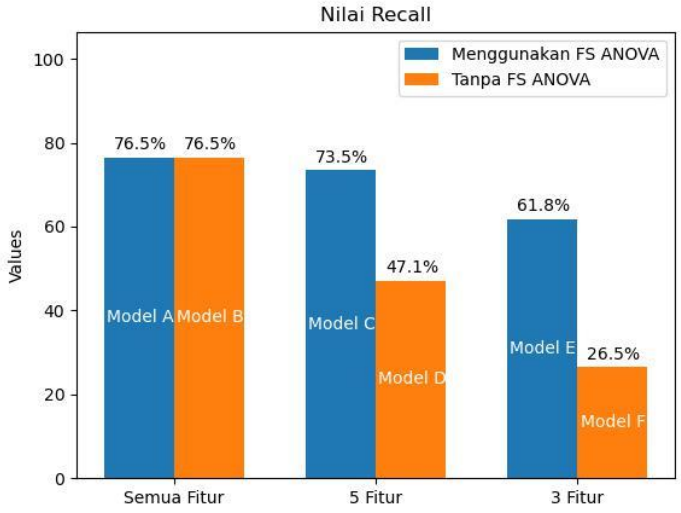
Berdasarkan dari uraian di atas, dapat disimpulkan bahwa pemilihan fitur yang tepat dapat mempengaruhi kinerja model klasifikasi Gaussian Naive Bayes. Model dengan semua fitur tanpa dan dengan seleksi fitur Anova (Model A dan Model B) memberikan tingkat presisi yang serupa. Namun, mengurangi jumlah fitur dan memilih fitur secara acak tanpa metode seleksi yang tepat (seperti pada Model D dan Model F) dapat mengurangi kinerja model dalam melakukan klasifikasi dengan tingkat presisi yang lebih rendah.

Selain itu, hasil penelitian untuk matrik presisi ini juga menunjukkan bahwa metode Anova Feature Selection memberikan peningkatan dalam hal presisi untuk Model C dan Model E. Ini menunjukkan bahwa memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan kinerja model dalam melakukan klasifikasi dengan tingkat presisi yang lebih tinggi.

4.3.3 Recall

Recall atau sensitivitas mengukur sejauh mana model mampu mengidentifikasi semua kasus positif yang sebenarnya. Dalam klasifikasi penyakit, recall memainkan peran penting dalam memastikan bahwa model dapat mendeteksi pasien yang sebenarnya menderita penyakit tersebut. Recall yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali pasien yang membutuhkan perawatan atau pengujian lebih lanjut. Hal ini dapat membantu dalam memperoleh diagnosis yang tepat dan menghindari

kejadian di mana pasien yang sebenarnya sakit diabaikan. recall dihitung dengan membagi jumlah contoh positif yang benar diidentifikasi oleh model (true positive) dengan jumlah total contoh positif yang sebenarnya ada dalam dataset (true positeve + false_negative). Berikut adalah hasil nilai recall untuk setiap model yang diuji.



Gambar 4.9 Nilai Recall

Pada Model A menggunakan semua fitur yang tersedia dengan melakukan seleksi fitur menggunakan metode Anova. Dalam hal recall, model ini mencapai nilai 76,5%, yang mengindikasikan kemampuan model untuk mengidentifikasi jumlah positif yang sebenarnya dengan akurasi yang cukup baik. Model B juga menggunakan semua fitur yang tersedia, tetapi kali ini tanpa dilakukan seleksi fitur menggunakan metode Anova. Meskipun metode seleksi fitur digunakan, recall tetap sama dengan Model A, yaitu 76,5%. Hal ini menunjukkan bahwa Anova Feature Selection tidak memberikan perbaikan dalam hal recall untuk model yang menggunakan semua fitur.

Pada Model C melakukan seleksi fitur dengan menggunakan metode Anova Feature Selection dan memilih 5 fitur terbaik. Hasilnya, recall menurun menjadi 73,5%. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan meskipun menggunakan metode seleksi fitur Anova Feature Selection dapat mengurangi kemampuan model dalam mengidentifikasi jumlah positif yang sebenarnya. Model D juga menggunakan 5 fitur yang dipilih secara acak tetapi tanpa melakukan seleksi fitur menggunakan metode Anova Feature Selection. Dalam hal recall, model ini mencapai nilai 47,1%, yang jauh lebih rendah dibandingkan dengan model-model sebelumnya. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat mengurangi kinerja model.

Pada Model E melakukan seleksi fitur dengan menggunakan metode Anova Feature Selection dan memilih 3 fitur terbaik. Recall pada model ini adalah 61,8%, yang lebih rendah dibandingkan dengan Model A, Model B, dan Model C tetapi lebih tinggi daripada Model D. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat mempengaruhi kinerja model, namun jumlah fitur yang lebih sedikit tidak selalu menghasilkan performa yang lebih baik. Model F juga menggunakan 3 fitur yang dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova. Recall pada model ini adalah 26,5%, yang merupakan nilai paling rendah di antara semua model yang telah dibangun. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat sangat merugikan kinerja model.

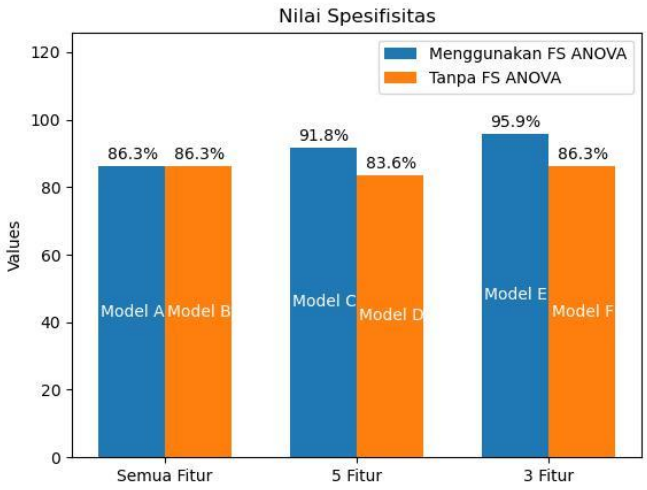
Berdasarkan uraian di atas, dapat disimpulkan bahwa pemilihan fitur yang tepat dapat mempengaruhi kinerja model klasifikasi Gaussian Naive Bayes. Model dengan semua fitur tanpa dan dengan seleksi fitur Anova (Model A dan Model B) memberikan recall yang cukup baik. Namun, membatasi jumlah fitur atau memilih fitur secara acak tanpa metode seleksi yang tepat (seperti pada Model D dan Model F) dapat mengurangi kinerja model.

Selain itu, hasil penelitian untuk matrik recall ini juga menunjukkan bahwa metode Anova Feature Selection tidak memberikan perbaikan yang signifikan dalam hal recall untuk dataset yang digunakan. Oleh karena itu, perlu mempertimbangkan metode seleksi fitur alternatif atau menggunakan pendekatan yang lebih canggih untuk meningkatkan kinerja model.

4.3.4 Spesifisitas

Spesifisitas mengukur sejauh mana model dapat mengidentifikasi kasus negatif dengan benar. Dalam klasifikasi penyakit, spesifisitas penting untuk memastikan bahwa model dapat mengklasifikasikan pasien yang sehat sebagai negatif dengan benar. Spesifisitas yang tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali pasien yang sehat dan menghindari kesalahan dalam mengklasifikasikan mereka sebagai positif. Ini membantu dalam mencegah diagnosis yang tidak perlu atau pengujian yang berlebihan pada pasien yang sebenarnya tidak memerlukan perhatian medis. Spesifisitas dihitung dengan membagi jumlah contoh negatif yang benar diidentifikasi oleh model (true negative) dengan jumlah total contoh negatif yang

sebenarnya ada dalam dataset (true negative + false positive). Berikut adalah hasil nilai spesifisitas untuk setiap model yang diuji.



Gambar 4.10 Nilai Spesifisitas

Pada Model A menggunakan semua fitur yang tersedia dengan melakukan seleksi fitur menggunakan metode Anova. Dalam hal spesifisitas, model ini mencapai nilai 86,3%, yang mengindikasikan kemampuan model untuk mengklasifikasikan dengan tingkat kebenaran yang baik untuk kelas negatif. Model B juga menggunakan semua fitur yang tersedia, tetapi kali ini tanpa dilakukan seleksi fitur menggunakan metode Anova. Meskipun metode seleksi fitur tidak digunakan, spesifisitas tetap sama dengan Model A, yaitu 86,3%. Hal ini menunjukkan bahwa Anova Feature Selection tidak memberikan perbaikan dalam hal spesifisitas untuk model yang menggunakan semua fitur dalam dataset yang digunakan.

Pada Model C melakukan seleksi fitur dengan menggunakan metode Anova dan memilih 5 fitur terbaik. Hasilnya, spesifisitas meningkat menjadi

91,8%. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat meningkatkan kinerja model dalam melakukan klasifikasi dengan tingkat spesifisitas yang lebih tinggi untuk kelas negatif. Model D juga menggunakan 5 fitur tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova. Dalam hal spesifisitas, model ini mencapai nilai 83,6%, yang lebih rendah dibandingkan dengan model-model sebelumnya. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat menurunkan kinerja model dalam melakukan klasifikasi dengan tingkat spesifisitas yang lebih rendah untuk kelas negatif.

Pada Model E melakukan seleksi fitur dengan menggunakan metode Anova dan memilih 3 fitur terbaik. Spesifisitas pada model ini adalah 95,9%, yang lebih tinggi dibandingkan dengan Model A, Model B, dan Model C. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat mempengaruhi kinerja model, dan memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan tingkat spesifisitas untuk kelas negatif. Model F juga menggunakan 3 fitur tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova. Spesifisitas pada model ini adalah 86,3%, yang sama dengan Model A dan Model B. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat tidak memberikan perbaikan dalam hal spesifisitas untuk kelas negatif.

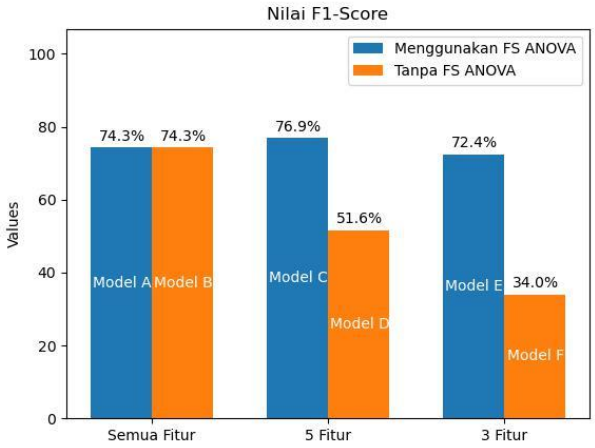
Berdasarkan hasil uraian diatas, dapat disimpulkan bahwa pemilihan fitur yang tepat dapat mempengaruhi kinerja model klasifikasi Gaussian Naive Bayes dalam hal spesifisitas untuk kelas negatif. Model dengan semua fitur tanpa seleksi

fitur Anova (Model A dan Model B) memberikan tingkat spesifisitas yang serupa. Namun, membatasi jumlah fitur (seperti pada Model C dan Model E) dapat meningkatkan kinerja model dalam melakukan klasifikasi dengan tingkat spesifisitas yang lebih tinggi untuk kelas negatif.

Selain itu, hasil penelitian untuk matrik spesifisitas ini juga menunjukkan bahwa metode Anova Feature Selection memberikan peningkatan dalam hal spesifisitas untuk Model C dan Model E. Ini menunjukkan bahwa memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan kinerja model dalam melakukan klasifikasi dengan tingkat spesifisitas yang lebih tinggi untuk kelas negatif.

4.3.5 F1-Score

F1-Score menggabungkan informasi tentang presisi dan recall dalam satu angka yang merepresentasikan performa keseluruhan model. Dalam klasifikasi penyakit, F1-Score memberikan gambaran yang seimbang tentang kemampuan model dalam mengenali pasien yang sebenarnya menderita penyakit (recall) dan memastikan bahwa pasien yang diidentifikasi memiliki penyakit adalah relevan (presisi). F1-Score sangat berguna ketika ada ketidakseimbangan antara jumlah pasien positif dan negatif dalam dataset, sehingga memberikan gambaran yang lebih komprehensif tentang performa model secara keseluruhan. Berikut adalah hasil F1-Score untuk setiap model pengujian.



Gambar 4.11 Nilai F1-Score

Pada Model A menggunakan semua fitur yang tersedia dengan melakukan seleksi fitur menggunakan metode Anova. Dalam hal f1-score, model ini mencapai nilai 74,3%, yang mengindikasikan kemampuan model dalam melakukan klasifikasi dengan tingkat kebenaran yang cukup baik secara keseluruhan untuk kelas positif dan negatif. Model B juga menggunakan semua fitur yang tersedia, tetapi kali tidak dilakukan seleksi fitur menggunakan metode Anova. Meskipun metode seleksi fitur tidak digunakan, f1-score tetap sama dengan Model A, yaitu 74,3%. Hal ini menunjukkan bahwa Anova Feature Selection tidak memberikan perbaikan dalam hal f1-score untuk dataset yang digunakan.

Pada Model C melakukan seleksi fitur dengan menggunakan metode Anova dan memilih 5 fitur terbaik. Hasilnya, f1-score meningkat menjadi 76,9%. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dan memilih fitur yang paling relevan menggunakan metode seleksi fitur dapat meningkatkan

kinerja model dalam melakukan klasifikasi secara keseluruhan. Model D juga menggunakan 5 fitur tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova. Dalam hal f1-score, model ini mencapai nilai 51,6%, yang lebih rendah dibandingkan dengan model-model sebelumnya. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat merugikan kinerja model dalam melakukan klasifikasi secara keseluruhan.

Pada Model E melakukan seleksi fitur dengan menggunakan metode Anova dan memilih 3 fitur terbaik. F1-score pada model ini adalah 72,4%, yang lebih rendah dibandingkan dengan Model A dan Model C tetapi lebih tinggi dibandingkan dengan model B dan D. Ini menunjukkan bahwa membatasi jumlah fitur yang digunakan dapat mempengaruhi kinerja model, dan memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan f1-score. Model F juga menggunakan 3 fitur tetapi dipilih secara acak tanpa melakukan seleksi fitur menggunakan metode Anova. F1-score pada model ini sangat rendah, yaitu 34,0%. Hal ini menunjukkan bahwa pemilihan fitur acak tanpa metode seleksi yang tepat dapat sangat merugikan kinerja model dalam melakukan klasifikasi secara keseluruhan.

Berdasarkan uraian di atas, dapat disimpulkan bahwa pemilihan fitur yang tepat dapat mempengaruhi kinerja model klasifikasi Gaussian Naive Bayes dalam hal f1-score secara keseluruhan. Model dengan semua fitur tanpa seleksi fitur Anova (Model A dan Model B) memberikan tingkat f1-score yang serupa. Namun, membatasi jumlah fitur seperti pada Model C dapat mempengaruhi

kinerja model dalam melakukan klasifikasi secara keseluruhan dengan f1-score yang lebih tinggi.

Selain itu, hasil penelitian untuk matrik f1-score ini juga menunjukkan bahwa metode Anova Feature Selection memberikan peningkatan dalam hal f1-score untuk Model C. Ini menunjukkan bahwa memilih fitur yang relevan menggunakan metode seleksi fitur dapat meningkatkan kinerja model dalam melakukan klasifikasi secara keseluruhan.

4.4 Pembahasan

Berdasarkan hasil evaluasi kinerja model yang telah dilakukan, dapat disimpulkan bahwa Model C dengan 5 fitur terbaik hasil Anova feature selection adalah model yang memiliki kinerja terbaik dalam hal akurasi, recall, dan f1-score. Dengan akurasi sebesar 86,0%, model ini mampu mengklasifikasikan data dengan tingkat kebenaran yang lebih tinggi dibandingkan model lainnya. Selain itu, dengan spesifisitas sebesar 91,8%, model ini juga mampu mengidentifikasi kelas negatif dengan akurasi yang tinggi.

Selanjutnya, model yang mendekati kinerja Model C adalah Model E, yang menggunakan 3 fitur terbaik hasil Anova feature selection. Model ini mencapai akurasi sebesar 85,0% dan nilai spesifisitas tertinggi sebesar 95,9%. Hal ini menunjukkan bahwa meskipun menggunakan jumlah fitur yang lebih sedikit, Model E mampu memberikan hasil yang kompetitif dan efektif dalam klasifikasi data.

Di sisi lain, Model D dan Model F, yang menggunakan fitur acak tanpa seleksi fitur metode Anova, menunjukkan kinerja yang lebih rendah dalam semua

matrik evaluasi. Model D mencapai akurasi sebesar 72,0% dan f1-score sebesar 51,6%, sedangkan Model F mencapai akurasi sebesar 67,3% dan f1-score sebesar 34,0%. Hasil ini menunjukkan bahwa pemilihan fitur yang acak tanpa mempertimbangkan metode seleksi fitur dapat mengurangi kinerja model klasifikasi.

Dalam hal persentase kenaikan performa, metode Anova feature selection, telah terbukti memberikan kenaikan yang signifikan dalam performa model klasifikasi Gaussian Naive Bayes. Dibandingkan dengan model tanpa seleksi fitur, model yang menggunakan seleksi fitur menunjukkan peningkatan terutama dalam akurasi, presisi, spesifisitas, dan f1-score. Penggunaan seleksi fitur dapat menghasilkan peningkatan performa antara 4% hingga 8% dalam beberapa kasus.

Sebagai contoh, jika kita membandingkan Model C yang menggunakan seleksi fitur dengan Model D yang tidak menggunakan seleksi fitur, terlihat bahwa Model C memberikan peningkatan akurasi sebesar 5%, presisi sebesar 8%, recall sebesar 4%, nilai f1-score sebesar 6%, dan spesifisitas sebesar 6%. Peningkatan ini menunjukkan bahwa seleksi fitur dengan Anova feature selection dapat membantu memperbaiki kemampuan model untuk mengklasifikasikan data dengan lebih baik dan mengidentifikasi pola yang relevan.

Ada beberapa faktor yang mempengaruhi peningkatan performa dalam penggunakan metode seleksi fitur. Misalnya, seleksi fitur membantu mengurangi dimensi fitur dalam model. Dalam banyak kasus, data yang digunakan memiliki jumlah fitur yang besar, tetapi tidak semua fitur tersebut berkontribusi secara signifikan terhadap pemisahan kelas target. Dengan menggunakan seleksi fitur,

kita dapat mengurangi kompleksitas model dan memori yang dibutuhkan, sehingga mengurangi risiko overfitting dan meningkatkan efisiensi komputasi.

Seleksi fitur membantu menghilangkan fitur-fitur yang tidak informatif. Terkadang, beberapa fitur dalam data tidak memiliki hubungan yang kuat dengan kelas target. Dengan menghilangkan fitur-fitur ini, kita dapat meningkatkan kemampuan model untuk menemukan pola dan hubungan yang lebih jelas antara fitur-fitur yang penting dan kelas target. Dengan demikian, model dapat membuat keputusan yang lebih baik dalam mengklasifikasikan data dengan meningkatkan akurasi dan meminimalkan kesalahan.

Selain itu, seleksi fitur juga membantu meningkatkan representasi fitur yang signifikan. Dengan memilih subset fitur yang paling informatif dan relevan, model dapat memperoleh representasi yang lebih baik dari fitur-fitur tersebut. Dalam hal ini, Anova feature selection memungkinkan model untuk menggambarkan variasi yang lebih besar dalam data, sehingga memungkinkan identifikasi pola-pola yang lebih kuat dan meningkatkan kemampuan model untuk mengklasifikasikan data dengan akurasi yang lebih tinggi.

Secara keseluruhan, penggunaan seleksi fitur, khususnya metode Anova feature selection dalam kasus ini, memiliki pengaruh yang positif terhadap performa model klasifikasi Gaussian Naive Bayes. Dengan memilih subset fitur yang paling informatif dan relevan, kita dapat meningkatkan akurasi, presisi, recall, spesifisitas, dan f1-score model. Seleksi fitur membantu menghilangkan noise dan mengoptimalkan representasi fitur, sehingga memungkinkan model

untuk membuat keputusan yang lebih baik dalam mengklasifikasikan data dan mengidentifikasi pola yang penting untuk pemisahan kelas target.

4.5 Integrasi Islam

Penyakit diabetes mellitus merupakan suatu kondisi kesehatan yang memerlukan pemahaman mendalam untuk pencegahan dan pengelolaan yang efektif. Dalam konteks ini, pengetahuan tentang faktor risiko, gejala, dan metode klasifikasi serta diagnosis memiliki peran krusial. Agama Islam, dengan ajarannya yang menyeluruh, memberikan pandangan yang mendukung pengetahuan sebagai sarana untuk meningkatkan kualitas hidup manusia.

Dalam konteks diabetes mellitus, pemahaman tentang penyakit ini dapat dianggap sebagai suatu bentuk ilmu pengetahuan yang dapat memberikan manfaat besar bagi umat manusia. Dengan mengetahui faktor risiko, seseorang dapat mengambil langkah-langkah pencegahan yang diperlukan untuk mengurangi kemungkinan terkena diabetes. Demikian pula, pemahaman tentang gejala dapat memungkinkan deteksi dini, memungkinkan pengelolaan yang lebih efektif.

Al-Quran dan hadis Rasulullah SAW juga menekankan nilai-nilai seperti kesehatan dan kesejahteraan, seperti pada surah Al-A'raf ayat 31:

□ الْمُسْرِفِيْنَ يُحِبُّ لَا اِنَّهٗ ۚ تُسْرِفُوْٓا وَلَا وَاشْرَبُوْا وَّكُلُوْا مَسْجِدٍ كُلِّ عِنْدَ زِيْنَتَكُمْ خُذُوْا اٰدَمَ يٰبَنِيْٓ

"Wahai anak cucu Adam! Pakailah pakaianmu yang bagus pada setiap (memasuki) masjid, makan dan minumlah, tetapi jangan berlebihan. Sungguh, Allah tidak menyukai orang yang berlebih-lebihan". (QS. Al-A'raf : 31).

Menurut tafsir Al-Munir ayat tersebut mengajarkan untuk memilih makanan yang halal, lezat, bergizi, dan bermanfaat bagi tubuh. Ketika datang ke

minuman, kita diperbolehkan meminum apa pun yang kita sukai selama tidak membuat mabuk atau merusak kesehatan. Namun, kita juga diingatkan untuk tidak berlebihan. Berlebihan dalam hal ini bisa berarti melebihi batas dalam hal kemewahan dalam makan, minum, dan berpakaian, atau melampaui batas yang diizinkan dari yang halal ke yang haram. Oleh karena itu, penting bagi kita untuk menemukan keseimbangan yang tepat, tidak terlalu pelit atau berlebihan dalam hal-hal tersebut (U. Rosyidah, & L. Mas'udah, 2022).

Dalam konteks tafsir di atas, perlu dicatat bahwa diabetes adalah salah satu kondisi kesehatan yang terkait dengan pola makan dan minum yang tidak seimbang. Diabetes adalah penyakit kronis yang ditandai oleh tingginya kadar gula darah dalam tubuh. Oleh karena itu, dalam mengamalkan ajaran tersebut, penting bagi setiap individu untuk memperhatikan jenis makanan dan minuman yang mereka konsumsi.

4.4.1 *Muamalah Ma'a Allah*

Penyakit adalah kenyataan yang tak terhindarkan dalam kehidupan manusia. Baik itu penyakit ringan maupun penyakit serius, setiap orang pasti pernah mengalami kondisi kesehatan yang memburuk. Dalam menghadapi berbagai penyakit ini, ayat ke-80 dalam Surah Asy-Syuara menawarkan keyakinan yang mendalam bahwa Allah adalah Penyembuh sejati yang memiliki kekuasaan untuk menyembuhkan penyakit apa pun. Surah Asy Syuara ayat 80:

وَاِذَا مَرِضْتُ فَهُوَ يَشْفِيْنِۙ

"Dan apabila aku sakit, Dialah yang menyembuhkan aku." (QS. Asy Syuara : 80).

Menurut Kemenag RI ayat ini menjelaskan bahwa Allah yang menyembuhkan manusia apabila ia sakit. Allah berkuasa menyembuhkan penyakit apa saja yang diderita seseorang. Meskipun begitu manusia juga harus mencari tahu cara untuk memperoleh kesembuhan itu (N. Khumaedah, 2020)

Ayat ini mengingatkan akan pentingnya menjaga kesehatan secara menyeluruh. Meskipun penyakit adalah bagian tak terpisahkan dari kehidupan, manusia juga memiliki tanggung jawab untuk menjaga kesehatannya sendiri. Ayat ini mengajarkan bahwa manusia harus berusaha menjaga keseimbangan, menerapkan gaya hidup sehat, dan mengambil tindakan pencegahan yang diperlukan. Dengan menjaga kesehatan, manusia dapat mengurangi risiko penyakit dan memperkuat ketahanan tubuh.

4.4.2 *Muamalah Ma'a an-Nas*

Meskipun Allah sebagai Penyembuh yang Mahakuasa, manusia juga memiliki tanggung jawab untuk menjaga kesehatan dan mengambil langkah pencegahan. Dalam usaha menjaga kesehatan, penting bagi manusia untuk menemukan keseimbangan antara kepercayaan pada Allah dan usaha yang dilakukan secara aktif. Meskipun Allah memiliki kekuatan untuk menyembuhkan, hal itu tidak berarti manusia dapat mengabaikan tanggung jawab mereka untuk berusaha semaksimal mungkin. Dengan menggabungkan keyakinan yang kuat pada Allah dengan tindakan konkret dalam menjaga kesehatan dan mengikuti pengobatan yang direkomendasikan, manusia dapat menghadapi penyakit dengan sikap yang seimbang dan optimis.

Sejalan dengan prinsip menjaga kesehatan dan mengambil tindakan pencegahan, terdapat pula sebuah hadis Rasulullah SAW:

اِغْتَنِمْ خَمْسًا قَبْلَ خَمْسٍ : شَبَابَكَ قَبْلَ هَرَمِكَ وَ صِحَّتَكَ قَبْلَ سَقَمِكَ وَ غِنَاكَ قَبْلَ فَقْرِكَ وَ فَرَاغَكَ قَبْلَ شُغْلِكَ وَ حَيَاتَكَ قَبْلَ مَوْتِكَ

"Manfaatkanlah lima perkara sebelum lima perkara, waktu mudamu sebelum datang waktu tuamu, waktu sehatmu sebelum waktu sakitmu, masa kayamu sebelum datang masa kefakiranmu, masa luangmu sebelum datang masa sibukmu, dan hidupmu sebelum datang matimu." (HR Al Hakim dalam Al Mustadrak-nya).

Hadis tersebut menekankan pentingnya menjaga kesehatan sebelum penyakit datang, serta memanfaatkan waktu dan sumber daya yang dimiliki dengan baik. Menurut (I. Heriani et al., 2020), hadis tersebut mengajarkan bahwa prinsip-prinsip kesehatan, kebersihan, dan kesucian merupakan syarat penting untuk mencapai kehidupan yang sejahtera di dunia dan kebahagiaan di akhirat.

Dalam penelitian ini, dilakukan analisis pengaruh seleksi fitur ANOVA terhadap performa model klasifikasi Gaussian Naïve Bayes. Tujuan dari penelitian ini adalah untuk menentukan subset fitur yang paling informatif dalam memprediksi penyakit diabetes, sehingga dapat membangun model klasifikasi yang lebih efektif. Dan diharapkan hasil penelitian ini dapat memberikan wawasan yang berharga dalam pencegahan dan penanganan penyakit diabetes mellitus, seperti yang ditekankan dalam hadis diatas yaitu pentingnya menjaga kesehatan sebelum penyakit datang.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Secara keseluruhan, hasil penelitian menunjukkan bahwa penggunaan ANOVA Feature Selection dapat meningkatkan kinerja model dalam beberapa kasus, terutama dalam hal akurasi, presisi, spesifisitas dan f1-score. Pemilihan fitur yang lebih baik dapat membantu model dalam mengklasifikasikan dengan lebih baik. Dengan akurasi sebesar 86,0%, model C yang hanya menggunakan lima fitur terbaik berdasarkan seleksi fitur ANOVA mampu mengklasifikasikan data dengan tingkat akurasi yang lebih tinggi dibandingkan model lainnya. Namun, perlu diperhatikan bahwa tidak semua model yang menggunakan fitur yang lebih sedikit memiliki kinerja yang lebih baik. Hal ini menunjukkan bahwa pemilihan fitur yang tepat sangat penting.

Selain itu, penggunaan fitur yang dipilih secara acak tanpa mempertimbangkan seleksi fitur menghasilkan penurunan kinerja model secara signifikan, terutama dalam hal akurasi, presisi, recall dan f1-score. Fitur-fitur acak tanpa mempertimbangkan metode seleksi fitur tidak memberikan informasi yang relevan dan sulit bagi model untuk mengklasifikasikan dengan benar.

5.2 Saran

Penting untuk dicatat bahwa pemilihan model terbaik harus dipertimbangkan berdasarkan konteks dan tujuan spesifik dari aplikasi atau penerapan dalam dunia nyata. Meskipun Model C dan Model E memiliki kinerja

yang lebih baik dalam penelitian ini, ada faktor lain yang perlu dipertimbangkan seperti kompleksitas model, waktu komputasi, dan biaya implementasi. Selain itu, hasil penelitian ini berlaku untuk dataset yang digunakan dalam penelitian tersebut. Ketika menerapkan model pada dataset yang berbeda, hasilnya bisa juga bervariasi. Oleh karena itu, disarankan untuk melakukan validasi tambahan dan evaluasi kinerja model pada dataset yang beragam untuk memastikan generalisasi yang baik.

DAFTAR PUSTAKA

Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima PIP. JATI (Jurnal Mahasiswa Teknik Informatika). 7(1). 452-457 https://doi.org/10.36040/jati.v7i1.6303

Nugroho, H., Ernawilis, E., Suheti, S., & Syamlan, S. F. (2023). Penyuluhan Kesehatan tentang Pengetahuan Pencegahan Diabetes Militus di Desa Rawat Rengas. Jurnal Peduli Masyarakat, 5(4), 1063-1070.

Saragih, M., Sipayung, R., & Pardede, J. A. (2023). Skrining Dan Pemeriksaan Kesehatan Pada Masyarakat Dengan Masalah Diabetes Mellitus Di Desa Kramat Gajah Deli Serdang. Jurnal Abdimas Mutiara, 4(2), 228-232.

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157-16173. https://doi.org/10.1007/s00521-022-07049-z

Shakeela, S., Shankar, N. S., Reddy, P. M., Tulasi, T. K., & Koneru, M. M. (2021). Optimal ensemble learning based on distinctive feature selection by univariate ANOVA-F statistics for IDS. International Journal of Electronics and Telecommunications, 267-275.

Rosyidah, U., & Mas'udah, L. (2022). LARANGAN BERLEBIH-LEBIHAN DALAM AL-QUR'AN. JADID: Journal of Quranic Studies and Islamic Communication, 2(1), 138-162.

Alassaf, M., & Qamar, A. M. (2022). Improving sentiment analysis of Arabic tweets by One-Way ANOVA. Journal of King Saud University-Computer and Information Sciences, 34(6), 2849-2859. https://doi.org/10.1016/j.jksuci.2020.10.023

Harjadinata, R. (2022). Analisis Sentimen Masyarakat Mengenai "Kebijakan Pemerintah Indonesia Dalam Menanggapi Covid-19" Dengan Menggunakan Metode Naïve Bayes Pada Media Sosial Facebook Dan Twitter. Repository Unja.

Id, I. D. (2021). Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python (Vol. 1). Unri Press.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Abedini, M., Bijari, A., & Banirostam, T. (2020). Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network. Int. J. Adv. Res. Comput. Commun. Eng, 9(7), 7-10.

Ilmi, A. F., Utari, D. M. (2020). Hubungan Lingkar Pinggang Dan Rasio Lingkar Pinggang-Panggul (RLPP) Terhadap Kadar Gula Puasa pada Mahasiswa Prodi Kesehatan Masyarakat STIKes Kharisma Persada. Journal of Nutrition College. 9(3). 223-226. https://doi.org/10.14710/jnc.v9i3.27658

Rezaeian, N., & Novikova, G. (2020). Persian text classification using naive bayes algorithms and support vector machine algorithm. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), 178-188.

Khumaedah, N. (2020). Pengaruh Intervensi Murotal Al Qur'an (Surat Asy-Syu'ara) Terhadap Tekanan Darah Pada Penderita Hipertensi Stage 1 Di Wilayah Kerja Puskesmas Wanadadi I.

Mazlan, A. D. N., Hairuddin, M. A., Md Tahir, N., Khirul Ashar, N. D., & Jusoh, M. H. (2020). Comparative analysis of PCA and ANOVA for assessing the subset feature selection of the geomagnetic Disturbance Storm Time. Journal of Electrical and Electronic Systems Research (JEESR), 17, 8-16.

Heriani, I., Hamid, A., Megasari, I. D., & Munajah, M. (2020). Konsep Kesehatan Lingkungan dalam Hukum Kesehatan dan Perspektif Hukum Islam. Prosiding Penelitian Dosen UNISKA MAB.

Tantawi, R. (2019). Menyembuhkan Penyakit dengan Obat dan Doa. Universitas Medan Area.

Brownlee, J. (2019). How to choose a feature selection method for machine learning. Machine Learning Mastery, 10.

Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and information technologies, 19(1), 3-26.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge managementprocess,5(2),1.

LAMPIRAN

Lampiran 1 : Hasil Klasifikasi GNB Model A

Glucose	Insulin	Age	BMI	Skin Thickness	Pregnancies	Diabetes Pedigree Function	Blood Pressure	Actual	Predicted
162	100	25	53,2	56	0	0,759	76	1	1
180	90	35	36,5	26	0	0,314	90	1	1
86	109,4	25	35,8	32	0	0,238	68	0	0
100	71	26	38,5	25	2	0,324	68	0	0
88	16	22	28,4	26	2	0,766	58	0	0
112	94	26	34,1	22	2	0,315	68	0	0
84	115	28	36,9	23	1	0,471	64	0	0
99	86	24	25,6	19	3	0,154	54	0	0
127	155	28	34,5	11	4	0,598	88	0	0
129	125	43	38,5	49	7	0,439	68	1	1
139	160	25	31,6	35	5	0,361	80	1	0
122	200	26	35,9	27	2	0,483	76	0	0
174	194	36	32,9	22	3	0,593	58	1	1
61	53,2	46	34,4	28	3	0,243	82	0	0
108	278	22	25,3	10	2	0,881	62	0	0
87	52	25	32,7	16	2	0,166	58	0	0
104	116	23	27,8	23	0	0,454	64	0	0
86	65	29	41,3	52	1	0,917	66	0	0
80	70	27	34,2	31	3	1,292	82	1	0
112	132,6	25	31,6	30	3	0,197	74	1	0
162	357,4	26	49,6	36	0	0,364	76	1	1
104,2	23	21	27,7	20	1	0,299	74	0	0
107	74	25	36,6	30	0	0,757	62	1	0
80	60	22	30	11	1	0,527	74	0	0
107	117,6	23	26,4	25	0	0,133	60	0	0
83	66	24	36,8	28	2	0,629	65	0	0
187	392	34	33,9	33	7	0,826	50	1	1
173	265	58	46,5	32	0	1,159	78	0	1
137	148	21	24,8	14	0	0,143	68	0	0
198	274	28	41,3	32	0	0,502	66	1	1
95	105	22	44,6	39	0	0,366	64	0	0
149	127	42	29,3	29	1	0,349	68	1	0
93	72	35	43,4	39	0	1,021	100	0	1
179	230,6	60	34,2	31	7	0,164	95	0	1
103	82	22	19,4	11	1	0,491	80	0	0
100	50	21	30,8	26	0	0,597	70	0	0
133	155	37	32,4	15	7	0,262	88	0	0
121	165	33	34,3	30	0	0,203	66	1	0
90	59	25	25,1	18	1	1,268	62	0	0

106	148	22	39,4	37	0	0,605	70	0	0
71	45,2	22	28	27	2	0,586	70	0	0
129	122	39	35,9	28	10	0,28	76	0	1
126	75	39	25,9	38	8	0,162	74	0	0
177	478	21	34,6	29	0	1,072	60	1	1
154	140	27	46,1	41	6	0,571	78	0	1
100	105	24	37,8	28	2	0,498	54	0	0
119	170	26	45,3	41	1	0,507	88	0	0
141	128	24	25,4	34	2	0,699	58	0	0
79	37	22	25,4	25	1	0,583	80	0	0
181	180	38	34,1	30	1	0,328	64	1	1
127	335	22	34,4	21	2	0,176	46	0	0
137	168	33	43,1	35	0	2,288	40	1	1
110	125	27	32,4	29	2	0,698	74	0	0
81	54,8	25	27,7	22	2	0,29	60	0	0
103	135,2	42	46,2	40	11	0,126	68	0	1
151	330,4	21	42,1	46	0	0,371	90	1	1
144	178,2	37	38,5	32	4	0,554	82	1	1
99	94	21	24,6	15	2	0,637	52	0	0
187	304	41	37,7	39	7	0,254	68	1	1
189	846	59	30,1	23	1	0,398	60	1	1
93	92	22	28,7	25	0	0,532	60	0	0
107	85	24	26,5	19	1	0,165	68	0	0
147	250,4	27	49,3	41	1	0,358	94	1	1
87	32	22	34,6	27	1	0,101	78	0	0
98	120	22	34,7	17	2	0,198	60	0	0
184	169	49	30	15	9	1,213	85	1	1
126	215	24	30,7	29	0	0,52	84	0	0
118	89	21	31,64	23	0	1,731	64	0	1
113	85,2	22	22,4	13	3	0,14	44	0	0
110	100	27	28,4	20	4	0,118	76	0	0
144	285	58	32	26	5	0,452	82	1	1
94	115	21	43,5	27	0	0,347	70	0	0
109	135	23	25,2	21	1	0,833	56	0	0
116	105	24	26,3	15	3	0,107	74	0	0
90	62,4	22	27,3	17	2	0,085	70	0	0
111	150	45	46,8	40	11	0,925	84	1	1
99	54	32	26,9	19	6	0,497	60	0	0
111	75,4	23	30,1	19	1	0,143	86	0	0
95	73	36	25,9	21	1	0,673	74	0	0
114	110	31	23,8	17	7	0,466	76	0	0
129	130	26	67,1	46	0	0,319	110	1	1
142	190	61	28,8	33	7	0,687	60	0	1

144	140	37	29,5	28	4	0,287	58	0	0
168	321	40	38,2	42	7	0,787	88	1	1
98	84	22	25,2	15	0	0,299	82	0	0
137	108,2	39	32	41	7	0,391	90	0	1
106	119	34	30,5	35	2	1,4	64	0	0
102	136,6	46	32,9	37	9	0,665	76	1	1
99	139,8	32	29	27	5	0,203	74	0	0
136	130	42	28,3	35	11	0,26	84	1	1
104	156	41	29,9	18	6	0,722	74	1	0
117	106	27	33,8	23	1	0,466	60	0	0
108	96,2	32	27,3	20	0	0,787	68	0	0
186	225	37	34,5	35	8	0,423	90	1	1
95	38	25	19,6	13	1	0,334	66	0	0
173	465	25	38,4	48	3	2,137	82	1	1
112	160	28	38,4	42	2	0,246	86	0	0
189	325	29	31,2	33	5	0,583	64	1	1
112	132	24	34,8	45	1	0,217	80	0	0
108	178	24	35,5	46	1	0,415	60	0	0
90	113,4	27	38,2	42	2	0,503	68	1	0
122	158	28	36,2	43	2	0,816	52	0	0
95	92	26	36,5	45	0	0,33	80	0	0
122	205,4	45	27,6	31	10	0,512	78	0	1
111	127	27	23,9	28	5	0,407	72	0	0
114	85,2	25	28,7	22	2	0,092	68	0	0
189	180,6	41	34,3	25	0	0,435	104	1	1

Lampiran 2 : Hasil Klasifikasi GNB Model B

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Actual	Predicted
0	162	76	56	100	53,2	0,759	25	1	1
0	180	90	26	90	36,5	0,314	35	1	1
0	86	68	32	109,4	35,8	0,238	25	0	0
2	100	68	25	71	38,5	0,324	26	0	0
2	88	58	26	16	28,4	0,766	22	0	0
2	112	68	22	94	34,1	0,315	26	0	0
1	84	64	23	115	36,9	0,471	28	0	0
3	99	54	19	86	25,6	0,154	24	0	0
4	127	88	11	155	34,5	0,598	28	0	0
7	129	68	49	125	38,5	0,439	43	1	1
5	139	80	35	160	31,6	0,361	25	1	0
2	122	76	27	200	35,9	0,483	26	0	0

3	174	58	22	194	32,9	0,593	36	1	1
3	61	82	28	53,2	34,4	0,243	46	0	0
2	108	62	10	278	25,3	0,881	22	0	0
2	87	58	16	52	32,7	0,166	25	0	0
0	104	64	23	116	27,8	0,454	23	0	0
1	86	66	52	65	41,3	0,917	29	0	0
3	80	82	31	70	34,2	1,292	27	1	0
3	112	74	30	132,6	31,6	0,197	25	1	0
0	162	76	36	357,4	49,6	0,364	26	1	1
1	104,2	74	20	23	27,7	0,299	21	0	0
0	107	62	30	74	36,6	0,757	25	1	0
1	80	74	11	60	30	0,527	22	0	0
0	107	60	25	117,6	26,4	0,133	23	0	0
2	83	65	28	66	36,8	0,629	24	0	0
7	187	50	33	392	33,9	0,826	34	1	1
0	173	78	32	265	46,5	1,159	58	0	1
0	137	68	14	148	24,8	0,143	21	0	0
0	198	66	32	274	41,3	0,502	28	1	1
0	95	64	39	105	44,6	0,366	22	0	0
1	149	68	29	127	29,3	0,349	42	1	0
0	93	100	39	72	43,4	1,021	35	0	1
7	179	95	31	230,6	34,2	0,164	60	0	1
1	103	80	11	82	19,4	0,491	22	0	0
0	100	70	26	50	30,8	0,597	21	0	0
7	133	88	15	155	32,4	0,262	37	0	0
0	121	66	30	165	34,3	0,203	33	1	0
1	90	62	18	59	25,1	1,268	25	0	0
0	106	70	37	148	39,4	0,605	22	0	0
2	71	70	27	45,2	28	0,586	22	0	0
10	129	76	28	122	35,9	0,28	39	0	1
8	126	74	38	75	25,9	0,162	39	0	0
0	177	60	29	478	34,6	1,072	21	1	1
6	154	78	41	140	46,1	0,571	27	0	1
2	100	54	28	105	37,8	0,498	24	0	0
1	119	88	41	170	45,3	0,507	26	0	0
2	141	58	34	128	25,4	0,699	24	0	0
1	79	80	25	37	25,4	0,583	22	0	0
1	181	64	30	180	34,1	0,328	38	1	1
2	127	46	21	335	34,4	0,176	22	0	0
0	137	40	35	168	43,1	2,288	33	1	1
2	110	74	29	125	32,4	0,698	27	0	0
2	81	60	22	54,8	27,7	0,29	25	0	0
11	103	68	40	135,2	46,2	0,126	42	0	1

0	151	90	46	330,4	42,1	0,371	21	1	1
4	144	82	32	178,2	38,5	0,554	37	1	1
2	99	52	15	94	24,6	0,637	21	0	0
7	187	68	39	304	37,7	0,254	41	1	1
1	189	60	23	846	30,1	0,398	59	1	1
0	93	60	25	92	28,7	0,532	22	0	0
1	107	68	19	85	26,5	0,165	24	0	0
1	147	94	41	250,4	49,3	0,358	27	1	1
1	87	78	27	32	34,6	0,101	22	0	0
2	98	60	17	120	34,7	0,198	22	0	0
9	184	85	15	169	30	1,213	49	1	1
0	126	84	29	215	30,7	0,52	24	0	0
0	118	64	23	89	31,64	1,731	21	0	1
3	113	44	13	85,2	22,4	0,14	22	0	0
4	110	76	20	100	28,4	0,118	27	0	0
5	144	82	26	285	32	0,452	58	1	1
0	94	70	27	115	43,5	0,347	21	0	0
1	109	56	21	135	25,2	0,833	23	0	0
3	116	74	15	105	26,3	0,107	24	0	0
2	90	70	17	62,4	27,3	0,085	22	0	0
11	111	84	40	150	46,8	0,925	45	1	1
6	99	60	19	54	26,9	0,497	32	0	0
1	111	86	19	75,4	30,1	0,143	23	0	0
1	95	74	21	73	25,9	0,673	36	0	0
7	114	76	17	110	23,8	0,466	31	0	0
0	129	110	46	130	67,1	0,319	26	1	1
7	142	60	33	190	28,8	0,687	61	0	1
4	144	58	28	140	29,5	0,287	37	0	0
7	168	88	42	321	38,2	0,787	40	1	1
0	98	82	15	84	25,2	0,299	22	0	0
7	137	90	41	108,2	32	0,391	39	0	1
2	106	64	35	119	30,5	1,4	34	0	0
9	102	76	37	136,6	32,9	0,665	46	1	1
5	99	74	27	139,8	29	0,203	32	0	0
11	136	84	35	130	28,3	0,26	42	1	1
6	104	74	18	156	29,9	0,722	41	1	0
1	117	60	23	106	33,8	0,466	27	0	0
0	108	68	20	96,2	27,3	0,787	32	0	0
8	186	90	35	225	34,5	0,423	37	1	1
1	95	66	13	38	19,6	0,334	25	0	0
3	173	82	48	465	38,4	2,137	25	1	1
2	112	86	42	160	38,4	0,246	28	0	0
5	189	64	33	325	31,2	0,583	29	1	1

1	112	80	45	132	34,8	0,217	24	0	0
1	108	60	46	178	35,5	0,415	24	0	0
2	90	68	42	113,4	38,2	0,503	27	1	0
2	122	52	43	158	36,2	0,816	28	0	0
0	95	80	45	92	36,5	0,33	26	0	0
10	122	78	31	205,4	27,6	0,512	45	0	1
5	111	72	28	127	23,9	0,407	27	0	0
2	114	68	22	85,2	28,7	0,092	25	0	0
0	189	104	25	180,6	34,3	0,435	41	1	1

Lampiran 3 : Hasil Klasifikasi GNB Model C

Glucose	Skin Thickness	Insulin	BMI	Age	Actual	Predicted
162	56	100	53,2	25	1	1
180	26	90	36,5	35	1	1
86	32	109,4	35,8	25	0	0
100	25	71	38,5	26	0	0
88	26	16	28,4	22	0	0
112	22	94	34,1	26	0	0
84	23	115	36,9	28	0	0
99	19	86	25,6	24	0	0
127	11	155	34,5	28	0	0
129	49	125	38,5	43	1	1
139	35	160	31,6	25	1	0
122	27	200	35,9	26	0	0
174	22	194	32,9	36	1	1
61	28	53,2	34,4	46	0	0
108	10	278	25,3	22	0	0
87	16	52	32,7	25	0	0
104	23	116	27,8	23	0	0
86	52	65	41,3	29	0	0
80	31	70	34,2	27	1	0
112	30	132,6	31,6	25	1	0
162	36	357,4	49,6	26	1	1
104,2	20	23	27,7	21	0	0
107	30	74	36,6	25	1	0
80	11	60	30	22	0	0
107	25	117,6	26,4	23	0	0
83	28	66	36,8	24	0	0
187	33	392	33,9	34	1	1
173	32	265	46,5	58	0	1

Glucose	ST	Insulin	BMI	Ag	A	P
103	40	135,2	46,2	42	0	1
151	46	330,4	42,1	21	1	1
144	32	178,2	38,5	37	1	1
99	15	94	24,6	21	0	0
187	39	304	37,7	41	1	1
189	23	846	30,1	59	1	1
93	25	92	28,7	22	0	0
107	19	85	26,5	24	0	0
147	41	250,4	49,3	27	1	1
87	27	32	34,6	22	0	0
98	17	120	34,7	22	0	0
184	15	169	30	49	1	1
126	29	215	30,7	24	0	0
118	23	89	31,64	21	0	0
113	13	85,2	22,4	22	0	0
110	20	100	28,4	27	0	0
144	26	285	32	58	1	1
94	27	115	43,5	21	0	0
109	21	135	25,2	23	0	0
116	15	105	26,3	24	0	0
90	17	62,4	27,3	22	0	0
111	40	150	46,8	45	1	1
99	19	54	26,9	32	0	0
111	19	75,4	30,1	23	0	0
95	21	73	25,9	36	0	0
114	17	110	23,8	31	0	0
129	46	130	67,1	26	1	1
142	33	190	28,8	61	0	1

137	14	148	24,8	21	0	0
198	32	274	41,3	28	1	1
95	39	105	44,6	22	0	0
149	29	127	29,3	42	1	1
93	39	72	43,4	35	0	0
179	31	230,6	34,2	60	0	1
103	11	82	19,4	22	0	0
100	26	50	30,8	21	0	0
133	15	155	32,4	37	0	0
121	30	165	34,3	33	1	0
90	18	59	25,1	25	0	0
106	37	148	39,4	22	0	0
71	27	45,2	28	22	0	0
129	28	122	35,9	39	0	0
126	38	75	25,9	39	0	0
177	29	478	34,6	21	1	1
154	41	140	46,1	27	0	1
100	28	105	37,8	24	0	0
119	41	170	45,3	26	0	0
141	34	128	25,4	24	0	0
79	25	37	25,4	22	0	0
181	30	180	34,1	38	1	1
127	21	335	34,4	22	0	0
137	35	168	43,1	33	1	1
110	29	125	32,4	27	0	0

144	28	140	29,5	37	0	0
168	42	321	38,2	40	1	1
98	15	84	25,2	22	0	0
137	41	108,2	32	39	0	1
106	35	119	30,5	34	0	0
102	37	136,6	32,9	46	1	0
99	27	139,8	29	32	0	0
136	35	130	28,3	42	1	0
104	18	156	29,9	41	1	0
117	23	106	33,8	27	0	0
108	20	96,2	27,3	32	0	0
186	35	225	34,5	37	1	1
95	13	38	19,6	25	0	0
173	48	465	38,4	25	1	1
112	42	160	38,4	28	0	0
189	33	325	31,2	29	1	1
112	45	132	34,8	24	0	0
108	46	178	35,5	24	0	0
90	42	113,4	38,2	27	1	0
122	43	158	36,2	28	0	0
95	45	92	36,5	26	0	0
122	31	205,4	27,6	45	0	0
111	28	127	23,9	27	0	0
114	22	85,2	28,7	25	0	0
189	25	180,6	34,3	41	1	1

Lampiran 4 : Hasil Klasifikasi GNB Model D

Pregnancies	**Blood Pressure**	**Skin Thickness**	**BMI**	**DPF**	**Actual**	**Predicted**
0	76	56	53,2	0,759	1	1
0	90	26	36,5	0,314	1	0
0	68	32	35,8	0,238	0	0
2	68	25	38,5	0,324	0	0
2	58	26	28,4	0,766	0	0
2	68	22	34,1	0,315	0	0
1	64	23	36,9	0,471	0	0
3	54	19	25,6	0,154	0	0
4	88	11	34,5	0,598	0	0
7	68	49	38,5	0,439	1	1
5	80	35	31,6	0,361	1	0
2	76	27	35,9	0,483	0	0

P	**BP**	**ST**	**BMI**	**DPF**	**A**	**P**
11	68	40	46,2	0,126	0	1
0	90	46	42,1	0,371	1	1
4	82	32	38,5	0,554	1	0
2	52	15	24,6	0,637	0	0
7	68	39	37,7	0,254	1	1
1	60	23	30,1	0,398	1	0
0	60	25	28,7	0,532	0	0
1	68	19	26,5	0,165	0	0
1	94	41	49,3	0,358	1	1
1	78	27	34,6	0,101	0	0
2	60	17	34,7	0,198	0	0
9	85	15	30	1,213	1	1

3	58	22	32,9	0,593	1	0
3	82	28	34,4	0,243	0	0
2	62	10	25,3	0,881	0	0
2	58	16	32,7	0,166	0	0
0	64	23	27,8	0,454	0	0
1	66	52	41,3	0,917	0	1
3	82	31	34,2	1,292	1	1
3	74	30	31,6	0,197	1	0
0	76	36	49,6	0,364	1	1
1	74	20	27,7	0,299	0	0
0	62	30	36,6	0,757	1	0
1	74	11	30	0,527	0	0
0	60	25	26,4	0,133	0	0
2	65	28	36,8	0,629	0	0
7	50	33	33,9	0,826	1	0
0	78	32	46,5	1,159	0	1
0	68	14	24,8	0,143	0	0
0	66	32	41,3	0,502	1	0
0	64	39	44,6	0,366	0	0
1	68	29	29,3	0,349	1	0
0	100	39	43,4	1,021	0	1
7	95	31	34,2	0,164	0	1
1	80	11	19,4	0,491	0	0
0	70	26	30,8	0,597	0	0
7	88	15	32,4	0,262	0	0
0	66	30	34,3	0,203	1	0
1	62	18	25,1	1,268	0	0
0	70	37	39,4	0,605	0	0
2	70	27	28	0,586	0	0
10	76	28	35,9	0,28	0	1
8	74	38	25,9	0,162	0	0
0	60	29	34,6	1,072	1	0
6	78	41	46,1	0,571	0	1
2	54	28	37,8	0,498	0	0
1	88	41	45,3	0,507	0	1
2	58	34	25,4	0,699	0	0
1	80	25	25,4	0,583	0	0
1	64	30	34,1	0,328	1	0
2	46	21	34,4	0,176	0	0
0	40	35	43,1	2,288	1	1
2	74	29	32,4	0,698	0	0

0	84	29	30,7	0,52	0	0
0	64	23	31,64	1,731	0	1
3	44	13	22,4	0,14	0	0
4	76	20	28,4	0,118	0	0
5	82	26	32	0,452	1	0
0	70	27	43,5	0,347	0	0
1	56	21	25,2	0,833	0	0
3	74	15	26,3	0,107	0	0
2	70	17	27,3	0,085	0	0
11	84	40	46,8	0,925	1	1
6	60	19	26,9	0,497	0	0
1	86	19	30,1	0,143	0	0
1	74	21	25,9	0,673	0	0
7	76	17	23,8	0,466	0	0
0	110	46	67,1	0,319	1	1
7	60	33	28,8	0,687	0	0
4	58	28	29,5	0,287	0	0
7	88	42	38,2	0,787	1	1
0	82	15	25,2	0,299	0	0
7	90	41	32	0,391	0	1
2	64	35	30,5	1,4	0	1
9	76	37	32,9	0,665	1	1
5	74	27	29	0,203	0	0
11	84	35	28,3	0,26	1	1
6	74	18	29,9	0,722	1	0
1	60	23	33,8	0,466	0	0
0	68	20	27,3	0,787	0	0
8	90	35	34,5	0,423	1	1
1	66	13	19,6	0,334	0	0
3	82	48	38,4	2,137	1	1
2	86	42	38,4	0,246	0	0
5	64	33	31,2	0,583	1	0
1	80	45	34,8	0,217	0	0
1	60	46	35,5	0,415	0	0
2	68	42	38,2	0,503	1	0
2	52	43	36,2	0,816	0	0
0	80	45	36,5	0,33	0	0
10	78	31	27,6	0,512	0	1
5	72	28	23,9	0,407	0	0
2	68	22	28,7	0,092	0	0
0	104	25	34,3	0,435	1	0

Lampiran 5 : Hasil Klasifikasi GNB Model E

Glucose	Insulin	Age	Actual	Predicted
162	100	25	1	1
180	90	35	1	1
86	109,4	25	0	0
100	71	26	0	0
88	16	22	0	0
112	94	26	0	0
84	115	28	0	0
99	86	24	0	0
127	155	28	0	0
129	125	43	1	0
139	160	25	1	0
122	200	26	0	0
174	194	36	1	1
61	53,2	46	0	0
108	278	22	0	0
87	52	25	0	0
104	116	23	0	0
86	65	29	0	0
80	70	27	1	0
112	132,6	25	1	0
162	357,4	26	1	1
104,2	23	21	0	0
107	74	25	1	0
80	60	22	0	0
107	117,6	23	0	0
83	66	24	0	0
187	392	34	1	1
173	265	58	0	1
137	148	21	0	0
198	274	28	1	1
95	105	22	0	0
149	127	42	1	1
93	72	35	0	0
179	230,6	60	0	1
103	82	22	0	0

G	Ins	A	Ac	P
133	155	37	0	0
121	165	33	1	0
90	59	25	0	0
106	148	22	0	0
71	45,2	22	0	0
129	122	39	0	0
126	75	39	0	0
177	478	21	1	1
154	140	27	0	0
100	105	24	0	0
119	170	26	0	0
141	128	24	0	0
79	37	22	0	0
181	180	38	1	1
127	335	22	0	0
137	168	33	1	0
110	125	27	0	0
81	54,8	25	0	0
103	135,2	42	0	0
151	330,4	21	1	1
144	178,2	37	1	1
99	94	21	0	0
187	304	41	1	1
189	846	59	1	1
93	92	22	0	0
107	85	24	0	0
147	250,4	27	1	1
87	32	22	0	0
98	120	22	0	0
184	169	49	1	1
126	215	24	0	0
118	89	21	0	0
113	85,2	22	0	0
110	100	27	0	0
144	285	58	1	1

G	Ins	A	Ac	P
109	135	23	0	0
116	105	24	0	0
90	62,4	22	0	0
111	150	45	1	0
99	54	32	0	0
111	75,4	23	0	0
95	73	36	0	0
114	110	31	0	0
129	130	26	1	0
142	190	61	0	1
144	140	37	0	0
168	321	40	1	1
98	84	22	0	0
137	108,2	39	0	0
106	119	34	0	0
102	136,6	46	1	0
99	139,8	32	0	0
136	130	42	1	0
104	156	41	1	0
117	106	27	0	0
108	96,2	32	0	0
186	225	37	1	1
95	38	25	0	0
173	465	25	1	1
112	160	28	0	0
189	325	29	1	1
112	132	24	0	0
108	178	24	0	0
90	113,4	27	1	0
122	158	28	0	0
95	92	26	0	0
122	205,4	45	0	0
111	127	27	0	0
114	85,2	25	0	0
189	180,6	41	1	1

Lampiran 6 : Hasil Klasifikasi GNB Model F

Pregnancies	Blood Pressure	DPF	Actual	Predicted
0	76	0,759	1	0
0	90	0,314	1	0
0	68	0,238	0	0
2	68	0,324	0	0
2	58	0,766	0	0
2	68	0,315	0	0
1	64	0,471	0	0
3	54	0,154	0	0
4	88	0,598	0	0
7	68	0,439	1	0
5	80	0,361	1	0
2	76	0,483	0	0
3	58	0,593	1	0
3	82	0,243	0	0
2	62	0,881	0	0
2	58	0,166	0	0
0	64	0,454	0	0
1	66	0,917	0	0
3	82	1,292	1	1
3	74	0,197	1	0
0	76	0,364	1	0
1	74	0,299	0	0
0	62	0,757	1	0
1	74	0,527	0	0
0	60	0,133	0	0
2	65	0,629	0	0
7	50	0,826	1	0
0	78	1,159	0	1
0	68	0,143	0	0
0	66	0,502	1	0
0	64	0,366	0	0
1	68	0,349	1	0
0	100	1,021	0	1
7	95	0,164	0	1
1	80	0,491	0	0

P	BP	DPF	A	Pr
7	88	0,262	0	0
0	66	0,203	1	0
1	62	1,268	0	1
0	70	0,605	0	0
2	70	0,586	0	0
10	76	0,28	0	1
8	74	0,162	0	0
0	60	1,072	1	0
6	78	0,571	0	0
2	54	0,498	0	0
1	88	0,507	0	0
2	58	0,699	0	0
1	80	0,583	0	0
1	64	0,328	1	0
2	46	0,176	0	0
0	40	2,288	1	1
2	74	0,698	0	0
2	60	0,29	0	0
11	68	0,126	0	1
0	90	0,371	1	0
4	82	0,554	1	0
2	52	0,637	0	0
7	68	0,254	1	0
1	60	0,398	1	0
0	60	0,532	0	0
1	68	0,165	0	0
1	94	0,358	1	0
1	78	0,101	0	0
2	60	0,198	0	0
9	85	1,213	1	1
0	84	0,52	0	0
0	64	1,731	0	1
3	44	0,14	0	0
4	76	0,118	0	0
5	82	0,452	1	0

P	BP	DPF	A	P
1	56	0,833	0	0
3	74	0,107	0	0
2	70	0,085	0	0
11	84	0,925	1	1
6	60	0,497	0	0
1	86	0,143	0	0
1	74	0,673	0	0
7	76	0,466	0	0
0	110	0,319	1	0
7	60	0,687	0	0
4	58	0,287	0	0
7	88	0,787	1	1
0	82	0,299	0	0
7	90	0,391	0	1
2	64	1,4	0	1
9	76	0,665	1	1
5	74	0,203	0	0
11	84	0,26	1	1
6	74	0,722	1	0
1	60	0,466	0	0
0	68	0,787	0	0
8	90	0,423	1	1
1	66	0,334	0	0
3	82	2,137	1	1
2	86	0,246	0	0
5	64	0,583	1	0
1	80	0,217	0	0
1	60	0,415	0	0
2	68	0,503	1	0
2	52	0,816	0	0
0	80	0,33	0	0
10	78	0,512	0	1
5	72	0,407	0	0
2	68	0,092	0	0
0	104	0,435	1	0