KLASIFIKASI PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

THESIS

Oleh:
YUSRIL HAZA MAHENDRA
NIM. 230605220009

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

KLASIFIKASI PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

THESIS

Diajukan Kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
YUSRIL HAZA MAHENDRA
NIM. 230605220009

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

HALAMAN PENGAJUAN

KLASIFIKASI PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

THESIS

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
YUSRIL HAZA MAHENDRA
NIM. 230605220009

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

HALAMAN PERSETUJUAN

KLASIFIKASI PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

TESIS

Oleh:
YUSRIL HAZA MAHENDRA
NIM. 230605220009

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 22 Oktober 2025

Pembimbing I,	Pembimbing II,
Dr. Ririen Kusumawati, S.Si., M.Kom	Dr. M. Imamudin, Lc., MA
NIP.19720309200501002	NIP.197406022009011010

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

Prof. Dr. Ir. Muhammad Faisal, S. Kom., M.T
NIP.197405102005011007

HALAMAN PERSEMBAHAN

KLASIFIKASI PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

TESIS

Oleh:
YUSRIL HAZA MAHENDRA
NIM. 230605220009

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 22 Oktober 2025

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Dr.Ir Fresy Nugroho, S.T., M.T,IPM NIP 197107222011011001	( )
Penguji II	:	Dr. M. Amin Hariyadi, M.T NIP. 196701182005011001	( )
Pembimbing I	:	Dr. Ririen Kusumawati, S.Si., M.Kom NIP 197203092005012002	( )
Pembimbing II	:	Dr. M. Imamudin, Lc, MA NIP. 197406022009011010	( )

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

Prof. Dr. H. Muhammad Faisal, S. Kom., M.T
NIP.197405102005011007

HALAMAN PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama	: Yusril Haza Mahendra
NIM	: 230605220009
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

Dengan ini menyatakan bahwa tesis yang saya tulis benar-benar merupakan hasil karya saya sendiri. Tesis ini bukan merupakan pengambilalihan data, tulisan, atau pemikiran orang lain yang saya akui sebagai hasil karya saya, kecuali dengan mencantumkan sumber referensi pada daftar pustaka sesuai dengan kaidah ilmiah yang berlaku.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa tesis ini adalah hasil jiplakan, saya bersedia menerima segala sanksi yang berlaku atas perbuatan tersebut.

Malang, 22 Oktober 2025
Yang membuat pernyataan

Yusril Haza Mahendra
NIM.230605220009

MOTTO

إِنَّ اللَّهَ لَا يُغَيِّرُ مَا بِقَوْمٍ حَتَّى يُغَيِّرُوا مَا بِأَنْفُسِهِمْ

"Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum sebelum mereka mengubah keadaan diri mereka sendiri." (QS. Ar-Ra'd: 11)

"Indeed, Allah will not change the condition of a people until they change what is in themselves." (QS. Ar-Ra'd: 11)

HALAMAN PERSEMBAHAN

Kupersembahkan karyaku ini:

1. Kedua orang tua Bapak Butani dan Ibu Minawarah tercinta yang selalu memberi cinta yang tulus dan tidak perna akan putus sambugan do'a dan yang selalu mengalir terus.
2. KHR. Azaim Ibrahimy, Selaku pengasuh Ponpes Salafiya Syafi'iya, atas segala doa dan bimbingannya serta izin yang di berikan kepada kami.
3. Saudara dan keluarga besar yang selalu menjadi roda-roda ketika aku jatuh dan memberi semangat yang besar.
4. Segenap sivitas akademika Fakultas Sains dan Teknologi Universitas Ibrahimy.
5. Teman - Teman Magister Inforamatika semua yang angkatan IX khusunya, mereka teman perjuanganku.

Semoga karya sederhana ini dapat memberikan manfaat, baik bagi pembaca, maupun bagi saya secara pribadi.

Wassalamu'alaikum Wr. Wb.

Malang 22 Oktober 2025

Yusril Haza Mahendra

KATA PENGANTAR

Assalamu’alaikum Wr. Wb.

Syukur alhamdulillah saya panjatkan ke hadirat Allah SWT atas limpahan rahmat, hidayah, dan inayah-Nya, sehingga saya dapat menyelesaikan studi di Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, serta menyelesaikan tesis ini dengan baik.

Selanjutnya penulis haturkan ucapan terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah memberikan dukungan, bimbingan, serta doa dalam penyelesaian tesis ini. Ucapan terima kasih saya sampaikan kepada:

1. Ibu Dr. Ririen Kusumawati, S. Si., M.Kom dan Bapak Dr. Imamuddin, Lc., M.Ag, selaku dosen pembimbing Tesis, atas segala bimbingan, arahan, dan pengalaman berharga yang diberikan selama proses penulisan.
2. Segenap sivitas akademika Program Studi Magister Informatika, khususnya para dosen yang telah memberikan ilmu dan bimbingan selama masa studi.
3. Segenap keluarga, yang selalu memberikan semangat dan motivasi hingga saya dapat menyelesaikan tesis ini.
4. Semua pihak yang telah membantu, baik secara material maupun moral, yang tidak dapat saya sebutkan satu per satu.

Saya menyadari bahwa tesis ini masih memiliki kekurangan. Semoga karya ini dapat memberikan manfaat, baik bagi pembaca, maupun bagi saya secara pribadi.

Wassalamu’alaikum Wr. Wb.

Malang 22 Oktober 2025

Yusril Haza Mahendra

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Mahendra Haza Yusril 2025. **Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma *Naive Bayes* Dan *Random Forest.*** Tesis. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (1) Dr. Ririen Kusumawati, M.Kom (II) Dr. Imamudin, Lc., MA.

Diabetes Mellitus merupakan penyakit kronis dengan prevalensi yang terus meningkat dan membutuhkan sistem deteksi dini yang akurat untuk mendukung pengambilan keputusan medis secara cepat dan objektif. Penelitian ini bertujuan untuk menganalisis dan membandingkan performa algoritma *Naive Bayes* dan *Random Forest* dalam klasifikasi penyakit Diabetes Mellitus berdasarkan parameter klinis seperti kadar glukosa darah, Body Mass Index (BMI), tekanan darah, usia, dan riwayat penyakit. Data dibagi menjadi data latih dan data uji dengan rasio 70:30, kemudian melalui tahapan preprocessing, pemodelan, dan evaluasi menggunakan metrik akurasi, *precision*, recall, dan F1-score. Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memberikan performa terbaik dengan akurasi sebesar 98,00%, *precision* 97,82%, recall 98,10%, dan F1-score 97,96%, sedangkan algoritma *Naive Bayes* menghasilkan akurasi 90,00%, *precision* 89,30%, recall 88,75%, dan F1-score 89,02%. Hasil klasifikasi menunjukkan bahwa sebesar 91,5% pasien diprediksi tidak terkena diabetes dan 8,5% pasien diprediksi positif diabetes, menggambarkan dominasi kelas negatif diabetes pada populasi data. Temuan ini menunjukkan bahwa *Random Forest* lebih unggul dalam mengidentifikasi pola data yang kompleks serta memberikan prediksi yang lebih akurat dan andal dibandingkan *Naive Bayes*. Analisis fitur juga mengungkap bahwa BMI, usia, riwayat merokok, dan kadar glukosa darah merupakan variabel paling berpengaruh dalam menentukan risiko Diabetes Mellitus.

Kata Kunci: Diabetes Mellitus, *Naive Bayes, Random Forest,* Klasifikasi.

ABSTRACT

Mahendra Haza Yusril 2025. **Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma *Naive Bayes* Dan *Random Forest.*** Theses. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (1) Dr. Ririen Kusumawati, M.Kom (II) Dr. Imamudin, Lc., MA.

Diabetes Mellitus is a chronic disease with increasing prevalence and requires an accurate early detection system to support rapid and objective medical decision-making. This study aims to analyze and compare the performance of *the* Naive Bayes and Random Forest algorithms in the classification of Diabetes Mellitus disease based on clinical parameters such as blood glucose levels, Body Mass Index (BMI), blood pressure, age, and disease history. The data was divided into training data and test data with a 70:30 ratio, then through preprocessing, modeling, and evaluation stages using accuracy, *precision*, recall, and F1-score metrics. The test results showed that *the* Random Forest algorithm provided the best performance with an accuracy of 98.00%, precision of 97.82%, recall of 98.10%, and an F1-score of 97.96%, while the *Naive Bayes* algorithm produced an accuracy of 90.00%, precision of 89.30%, recall of 88.75%, and F1 score 89.02%. The classification results showed that 91.5% of patients were predicted not to have diabetes and 8.5% of patients were predicted to be diabetic, illustrating the dominance of the diabetic negative class in the data population. These findings show that Random Forest is superior at identifying complex data patterns and providing more accurate and reliable predictions than *Naive Bayes*. Feature analysis also revealed that BMI, age, smoking history, and blood glucose levels were the most influential variables in determining the risk of Diabetes Mellitus.

Keywords: Diabetes Mellitus, *Naive Bayes, Random Forest*, Classification.

الملخص

مهندرا هزا يسرل، ٢٠٢٥ .*تصنيف مرض السكري باستخدام خوارزميتي* بايز السادج ، *الغابات العشوائية*.رسالة الماجستير، برنامج ماجستير المعلوماتية، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية مولانا مالك إبراهيم مالانج. المشرفون: (١) الدكتورة ريرين كوسوماواتي، ماجستير في الحاسوب، (٢) الدكتور إمام الدين، ليسانس، ماجستير.

داء السكري هو مرض مزمن ينتشر بشكل متزايد ويتطلب نظاما دقيقا للكشف المبكر لدعم اتخاذ القرارات الطبية السريعة والموضوعية. تهدف هذه الدراسة إلى تحليل ومقارنة أداء خوارزميات ساذج بايز و *غابة عشوائية* في تصنيف مرض داء السكري بناء على المعايير السريرية مثل مستويات الجلوكوز في الدم ومؤشر كتلة الجسم (*مؤشر كتلة الجسم*) وضغط الدم والعمر وتاريخ المرض. تم تقسيم البيانات إلى بيانات التدريب وبيانات الاختبار بنسبة ٣٥:٧٥ ، ثم من خلال مراحل المعالجة المسبقة والنمذجة والتقييم باستخدام مقاييس الدقة *والدقة* والاستدعاء ودرجة F١. أظهرت نتائج الاختبار أن خوارزمية *الغابة العشوائية* قدمت أفضل أداء بدقة ٩٨,٠٠٪ ، ودقة ٩٧٠٨٢ ٪ ، واستدعاء ٠١.٩٨٪ ، ودرجة F١ ٩٦.٩٧٪ ، بينما *أنتجت خوارزمية* ساذج بايز دقة ٩٠.٠٠٪ ، ودقة ٨٩.٣٠٪ ، واستدعاء ٨٨.٧٥٪ ، و درجة F١ ٨٩.٠٢٪ أظهرت نتائج التصنيف أنه من المتوقع ألا يكون ٩١.٥٪ من المرضى مصابين بمرض السكري و ٨٥٪ من المرضى متوقعون أن يكونوا مصابين بالسكري ، مما يوضح هيمنة الفئة السلبية لمرضى السكري في مجتمع البيانات. تظهر هذه النتائج أن *غابة عشوائية* متفوقة في تحديد أنماط البيانات المعقدة وتقديم تنبؤات أكثر دقة وموثوقية من ساذج بايز. كشف تحليل الميزات أيضا أن مؤشر كتلة الجسم والعمر وتاريخ التدخين ومستويات الجلوكوز في الدم كانت المتغيرات الأكثر تأثيرا في تحديد خطر الإصابة بداء السكري.

الكلمات المفتاحية :داء السكري، بايز السادج ، *الغابات العشوائية* ، التصنيف.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus merupakan penyakit kronis yang menunjukkan peningkatan prevalensi secara signifikan di tingkat global. Kondisi ini ditandai dengan tingginya kadar glukosa dalam darah akibat ketidakseimbangan fungsi tubuh dalam memproduksi maupun memanfaatkan insulin secara efisien. Akibatnya, kadar gula darah meningkat dan menjadi sulit dikendalikan. Faktor risiko utama yang berkontribusi terhadap timbulnya diabetes meliputi obesitas, pola hidup sedentari, konsumsi makanan tinggi gula dan lemak, serta predisposisi genetik. Secara umum, diabetes berkembang secara progresif dan lebih banyak dialami oleh orang dewasa, meskipun dalam beberapa tahun terakhir kasusnya juga semakin sering ditemukan pada anak-anak dan remaja seiring meningkatnya prevalensi obesitas Geetha & Prasad. (2023).

Diabetes tipe 1 merupakan gangguan autoimun di mana sistem imun tubuh secara keliru menyerang dan merusak sel beta pankreas yang berperan dalam produksi insulin. Penyebabnya meliputi faktor *genetik,* lingkungan, dan gaya hidup. Diabetes tipe 2 terjadi karena resistensi *insulin*, dimana tubuh tidak merespons *insulin* dengan baik. Pankreas awalnya memproduksi lebih banyak *insulin* untuk mengatasi resistensi ini, tetapi akhirnya produksinya menurun, sehingga kadar gula darah meningkat.

Faktor penyebab utamanya adalah *genetik*, kelebihan berat badan, *obesitas*, dan gaya hidup yang kurang aktif Wu *et al.* (2023).

Diabetes gestasional muncul selama kehamilan akibat hormon yang mengganggu kerja *insulin*. Faktor risiko termasuk riwayat *prediabetes* dan keluarga dengan diabetes. Sekitar 50% kasus diabetes gestasional dapat dikonfirmasi dari sumber yang terpercaya Butt *et al*. (2021).

Terdapat ayat Al-Quran yang melarang manusia dalam berlebihan termasuk pada makanan. Allah berfirman di Al-Quran surah Al-A'raf ayat 31 yang artinya:

يَٰبَنِيٓ ءَادَمَ خُذُواْ زِينَتَكُمۡ عِندَ كُلِّ مَسۡجِدٖ وَكُلُواْ وَٱشۡرَبُواْ وَلَا تُسۡرِفُوٓاْۚ إِنَّهُۥ لَا يُحِبُّ ٱلۡمُسۡرِفِينَ

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

Menurut Tafsir Ibnu Katsir, ayat ini menegaskan bahwa manusia diperintahkan untuk menikmati makanan dan minuman yang halal, namun tetap dalam batas kewajaran tanpa berlebih-lebihan. Ibnu Katsir menjelaskan bahwa berlebihan dalam makan dan minum dapat merusak tubuh dan menimbulkan penyakit. Konteks ini sejalan dengan kondisi penderita diabetes yang umumnya dipicu oleh pola konsumsi yang tidak seimbang, khususnya akibat asupan gula dan lemak yang berlebihan. (Ibnu Katsir, *Tafsir al-Qur'an al-Azhim*, Juz 7). Imam Al-Qurthubi dalam tafsirnya juga menjelaskan bahwa larangan berlebihan (israf) dalam ayat ini tidak hanya berlaku untuk konsumsi makanan, tetapi juga untuk perilaku dalam kehidupan sehari-hari. Makan dan minum yang melampaui batas

dianggap sebagai perbuatan yang tidak diridhai Allah karena dapat mendatangkan mudarat bagi tubuh (Al-Qurthubi, *Tafsir al-Jami' li-Ahkam al-Qur'an*, Juz 7).

وَإِذَا مَرِضْتُ فَهُوَ يَشْفِينِ

"Dan apabila aku sakit, Dialah yang menyembuhkan aku" (Qs.Asy- Syu'ara. 80)

Ayat ini merupakan pernyataan Nabi Ibrahim a.s. yang menegaskan keimanannya kepada Allah sebagai satu-satunya pemberi kesembuhan. Dalam konteks ini, sakit adalah bagian dari ujian hidup, sedangkan kesembuhan datang dari Allah melalui sebab yang ditentukan-Nya, seperti pengobatan atau doa. Ayat ini mengajarkan pentingnya tawakal kepada Allah dalam menghadapi penyakit.

Dalam era digital dan kemajuan teknologi informasi, klasifikasi telah menjadi alat yang sangat berguna dalam bidang kesehatan. Klasifikasi memungkinkan peneliti dan praktisi kesehatan untuk mengeksplorasi dan klasifikasi data kesehatan dalam skala besar untuk mengidentifikasi pola, hubungan, dan tren yang mungkin sulit dideteksi secara manual. Dalam konteks deteksi *Diabetes mellitus*, klasifikasi memiliki potensi besar untuk membantu dalam pengembangan *model prediktif* yang dapat mengidentifikasi individu yang berisiko tinggi untuk mengembangkan diabetes atau membantu dalam manajemen penyakit pada individu yang sudah didiagnosis Chao & Li, (2022).

Dalam konteks ini, pemilihan metode yang tepat untuk klasifikasi diagnosis diabetes menjadi sangat penting. Dua metode yang sering digunakan dalam klasifikasi data medis adalah *Naive Bayes* dan *Random Forest*. Metode *Naive Bayes* adalah algoritma klasifikasi sederhana yang berbasis pada teori probabilitas.

Keunggulannya adalah kemampuannya untuk menangani *dataset* besar dan beragam, seperti data medis untuk diagnosis diabetes. *Naïve Bayes* cepat dalam proses pelatihan dan klasifikasi, serta efektif dalam klasifikasi meskipun ada asumsi independensi antar fitur. Meskipun jarang terjadi dalam kenyataan, asumsi ini tidak mengurangi akurasi, terutama pada data dengan jumlah sampel yang banyak dan distribusi yang stabil. *Random Forest* merupakan salah satu metode pembelajaran *ensemble* yang mengombinasikan sejumlah pohon keputusan untuk menghasilkan model klasifikasi dengan tingkat akurasi yang lebih tinggi. Keunggulan utama algoritma ini terletak pada kemampuannya dalam mengolah data yang bersifat kompleks dan melibatkan banyak variabel yang saling berhubungan, seperti yang sering dijumpai pada kasus diagnosis penyakit diabetes. *Random Forest* mengurangi risiko *overfitting* dan memberikan wawasan tentang pentingnya fitur-fitur dalam model, membantu memahami faktor utama dalam diagnosis diabetes Utomo *et al*. (2020).

Pendekatan *Naive Bayes* dan *Random Forest* untuk deteksi *Diabetes mellitus* melibatkan penggunaan berbagai teknik, termasuk tetapi tidak terbatas pada klasifikasi, *klastering,* dan asosiasi. Dengan memanfaatkan *dataset* yang mencakup informasi tentang faktor risiko seperti riwayat keluarga, pola makan, aktivitas fisik, dan hasil tes laboratorium, model-model *prediktif* dapat dikembangkan untuk mengidentifikasi individu yang rentan terhadap diabetes Sreehari & Babu. (2024).

Naive Bayes telah terbukti efektif dalam memberikan klasifikasi yang baik dalam beberapa studi diabetes. Dalam sebuah penelitian yang membandingkan

teknik data mining untuk diagnosis dini diabetes, *Naive Bayes* menunjukkan kinerja yang cukup baik dalam klasifikasi data dengan akurasi tinggi, meskipun *neural network* memberikan hasil yang sedikit lebih baik dalam hal akurasi Marques. (2021) *Random Forest* juga sering digunakan untuk klasifikasi diabetes karena kemampuannya mengatasi data yang kompleks. Dalam penelitian yang dilakukan oleh Assegie & Nair. (2020). *Random Forest* termasuk dalam teknik yang diuji untuk mendeteksi diabetes, dengan hasil yang cukup baik dalam mengidentifikasi variabel yang berhubungan dengan diabetes.

Tujuan penelitian ini berfokus pada perbandingan algoritma *Naive Bayes* dan *Random Forest* dalam klasifikasi deteksi *diabetes mellitus*, dengan pendekatan data mining untuk menganalisis tingkat akurasi kedua algoritma. Sementara banyak penelitian sebelumnya menilai algoritma-algoritma ini dalam konteks klasifikasi penyakit diabetes, penelitian ini tidak hanya melihat pada akurasi klasifikasi, tetapi juga bertujuan untuk menilai seberapa efektif kedua algoritma dalam memberikan hasil yang optimal dalam klasifikasi diabetes. Dengan demikian, penelitian ini bertujuan untuk menggali lebih dalam mengenai kemampuan kedua algoritma dalam meningkatkan akurasi diagnosis, yang pada gilirannya dapat mendukung upaya pencegahan lebih awal dan pengelolaan *diabetes mellitus*.

Pendekatan ini relevan dengan penelitian-penelitian sebelumnya yang menunjukkan bahwa *Random Forest* mampu menangani kompleksitas interaksi antar variabel lebih baik daripada *Naive Bayes*, namun *Naive Bayes* tetap memberikan hasil yang efisien dalam kondisi *dataset* besar dan beragam. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi lebih lanjut pada

pemahaman tentang algoritma mana yang lebih efektif dalam meningkatkan tingkat akurasi dalam klasifikasi diabetes, yang nantinya dapat berkontribusi dalam pengurangan beban kesehatan masyarakat akibat *diabetes mellitus*.

1.2 Pernyataan Masalah

1. Bagaimana mencari akurasi di dalam klasifikasi diagnosis penyakit *Diabetes mellitus* menggunakan algoritma *Naive Bayes* dan *Random Forest*?
2. Bagaimana klasifikasi diagnosis penyakit *Diabetes mellitus* ?

1.3 Tujuan Penelitian

1. Mendapatkan akurasi pada klasifikasi *Diabetes mellitus*, menggunakan algoritma *Naive Bayes* dan *Random Forest.*
2. Menghasilkan klasifikasi diagnosis penyakit *Diabetes mellitus*.

1.4 Batasan Masalah

Dalam penelitian ini, terdapat sejumlah batasan yang perlu diperhatikan untuk memperjelas ruang lingkup dan konteks dari klasifikasi dalam deteksi *Diabetes mellitus* :

1. Keterbatasan *Dataset* yang Digunakan

Penelitian yg di lakukan menggunakan *dataset* yang tersedia secara publik, seperti *dataset kaggle*, sehingga hasil klasifikasi dan *model prediktif* yang dikembangkan mungkin tidak mencerminkan populasi yang lebih luas atau faktor risiko yang unik di luar data tersebut.

2. Fokus pada metode *Naive Bayes* dan *Random Forest*

Penelitian ini akan terbatas pada penggunaan algoritma *Naive Bayes* dan *Random Forest* dalam pendekatan *data mining*, sehingga potensi metode lain yang mungkin lebih efektif atau efisien dalam mendeteksi *Diabetes mellitus* tidak akan dieksplorasi dalam studi ini.

3. Tidak menangani aspek klinis secara mendalam

Penelitian ini akan fokus pada pengembangan model prediktif dan klasifikasi data, namun tidak akan mengkaji secara mendalam aspek klinis atau *fisiologis* dari *Diabetes mellitus*, seperti pengaruh intervensi medis tertentu atau pengelolaan penyakit dalam jangka panjang.

1.5 Manfaat Penelitian

1. Bagi Peneliti

Untuk menambah wawasan serta ilmu pengetahuan peneliti dan memahami cara kerja proses *Naive Bayes* dan *Random Forest* klasifikasi diagnosis penyakit *Diabetes mellitus*.

2. Bagi akademis atau Pembaca

Penelitian ini diharapkan dapat memberikan kontribusi serta acuan (referensi) bagi para pembaca agar ilmu pengetahuannya bertambah terkait klasifikasi penyakit *Diabetes mellitus*.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini terbagi menjadi enam bab. Setiap bab terdiri dari beberapa subbab, yang dijelaskan sebagai berikut:

BAB I: PENDAHULUAN

Bab ini menjelaskan latar belakang, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan.

BAB II: LANDASAN TEORI

Bab ini memuat teori-teori yang relevan dengan penelitian, seperti penjelasan mengenai penyakit diabetes, konsep machine learning, *supervised learning*, algoritma *Random Forest*, *Naïve Bayes*, serta studi literatur terkait.

BAB III: METODOLOGI PENELITIAN

Bab ini membahas metode penelitian yang digunakan, termasuk metode pengumpulan data melalui studi pustaka dan studi literatur, serta alur penelitian yang dilakukan.

BAB IV: IMPLEMENTASI

Bab ini menjelaskan proses implementasi pada alur penelitian, termasuk penerapan model yang diusulkan.

BAB V: HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil implementasi serta pengujian model menggunakan data baru untuk mengevaluasi akurasi.

BAB VI: KESIMPULAN

Bab ini menyajikan kesimpulan dari proses Klasifikasi penyakit diabetes menggunakan algoritma *Random Forest* dan *Naïve Bayes,* serta memberikan saran untuk penelitian di masa mendatang.

BAB II

STUDI PUSTAKA

2.1 Klasifikasi Penyakit *Diabetes mellitus*

Penelitian yang dilakukan bertujuan membandingkan algoritma *Naive Bayes, Random Forest,* menggunakan *dataset Pima Indians Diabetes*. Hasilnya, *Random Forest* memberikan performa terbaik pada *dataset* lengkap dengan akurasi 79.57% *Naive Bayes* lebih unggul pada *dataset* dengan fitur yang diseleksi, mencapai akurasi 79.13%. Selain itu, *J48 Decision Tree* memiliki sensitivitas tertinggi (88.43%). Penelitian ini menunjukkan bahwa pemilihan algoritma sangat bergantung pada kompleksitas *dataset*, dimana *Naive Bayes* bekerja optimal pada fitur yang lebih spesifik, sedangkan *Random Forest* unggul dalam analisis *dataset* kompleks Chang. (2023).

Penelitian yang dilakukan bertujuan membandingkan algoritma *Naive Bayes* pada *dataset* diabetes yang terdiri dari 200 data. Hasilnya menunjukkan bahwa *Naive Bayes* memiliki akurasi tertinggi sebesar 80% unggul dalam *recall* (0.92), meskipun akurasinya lebih rendah (75%). Penelitian ini menegaskan keunggulan *Naive Bayes* lebih efektif dalam mendeteksi data positif. Secara keseluruhan, kedua penelitian menyimpulkan bahwa *Naive Bayes* dapat menjadi algoritma yang lebih andal untuk klasifikasi diagnosis diabetes, meskipun algoritma lain seperti dan *Random Forest* tetap memiliki keunggulan tergantung pada kebutuhan spesifik analisis Putry *et al*. (2022).

Penelitian yang dilakukan bertujuan untuk membandingkan tiga metode data mining, yaitu *Naive Bayes*, dan *Logistic Regression*, untuk memKlasifikasi penyakit diabetes menggunakan aplikasi *RapidMiner*. Hasil evaluasi menunjukkan bahwa *Logistic Regression* merupakan metode yang paling efektif dalam mendeteksi diabetes secara dini, dengan akurasi tertinggi sebesar 75.78% dan *AUC* 0.801. Metode *Naive Bayes* dan *Neural Network* memiliki akurasi yang lebih rendah, masing-masing 74.87% (AUC 0.799) dan 69.27% (*AUC* 0.736). Temuan ini menekankan pentingnya pemilihan algoritma yang tepat dalam pengembangan sistem deteksi dini penyakit diabetes untuk meningkatkan akurasi diagnosis Khanam & Foo. (2021).

Penelitian yang dilakukan berhasil menunjukkan bahwa algoritma *Naive Bayes* dapat digunakan secara efektif untuk mengklasifikasikan pasien diabetes melitus berdasarkan data yang tersedia. Dengan menggunakan data dari RS Dirgahayu Samarinda selama periode 2018 hingga 2021, penelitian ini mengevaluasi berbagai proporsi data *training* dan *testing*. Hasil analisis menunjukkan bahwa akurasi tertinggi yang dicapai adalah 92,31% pada proporsi data 70:30 dan 80:20, yang menunjukkan bahwa model ini memiliki kemampuan yang baik dalam memKlasifikasi status diabetes melitus pasien Khasanah *et al*. (2022).

Hasil penelitian menunjukkan membandingkan efektivitas algoritma *Naive Bayes* konvensional dengan algoritma *Naive Bayes* yang menggunakan pemilihan atribut berbasis *gain ratio* dalam memKlasifikasi komplikasi hipertensi. Hasil penelitian menunjukkan bahwa penerapan *gain ratio* dalam proses klasifikasi

meningkatkan akurasi dan performa algoritma *Naive Bayes*. Dengan menggunakan data rekam medis pasien hipertensi, algoritma *Naive Bayes gain ratio* berhasil memberikan hasil Klasifikasi yang lebih baik dibandingkan dengan metode konvensional, dengan peningkatan akurasi yang signifikan. Temuan ini menunjukkan potensi penggunaan teknik pemilihan atribut dalam meningkatkan efektivitas model prediktif untuk masalah kesehatan, khususnya dalam konteks pencegahan dan penanganan komplikasi hipertensi di Indonesia Arya *et al*. (2024).

Hasil penelitian menunjukkan bahwa melakukan penerapan metode *Particle Swarm Optimization* (PSO) sebagai teknik optimasi pada algoritma *Random Forest* terbukti efektif dalam meningkatkan akurasi klasifikasi diabetes. Akurasi meningkat dari 78.2% menjadi 82.1%, yang menunjukkan peningkatan signifikan sebesar 3.9%. Selain peningkatan akurasi, algoritma *Random Forest* yang dioptimalkan dengan PSO juga menunjukkan peningkatan dalam nilai *recall* sebesar 2.44% dan peningkatan *precision* sebesar 0.07%. Hal ini menunjukkan bahwa model yang dioptimalkan tidak hanya lebih akurat, tetapi juga lebih baik dalam mengidentifikasi kasus positif diabetes Pratama *et al*. (2023).

Penelitian yang dilakukan Penelitian ini menunjukkan bahwa algoritma *Random Forest* efektif dalam klasifikasikan penyakit diabetes menggunakan *dataset* dari gula karya medika. Penggunaan teknik normalisasi data, khususnya *Min-max normalization*, terbukti meningkatkan akurasi model secara signifikan, dengan model ini mencapai akurasi tertinggi sebesar 95.45%. Sebaliknya, model yang tidak menggunakan normalisasi data hanya mencapai akurasi 92% Abnoosian *et al*. (2023).

Hasil penelitian menunjukkan bahwa kombinasi metode *Synthetic Minority Oversampling Technique* (SMOTE) dengan algoritma *Random Forest* secara signifikan mampu meningkatkan kinerja model dalam melakukan klasifikasi pada kasus penyakit kanker paru-paru, dengan akurasi mencapai 94.1%, sensitivitas 94.5%, dan spesifisitas 93.7%. Sebagai perbandingan, tanpa menggunakan SMOTE, akurasi hanya 89.1% dan sensitivitasnya jauh lebih rendah, yaitu 55%. Peningkatan akurasi sebesar 5% dan sensitivitas sebesar 39% menunjukkan pentingnya mengatasi masalah ketidakseimbangan data dalam klasifikasi Lauw *et al.* (2023).

Berikut rangkuman penelitian-penelitian terkait yang telah dipaparkan dalam bentuk tabel yang ditunjukkan pada tabel 2.1:

Tabel 2.1 Penelitian terkait

No	**Peneliti**	**Topik**	**Metode**	**Hasil**
1	Chang. (2023)	*Klasifikasi diabetes*	*Naive Bayes, Random Forest,*	*Random Forest* akurasi tertinggi (79.57%), Naïve Bayes 79.13%. J48 Decision Tree tertinggi (88.43%)
2	Putry *et al*. (2022)	Mencari Akurasi pada *dataset* diabetes.	KNN dan *Naive Bayes*	Naive Bayes: Akurasi 80% KNN: Akurasi 75%
3	Khanam & Foo, (2021)	Algoritma pengembangan sistem deteksi dini penyakit diabetes	*Naive Bayes*, dan *Logistic Regression*,	Akurasi 74.87% (AUC 0.799) dan 69.27% (*AUC* 0.736)
4	Khasanah *et al*. (2022)	Klasifikasikan pasien diabetes melitus.	Algoritma *Naive Bayes Classifier*	Hasil penelitian menunjukkan akurasi terbaik yang dicapai adalah 92,31% pada proporsi data 60:40 dan 80:20
5	Arya *et al*. (2024)	Klasifikasi komplikasi hipertensi	Algoritma *Naive Bayes*	*Naive Bayes* konvensional, dengan peningkatan akurasi berkisar antara 0,03% hingga 16%.

Lanjutan				
6	Pratama *et a*l. (2023)	Peningkatan akurasi klasifikasi diabetes.	Algoritma *Random Forest*	Penelitian ini menunjukkan bahwa akurasi algoritma *Random Forest* tanpa optimasi adalah 78.2%.
7	Abnoosian *et a*l. (2023)	Klasifikasikan penyakit diabetes	Algoritma *Random Forest*	Model ini mencapai akurasi tertinggi sebesar 95.45%.
8	Lauw *et al*. (2023)	*Combination of Smote and Random Forest Methods for Lung Cancer Classification*	Metode *Random Forest*	Akurasi yang diperoleh adalah 89.1%, dengan *sensitivitas* 55%, dan *spesifisitas* 94.5%

2.2 Landasan Teori

Landasan teori adalah bagian dari penelitian yang menjelaskan konsep, teori, dan prinsip-prinsip yang mendasari atau menjadi dasar bagi penelitian yang dilakukan.

2.2.1 *Diabetes mellitus*

Diabetes mellitus adalah salah satu penyakit kronis yang ditandai dengan kadar glukosa darah yang tinggi akibat gangguan pada produksi *insulin* atau resistensi *insulin*. Penyakit ini berkembang perlahan dan umumnya terjadi pada orang dewasa, meskipun kini kasus pada remaja dan anak-anak semakin meningkat. Faktor risiko utama meliputi *obesitas*, gaya hidup tidak aktif, serta faktor *genetik*. Diagnosis dini dan penanganan yang tepat sangat penting untuk mencegah komplikasi serius, seperti penyakit jantung, kerusakan saraf, dan kerusakan ginjal Utomo *et al.* (2020).

A. Diabetes Tipe 1

Diabetes tipe 1 adalah penyakit *autoimun* dimana sistem kekebalan tubuh secara keliru menyerang sel *beta* di pankreas, menyebabkan kerusakan permanen sehingga sel-sel tersebut tidak bisa lagi memproduksi *insulin*. Penyebabnya meliputi faktor *genetik,* lingkungan, dan gaya hidup.

B. Diabetes Tipe 2

Diabetes tipe 2 terjadi karena resistensi *insulin*, dimana tubuh tidak merespons *insulin* dengan baik. Pankreas awalnya memproduksi lebih banyak *insulin* untuk mengatasi resistensi ini, tetapi akhirnya produksinya menurun, sehingga kadar gula darah meningkat. Faktor penyebab utamanya adalah *genetik*, kelebihan berat badan, obesitas, dan gaya hidup yang kurang aktif Wu *et al.* (2023).

C. Diabetes Gestasional

Diabetes gestasional muncul selama kehamilan akibat hormon yang mengganggu kerja *insulin*. Faktor risiko termasuk riwayat pradiabetes dan keluarga dengan diabetes. Sekitar 50% kasus diabetes gestasional dapat dikonfirmasi dari sumber yang terpercaya Butt *et al.* (2021).

2.2.2 Klasifikasi dalam Diagnosis Medis

Klasifikasi merupakan salah satu metode penting dalam diagnosis medis. Proses klasifikasi bertujuan untuk mengidentifikasi kategori atau kelas dari suatu data berdasarkan fitur-fitur tertentu. Dalam konteks diagnosis penyakit, klasifikasi digunakan untuk memKlasifikasi kemungkinan seseorang menderita suatu penyakit berdasarkan data medis, seperti riwayat kesehatan, tes laboratorium, dan faktor-faktor lain yang relevan Nurwijayanti *et al.* (2023).

2.2.3 Pembahasan Al-Qur'an dan Hadis Tentang Penyakit

Dalam Al-Qur'an, penyakit dibahas baik dalam konteks jasmani (fisik) maupun rohani (spiritual). Diabetes merupakan salah satu penyakit jasmani (fisik) dimana Al-Qur'an tidak hanya memberikan panduan tentang bagaimana menghadapi penyakit, tetapi juga menjelaskan hikmah di baliknya. Ayat-ayat dalam Al-Qur'an, seperti Surah Al-A'raf ayat 31, mengajarkan umat manusia untuk menjaga keseimbangan dalam hidup, termasuk dalam hal makan dan minum. Tafsir Ibnu Katsir menjelaskan bahwa larangan berlebih-lebihan ini berlaku untuk semua aspek kehidupan, dan salah satu bentuk israf (kelebihan) yang sering terjadi adalah dalam konsumsi makanan.

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

“Wahai anak cucu adam,pakailah pakayanmu yang indah pada setiap (memasuki) masjid dan makan serta minumla, tetapi jangan berleihan. Sesungguhnya Dia tidak menyukai orang-orang yang berlebihan” (Qs. Al-A'raf ayat 31,)

مَا أَنْزَلَ اللَّهُ دَاءً إِلَّا أَنْزَلَ لَهُ شِفَاءً

"Tidaklah Allah menurunkan suatu penyakit, melainkan Dia juga menurunkan penawarnya." (HR. Bukhari no. 5678, Muslim no. 2204)

Hadits ini menunjukkan bahwa setiap penyakit yang ada di dunia ini, termasuk diabetes melitus, pasti memiliki obat atau solusi. Ini mendorong umat Islam untuk aktif dalam riset dan pengembangan ilmu pengetahuan, termasuk melalui teknologi seperti machine learning dalam upaya deteksi dini dan klasifikasi penyakit.

مَنْ نَفَّسَ عَنْ مُؤْمِنٍ كُرْبَةً مِنْ كُرَبِ الدُّنْيَا، نَفَّسَ اللَّهُ عَنْهُ كُرْبَةً مِنْ كُرَبِ يَوْمِ الْقِيَامَةِ

"Barang siapa yang melepaskan satu kesusahan dari seorang mukmin dari

kesusahan-kesusahan dunia, maka Allah akan melepaskan darinya satu kesusahan dari kesusahan-kesusahan pada hari kiamat" (HR. Muslim, no. 2699)

Penerapan teknologi untuk mempermudah diagnosis dan klasifikasi penyakit adalah bentuk nyata menolong sesama**.** Dengan membantu pasien atau tenaga medis mengenali risiko diabetes secara lebih cepat dan akurat, maka peneliti ikut meringankan beban penderita. Islam sangat mengapresiasi kontribusi seperti ini karena berorientasi pada kemaslahatan umat**.**

Konsep tersebut sejalan dengan prinsip kesehatan modern yang menekankan pentingnya keseimbangan nutrisi serta penerapan sikap moderasi dalam berbagai aspek kehidupan. Dengan demikian, penelitian ini tidak hanya fokus pada aspek medis semata, tetapi juga hubungan dengan nilai-nilai keseimbangan yang menjadi dasar dalam menjaga kesehatan secara holistik, tetapi juga menghubungkannya dengan aspek spiritual yang dapat membantu dalam pengelolaan diabetes secara holistik *holistik* Care & Suppl. (2020).

A. Konsep muamalah ma'a allah (hubungan dengan Allah)

Penelitian ini bukan hanya aktivitas ilmiah biasa, tetapi merupakan bentuk ibadah dan pengabdian kepada Allah. Berikut penjabarannya:

1. Tawakal dan ikhtiar yang seimbang

وَإِذَا مَرِضْتُ فَهُوَ يَشْفِينِ

"Dan apabila aku sakit, Dialah yang menyembuhkan aku." (QS. Asy-Syu'ara': 80)

Ayat ini menanamkan keyakinan bahwa kesembuhan mutlak di tangan Allah. Namun, keyakinan ini tidak bertentangan dengan upaya penelitian. Justru,

penelitian Anda adalah sebab (wasilah) yang diciptakan dan dimudahkan oleh Allah untuk mencapai kesembuhan. Seorang peneliti Muslim meyakini bahwa algoritma *Random Forest* dan *Naive Bayes* yang berhasil dikembangkan adalah atas izin dan pertolongan-Nya. Ini mencegah kesombongan ilmiah dan senantiasa mengingatkan pada Sang Maha Penyembuh.

2. Penyakit sebagai bagian dari takdir dan ujian, serta kewajiban mencari obat

مَا أَنْزَلَ اللَّهُ دَاءً إِلَّا أَنْزَلَ لَهُ شِفَاءً

"Allah tidak menurunkan suatu penyakit, melainkan Dia turunkan juga obatnya." (HR. Al-Bukhari)

Hadis ini adalah landasan teologis utama bagi seluruh penelitian medis dan kesehatan dalam Islam. Diabetes adalah "disease" (penyakit) yang disebutkan dalam hadis ini. Tesis Anda yang bertujuan untuk mendeteksi diabetes secara lebih akurat adalah bagian dari proses "mencari obat" yang diperintahkan. Ini adalah bentuk ikhtiar yang sangat dianjurkan, sebagai wujud tawakal yang benar (berserah diri setelah berusaha maksimal).

B. Konsep muamalah ma'a an-nas (hubungan dengan sesama manusia)

1. Perintah tolong-menolong dalam kebaikan dan takwa (Al-Birr)

وَتَعَاوَنُوا عَلَى الْبِرِّ وَالتَّقْوَىٰ ۖ وَلَا تَعَاوَنُوا عَلَى الْإِثْمِ وَالْعُدْوَانِ

"Dan tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa." (QS. Al-Maidah: 2)

Menyelamatkan orang dari penyakit kronis adalah kebaikan (al-birr) yang sangat nyata. Kolaborasi antara dunia teknologi informasi (Anda sebagai peneliti Informatika) dengan dunia medis untuk menangani diabetes adalah bentuk ta'awun

'alal birri yang modern. Penelitian Anda adalah kontribusi nyata dari disiplin ilmu komputer untuk kemanusiaan.

2. Menjadi sebab terlepasnya kesusahan orang lain

مَنْ نَفَّسَ عَنْ مُؤْمِنٍ كُرْبَةً مِنْ كُرَبِ الدُّنْيَا، نَفَّسَ اللَّهُ عَنْهُ كُرْبَةً مِنْ كُرَبِ يَوْمِ الْقِيَامَةِ

"Barang siapa yang melepaskan satu kesusahan dari seorang mukmin di dunia, Allah akan melepaskan darinya satu kesusahan di hari kiamat." (HR. Muslim)

Seorang pasien yang tidak terdiagnosis diabetes atau terlambat diagnosis mengalami kesusahan fisik, mental, dan finansial. Model klasifikasi yang Anda bangun, dengan akurasi tinggi (96% untuk RF), dapat membantu tenaga medis mengurangi kesusahan ini dengan diagnosis yang lebih cepat dan akurat.

2.2.4 Algoritma *Naïve Bayes*

Naïve Bayes merupakan salah satu algoritma dalam bidang machine learning yang umum digunakan untuk menyelesaikan permasalahan klasifikasi. Algoritma ini berlandaskan pada penerapan Teorema *Bayes* dengan asumsi dasar yang bersifat “naif”, yaitu setiap fitur atau variabel input dalam model dianggap saling independen dan tidak memiliki hubungan ketergantungan satu sama lain. Meskipun asumsi ini seringkali tidak benar di dunia nyata, *Naive Bayes* tetap sangat efektif dalam banyak kasus praktik, terutama untuk klasifikasi teks, seperti *spam detection* atau *sentiment analysis* Lipsky *et al*. (2020).

Rumus *Naive Bayes* didasarkan pada Teorema Bayes:

$$P(C/X) = \frac{P(X/C) \cdot P(C)}{P(X)} \quad (1)$$

- *P(C/X)* = probabilitas kelas C diberikan fitur X
- *P(X/C)* = probabilitas mendapatkan fitur X diberikan kelas C
- *P(C)* = probabilitas dari kelas C (*prior probability*)
- *P(X)* = probabilitas dari fitur X (*evidence*)

Langkah-langkah penggunaan *Naive Bayes*: Pebdika *et al.* (2023)

1. *Preprocessing Data*:

- Bersihkan data (tangani *missing values*, hapus *noise*).
- Jika data berupa teks, lakukan tokenisasi, *stemming*, dan konversi ke representasi numerik.

2. Hitung *Probabilitas Prior* (*P(C)):*

- Probabilitas setiap kelas dihitung dari distribusi data pelatihan.

$$Jumlah\ data\ di\ kelas\ C\ \frac{P\ (c)}{Total\ Data} \quad (2)$$

3. Hitung *Likelihood (P(X/C)):*

- Untuk fitur kategorikal, hitung frekuensi kemunculan fitur dalam kelas tertentu.
- Untuk fitur kontinu, menggunakan distribusi Gaussian:

$$P\ (X|C = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma 2}\right) \quad (3)$$

4. Kalkulasi *Probabilitas Posterior (P(C/X)):*

- Gabungkan *prior* dan *likelihood*:

$$P\ (X|C\ \propto . \prod_{i=I}^{n} P\ (Xi|C) \quad (4)$$

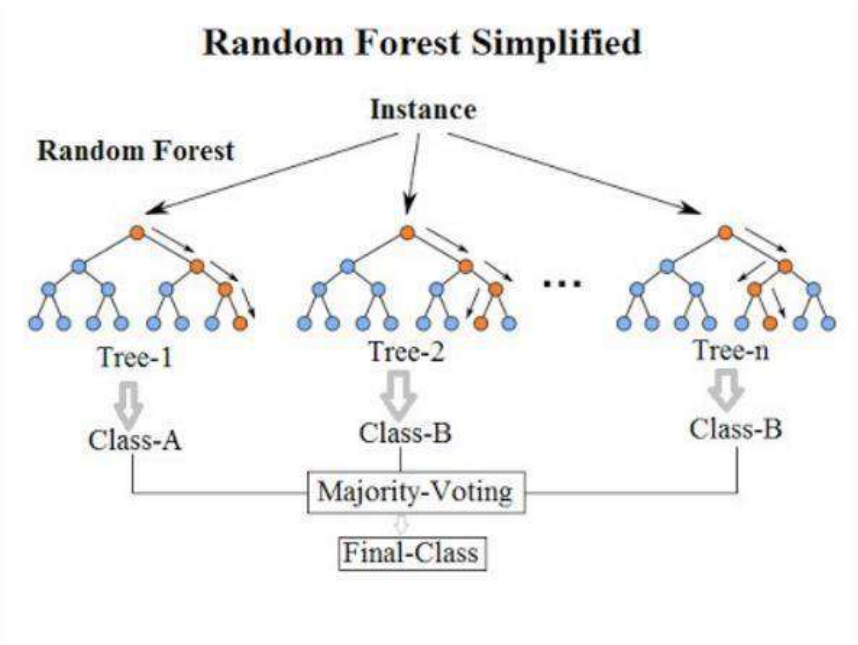
5. Klasifikasi:

- Pilih kelas *C* dengan probabilitas *P(C/X))* tertinggi. (5)

2.2.5 Algoritma *Random Forest*

Dikutip dari Carstensen *et al.* (2020) *Random Forest* adalah jenis algoritma *Supervised Learning* yang memanfaatkan *Decision Trees* untuk membuat Klasifikasi. *Random Forest* termasuk dalam *Ensemble Learning*, yaitu dimana

pendekatan dilakukan dengan beberapa algoritma yang digabungkan untuk meningkatkan hasil Klasifikasi. Dalam hal ini *Random Forest* menggunakan *Decision Tree.* Secara khusus, *Random Forest* digunakan untuk klasifikasi data kedalam kategori atau label yang telah ditentukan. *Random Forest*), *Decision Tree* yang dibangung akan menggunakan subset acak dari data pelatihan dengan penggantian sehingga menciptakan variasi yang membantu mengurangi *overfitting*. Setelah *Decision Trees* dibangun, Klasifikasi akhir diperoleh dengan *voting* dari semua *Decision Trees* Yesa *et al.* (2023).



Gambar 2.1 *Majority-Voting*
Pratama *et al.* (2023)

Berikut adalah langkah dan elemen penting dari algoritma Random Forest

1. Pemilihan Fitur

Setiap pohon dalam *Random Forest* dibangun menggunakan subset acak dari fitur. Jika total fitur adalah *M*, maka *subset* yang dipilih secara acak biasanya berjumlah *M* untuk klasifikasi.

$$F = RandomSubset\ (M) \quad (6)$$

2. *Entropy* atau *gini index* untuk *Spilt*

Setiap *node* pada pohon keputusan menggunakan fungsi berikut untuk memilih split terbaik berdasarkan fitur yang dipilih:

Entropi $Entropy\ (D) = -\sum_{i=I}^{n} pi \log\ (Pi)$ (7)

Gini Index $Gini\ (D) = 1 - \sum_{i=1}^{n} P_i^2$ ()

Dimana:

- *D*: *Dataset* pada node tertentu.
- *pi*: Proporsi data dari kelas iii dalam *D.*

Split terbaik adalah yang meminimalkan entropi atau Gini Index.

3. Klasifikasi *Voting* Mayoritas

Setelah semua pohon selesai dilatih, setiap pohon memberikan Klasifikasi *Yi* untuk data baru *X*. Klasifikasi akhir adalah voting mayoritas dari semua pohon:

$$y = mode\ (y1, y2 \ldots . yn) \quad (8)$$

Dimana:

- *n*: Jumlah total pohon.
- *mod*e: Nilai yang paling sering muncul (mayoritas).

Rumus Skor Akhir (Probabilitas untuk Setiap Kelas)

Jika ingin mendapatkan probabilitas untuk setiap kelas *Ck*, hitung rasio jumlah pohon yang memilih kelas tersebut terhadap total pohon:

$$P(Ck|X = \frac{Jumlah\ pohon\ yang\ memprdeksi\ \ Ck}{n}) \quad (9)$$

Menurut Pieske *et al.* (2020) kelebihan algoritma *Random Forest* adalah memiliki akurasi yang tinggi dalam klasifikasi, hal ini disebabkan karena *Random*

Forest menggabungkan banyak *decision tree* untuk menghasilkan hasil yang lebih baik. *Random Forest* juga memiliki kemampuan untuk mengatasi *overfitting*, hal ini karena metode pembagian data dan fitur secara acak yang membuat model lebih baik memKlasifikasi data baru. Serta memberikan *features importances* yang berguna membantu pengaruh masing-masing variabel dalam model.

2.2.6 *Naive Bayes* dan *Random Forest*

Setelah melakukan studi literatur, penulis memilih untuk membandingkan algoritma *Random Forest* dan *Naive Bayes* dalam Klasifikasi penyakit diabetes. Parameter yang digunakan dalam perbandingan ini meliputi *accuracy*, *precision,recall*, *F1-score*, dan *execution time*.

Accuracy merupakan tingkat keakuratan model terhadap Klasifikasi dibandingkan dengan nilai sebenarnya.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Precision adalah nilai proporsi Klasifikasi positif benar (*true positive*, TP) terhadap jumlah seluruh Klasifikasi positif.

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Rumus ini juga dapat dijelaskan sebagai berikut:

$$Precision = \frac{\text{Jumlah Klasifikasi benar (positif)}}{\text{jumlah Klasifikasi benar (positif)} + \text{ Jumlah Klasifikasi salah (positif)}}$$

Recall mengukur proporsi Klasifikasi positif benar terhadap total kasus positif, baik yang terKlasifikasi maupun yang tidak.

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

Rumus ini juga dapat dijelaskan sebagai berikut:

$$Precision = \frac{\text{Jumlah Klasifikasi benar (positif)}}{\text{umlah Klasifikasi benar (positif)} + \text{ Jumlah Klasifikasi salah (positif)}}$$

F1-Score merupakan rata-rata harmonis antara *precision* dan *recall*, memberikan keseimbangan antara keduanya.

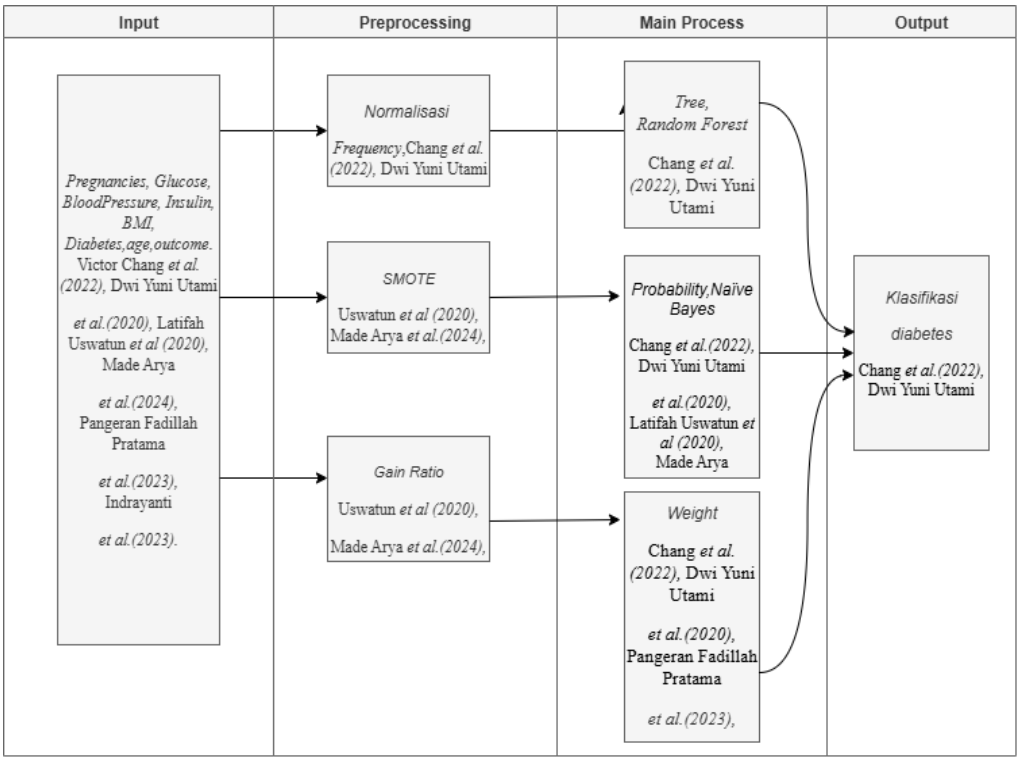
$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (14)$$

Execution Time adalah total waktu yang dibutuhkan oleh model untuk melakukan simulasi atau Klasifikasi. Parameter ini mengukur efisiensi waktu yang diperlukan untuk menyelesaikan tugas klasifikasi.

2.3 Kerangka Teori

Penelitian ini menggunakan data masukan seperti informasi pasien (usia, jenis kelamin, riwayat keluarga), parameter medis (kadar glukosa darah, tekanan darah), dan perilaku historis pasien (rekam medis sebelumnya). Melalui seleksi fitur, dua fitur utama yang digunakan adalah *recency* (waktu sejak pemeriksaan terakhir) dan *frequency* (frekuensi pemeriksaan kesehatan). Penelitian ini memanfaatkan empat pendekatan utama, yaitu metode berbasis pohon (*Random Forest*), berbasis jarak (*Euclidean Distance*), probabilistik (*Naive Bayes*), dan berbasis bobot (atribut penting seperti glukosa dan *BMI*). Dengan membandingkan kinerja *Naive Bayes* dan *Random Forest*, penelitian bertujuan untuk mengidentifikasi model terbaik

dalam klasifikasi diagnosis diabetes, sehingga dapat mendukung diagnosis dini dan pengelolaan penyakit secara lebih efektif. Dari studi literatur yang dilakukan diperoleh kerangka teori seperti terlihat pada Gambar 2.2.



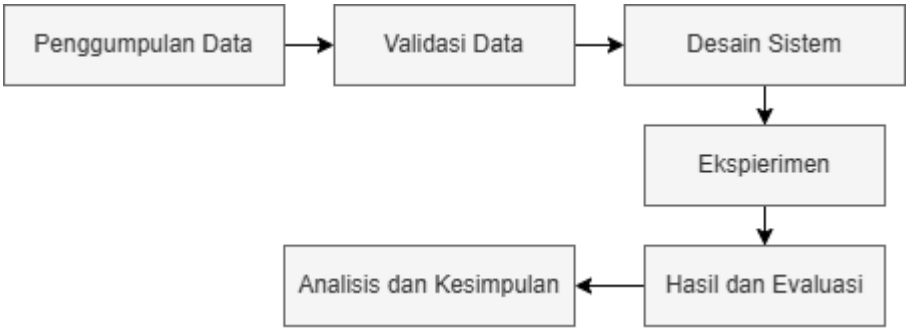
Gambar 2.2: Kerangka Teori

BAB III

METODOLOGI PENELITIAN

3.1 Prosedur Penelitian

Desain penelitian merupakan serangkaian tahapan sistematis yang disusun oleh peneliti untuk mencapai tujuan penelitian. Tahapan tersebut diawali dengan proses pengumpulan data, kemudian dilanjutkan dengan perancangan sistem yang memuat alur proses yang terjadi secara terstruktur dalam sistem. Proses-proses yang terdapat dalam rancangan tersebut selanjutnya diimplementasikan ke dalam sistem dengan tujuan utama membangun model agar mampu mempelajari pola dari data yang diberikan. Untuk memperoleh performa model terbaik serta mengidentifikasi faktor-faktor yang memengaruhi kinerjanya, diterapkan beberapa skenario eksperimen. Setiap skenario menghasilkan nilai performa tertentu sehingga dapat ditentukan skenario mana yang paling optimal. Tahapan akhir dalam desain penelitian adalah melakukan evaluasi dan analisis terhadap hasil eksperimen untuk menarik kesimpulan yang valid secara ilmiah. Seluruh rangkaian tahapan penelitian ini divisualisasikan dalam Gambar 3.1.



Gambar 3.1 Desain Penelitian

3.1.1 Pengumpulan Data

Proses pengumpulan data dilakukan dengan tujuan memperoleh informasi yang relevan guna mendukung pelaksanaan penelitian ini. Dalam upaya tersebut, penulis menggunakan metode studi pustaka sebagai teknik utama untuk mengumpulkan berbagai data dan informasi yang diperlukan. Salehi *et al.* (2020) Studi pustaka dilakukan dengan tujuan mengumpulkan berbagai informasi yang relevan dengan permasalahan yang menjadi fokus penelitian. Data dan informasi diperoleh melalui telaah terhadap penelitian-penelitian terdahulu serta jurnal-jurnal ilmiah yang dijadikan sebagai sumber referensi. Hasil dari kegiatan ini digunakan sebagai dasar dalam penyusunan landasan teori serta untuk membantu penulis dalam menemukan jawaban atas permasalahan yang diteliti. Adapun batasan penelitian dan jurnal yang penulis pilih yaitu: Liu *et al*. (2020).

a. Penelitian terkait yang digunakan oleh penulis yaitu penelitian dari 5 tahun terakhir dengan tema *Diabetes Melitus.*

b. Jurnal yang digunakan oleh penulis yaitu penelitian dari 5 tahun terakhir dengan tema *Diabetes Melitus.*

A. Atteribut Data

Atteribut data disebut juga atribut atau fitur, merupakan ciri khusus dari entitas data. Atribut data dalam *Data Science* atau klasifikasi merupakan variabel yang menjelaskan tentang catatan data atau contoh. Setiap atribut memiliki tipe tertentu data dan mewakili satu aspek dari entitas yang sedang dijelaskan. Detail dari atribut kumpulan data yang digunakan dalam penelitian ini ditunjukkan tabel 3.1.

Tabel 3.1 Data Atteribut

No	*Atteribut*	Tipe	Keterangan
1	*gender*	object	Jenis kelamin pasien (Male/Female)
2	*age*	float64	Usia pasien
3	*hypertension*	int64	Riwayat hipertensi (0 = tidak, 1 = ya)
4	*heart_disease*	int64	Riwayat penyakit jantung (0 = tidak, 1 = ya)
5	*smoking_history*	object	Riwayat merokok (never, current, etc.)
6	*bmi*	float64	Indeks massa tubuh
7	*HbA1c_level*	float64	Kadar HbA1c (hemoglobin terglikasi)
8	*blood_glucose_level*	int64	Kadar glukosa darah
9	*diabetes*	int64	Label target (0 = tidak diabetes, 1 = diabetes)

Untuk menemukan atribut yang tepat atau optimal bagi permasalahan tersebut, diperlukan klasifikasi menyeluruh terhadap semua atribut yang ada, serta mengabaikan atribut yang tidak relevan. *Dataset* input yang tercantum pada Tabel 3.1 mencakup berbagai atribut beserta deskripsinya. Pemilihan atribut yang sesuai dengan kualitasnya akan meningkatkan kualitas kumpulan data masukan dan hasil klasifikasi yang diharapkan.

3.1.2 Validasi Data

Pada tahap ini dilakukan pengecekan kualitas data untuk memastikan bahwa data bebas dari kesalahan, duplikasi, data kosong (missing values), serta memastikan distribusi data seimbang. Tahap validasi juga bertujuan untuk memastikan dataset layak digunakan dalam proses pelatihan model machine learning.

3.1.3 Desain Sistem

Tahapan ini berfokus pada perancangan model klasifikasi yang akan digunakan. Peneliti menetapkan algoritma yang akan diuji (*Naïve Bayes* dan *Random Forest*), arsitektur sistem, alur preprocessing data, teknik pembagian data (training dan testing), serta parameter evaluasi yang akan digunakan.

3.1.4 Eksperimen

Tahap eksperimen melibatkan proses pelatihan model (training) menggunakan data latih dan pengujian (testing) untuk melihat performa model. Pada tahap ini dilakukan pengujian algoritma untuk mengetahui sejauh mana masing-masing dapat mengklasifikasikan pasien sebagai diabetes atau non-diabetes.

3.1.5 Hasil dan Evaluasi

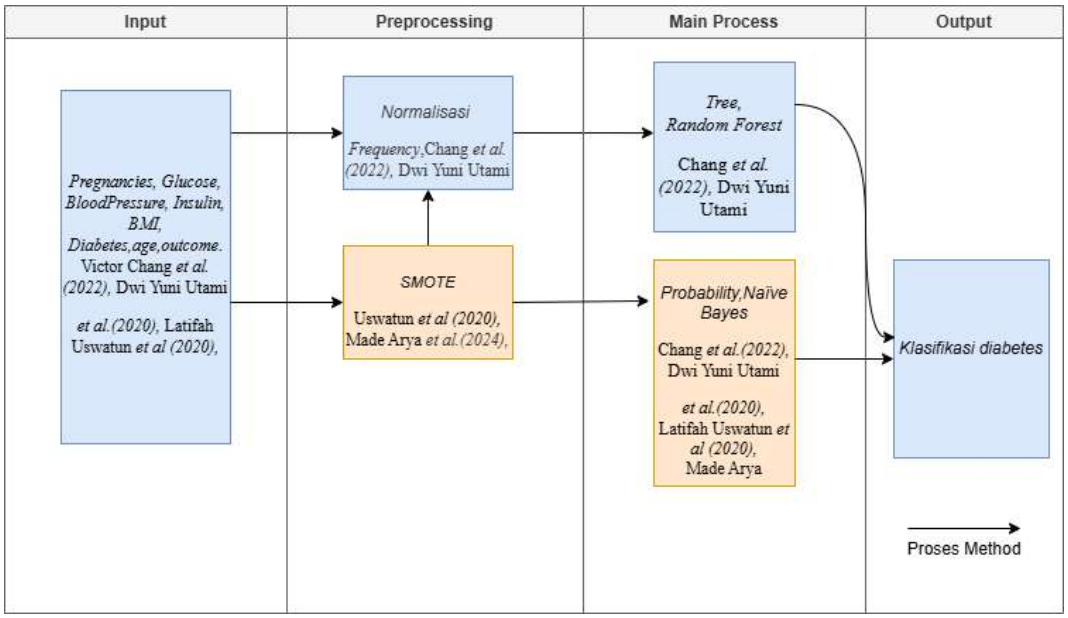
Output dari eksperimen dianalisis menggunakan metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan confusion matrix. Tahap ini menentukan algoritma mana yang memberikan hasil terbaik dan paling akurat dalam memprediksi diabetes.

3.1.6 Analisis dan Kesimpulan

Tahap akhir adalah melakukan analisis mendalam terhadap hasil evaluasi dan menarik kesimpulan berdasarkan perbandingan performa kedua algoritma. Kesimpulan ini menjadi dasar rekomendasi penggunaan algoritma terbaik dalam sistem pendukung keputusan medis.

3.2 Kerangka Konseptual

Kerangka konseptual menggambarkan tahapan proses serta keterkaitan antar komponen yang terlibat dalam penelitian ini. Melalui kerangka tersebut, alur mulai dari data masukan, proses prapengolahan, proses utama (yang mencakup ekstraksi fitur dan klasifikasi menggunakan algoritma *Naïve Bayes* dan *Random Forest*), hingga menghasilkan keluaran dapat dijelaskan secara sistematis. Kerangka ini memberikan pemahaman menyeluruh mengenai hubungan antar elemen dalam mencapai tujuan akhir, yaitu klasifikasi penyakit Diabetes Mellitus. Kerangka konsep disajikan pada Gambar 3.2.



Gambar 3.2 Kerangka Konseptual

Kerangka konseptual pada penelitian ini menggambarkan alur sistematis dalam proses klasifikasi diabetes mellitus berdasarkan data klinis yang terdiri dari variabel *pregnancies, glucose, blood pressure, insulin, BMI, age,* serta *outcome*. sehingga memiliki validitas yang kuat sebagai dasar pemodelan. Tahap preprocessing dilakukan untuk memastikan kualitas data yang optimal melalui

penerapan proses normalisasi guna menyamakan skala pada setiap fitur. Selain itu, digunakan pula teknik *Synthetic Minority Oversampling Technique* (SMOTE) sebagai upaya untuk mengatasi permasalahan ketidakseimbangan kelas pada data yang digunakan dalam penelitian. Langkah ini penting agar model machine learning dapat bekerja secara optimal dalam mengenali pola dan menghasilkan prediksi yang lebih akurat. Pada tahap utama (main process), data yang telah diproses kemudian dimodelkan menggunakan dua algoritma klasifikasi, yaitu *Naïve Bayes* dan *Random Forest*. Algoritma *Naïve Bayes* digunakan karena memiliki keunggulan dalam perhitungan probabilistik yang sederhana dan efisien, sedangkan *Random Forest* dipilih karena kemampuannya dalam menangani kompleksitas fitur dan menghasilkan prediksi yang stabil melalui pendekatan ensemble. Hasil dari kedua model ini selanjutnya dievaluasi menggunakan metrik akurasi, precision, recall, dan *F1-score* untuk menentukan algoritma terbaik dalam klasifikasi diabetes. Output akhir dari kerangka konseptual ini adalah keputusan klasifikasi "diabetes" atau "tidak diabetes", yang menjadi dasar untuk menarik kesimpulan dan memberikan rekomendasi berbasis kecerdasan buatan dalam mendukung diagnosis dini penyakit diabetes mellitus.

3.3 Skenario Uji Coba

Tujuan skenario uji coba adalah rancangan atau skema pengujian yang dibuat untuk memastikan bahwa sistem atau model yang dikembangkan diuji secara terstruktur, terukur, dan objektif berdasarkan tujuan penelitian. Dalam konteks penelitian Anda yang menggunakan algoritma *Random Forest* dan *Naïve Bayes* untuk klasifikasi penyakit Diabetes Mellitus, skenario uji coba menjelaskan

bagaimana proses pengujian dilakukan, data apa yang digunakan, tahapan evaluasinya, dan kriteria keberhasilan model.

Accuracy merupakan tingkat keakuratan model terhadap Klasifikasi dibandingkan dengan nilai sebenarnya.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision adalah nilai proporsi Klasifikasi positif benar (*true positive*, TP) terhadap jumlah seluruh Klasifikasi positif.

$$Precision = \frac{TP}{TP+FP}$$

Rumus ini juga dapat dijelaskan sebagai berikut:

$$Precision = \frac{\text{Jumlah Klasifikasi benar (positif)}}{\text{jumlah Klasifikasi benar (positif)} + \text{ Jumlah Klasifikasi salah (positif)}}$$

Recall mengukur proporsi klasifikasi positif benar terhadap total kasus positif, baik yang terklasifikasi maupun yang tidak.

$$Recall = \frac{TP}{TP+FN}$$

Rumus ini juga dapat dijelaskan sebagai berikut:

$$Precision = \frac{\text{Jumlah Klasifikasi benar (positif)}}{\text{umlah Klasifikasi benar (positif)} + \text{ Jumlah Klasifikasi salah (positif)}}$$

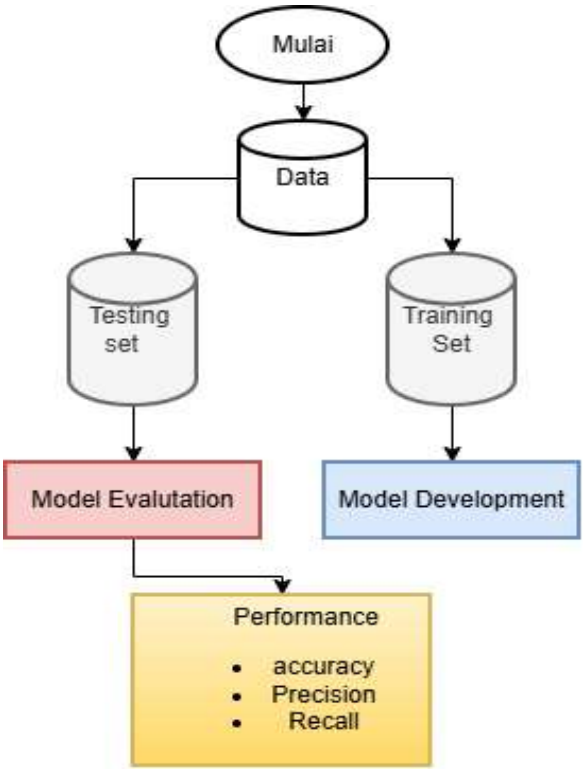
F1-Score merupakan rata-rata harmonis antara *precision* dan *recall*, memberikan keseimbangan antara keduanya.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Execution Time adalah total waktu yang dibutuhkan oleh model untuk melakukan simulasi atau klasifikasi. Parameter ini mengukur efisiensi waktu yang diperlukan untuk menyelesaikan tugas klasifikasi.

3.4 Skenario Pengujian Hasil

Skenario pengujian hasil bertujuan untuk mengevaluasi performa algoritma yang digunakan dalam penelitian, yaitu *Naïve Bayes* dan *Random Forest,* dalam mengklasifikasikan penyakit Diabetes Mellitus berdasarkan data pasien. Pengujian dilakukan melalui beberapa skenario untuk memastikan tingkat akurasi, keandalan, dan konsistensi model.

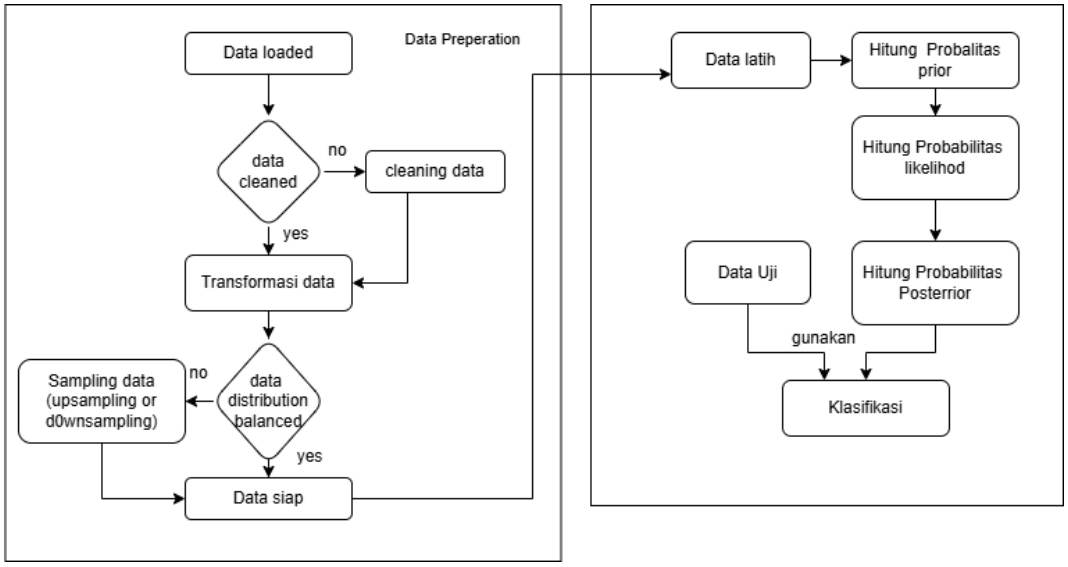


Gambar 3.3 Pengujian Hasil

BAB IV

IMPLEMENTASI METODE *NAÏVE BAYES*

4.1 Desain Sistem Naïve Bayes



Gambar 4.1 *Flowchart Naïve Bayes*
Fakhri *et al.* (2023)

1. Tahap Data Preparation (Persiapan Data)

a. Data Loaded (Data Dimuat)

Pada tahap awal, data diabetes dimasukkan ke dalam sistem (melalui Python atau aplikasi pengolahan data lainnya). Data ini masih dalam bentuk mentah (raw data).

b. Data Cleaned (Pembersihan Data)

Flowchart menunjukkan pengambilan keputusan: apakah data sudah bersih atau belum?

Jika belum, maka dilakukan proses cleaning data seperti:

- Menghapus data duplikat

- Mengatasi missing value,
- Mengatasi data yang tidak konsisten.
- Jika sudah bersih, masuk ke tahap selanjutnya.

c. Transformasi Data

Data mentah selanjutnya diubah menjadi bentuk yang dapat diproses oleh algoritma machine learning. Transformasi yang dilakukan dapat berupa:

- Normalisasi atau standardisasi numerik,
- Encoding pada fitur kategorikal,
- Scaling fitur agar model bekerja optimal.

d. Pengecekan Distribusi Data

- Pada langkah ini dicek apakah distribusi kelas (misalnya: diabetes positif dan negatif) sudah seimbang (balanced).
- Jika tidak seimbang, dilakukan sampling data:
- Oversampling jika data minoritas lebih sedikit,
- Undersampling jika data mayoritas terlalu dominan.
- Jika distribusi sudah seimbang, maka data dianggap siap digunakan (data siap).

2. Tahap Modeling dengan Algoritma Naïve Bayes

a. Pemisahan Data

Data yang telah siap dibagi menjadi:

- Data latih (training data): digunakan untuk membangun model.
- Data uji (testing data): digunakan untuk menguji performa model.

b. Perhitungan Probabilitas Prior

Probabilitas prior dihitung dari proporsi jumlah kelas dalam data latih. Misalnya:

- P(diabetes) dan P(tidak diabetes).

  Ini adalah tingkat kemungkinan suatu kelas muncul sebelum mempertimbangkan fitur lainnya.

c. Perhitungan Probabilitas Likelihood

Likelihood menghitung probabilitas kemunculan fitur tertentu berdasarkan masing-masing kelas.

d. Perhitungan Probabilitas Posterior

Menggunakan Teorema Bayes, probabilitas prior dan likelihood digabungkan untuk menghitung probabilitas akhir (posterior).

Inilah inti prediksi:

Posterior = prior × likelihood

e. Proses Klasifikasi

- Data uji dimasukkan ke dalam model,
- Model menggunakan probabilitas posterior untuk menentukan apakah pasien termasuk kategori diabetes atau tidak diabetes,
- Hasil akhirnya berupa prediksi klasifikasi.

4.2 Implementasi *Naive Bayes*

Adalah algoritma yang digunakan untuk klasifikasi dengan pendekatan probabilistik. Algoritma ini efektif karena hanya memerlukan jumlah data yang sedikit untuk menentukan parameter klasifikasi. Berikut tahapan untuk melakukan perhitungan menggunakan algoritma *Naive Bayes* dengan jumlah data sebanyak

100.000. Tahapan dalam menggunakan algoritma *Naive Bayes* adalah sebagai berikut:

1. Teorema *Naive Bayes*

$$P(X \mid C) = \frac{P(C\ I\ X\ ).P(C)}{P(X)}$$

- *P(C/X)* = probabilitas kelas *C* diberikan fitur *X*
- *P(X/C)* = probabilitas mendapatkan fitur *X* diberikan kelas *C*
- *P(C)* = probabilitas dari kelas *C* (*prior probability*)
- *P(X)* = probabilitas dari fitur *X* (*evidence*)

2. Menghitung Probabilitas Prior

   Dihitung berdasarkan jumlah data dalam setiap kelas dibandingkan total sampel.

$$P(diabetes = yes) = \frac{jumlah\ yes}{total\ sampel}\frac{8500}{100000} = 0.085$$

$$P(diabetes = no) = \frac{jumlah\ no}{total\ sampel}\ \frac{91500}{100000} = 0.915$$

3. Menghitung Mean (μ)

   Untuk kelas "yes":

   μumur, yes $= \frac{44..+\cdots+61}{8500} = 70{,}64$

   μbmi, yes $= \frac{19.31+\cdots+30.11}{8500} = 33.98$

   μHbA1c_level, yes $= \frac{65+\cdots+62}{8500} = 6.93$

   μblood, yes $= \frac{200+\cdots+250}{8500} = 194.09$

   Untuk kelas "no":

   μumur, no $= \frac{44..+\cdots+61}{91500} = 50{,}64$

   μbmi, no $= \frac{19.31+\cdots+30.11}{91500} = 28.84$

$\mu HbA1c\_level, no = \frac{65+\cdots+62}{91500} = 5.93$

$\mu HbA1c\_level, no = \frac{200+\cdots+250}{91500} = 143.66$

4. *Menghitung Likelihood P(X | C)*

Likelihood dihitung menggunakan distribusi *Gaussian*:

- x adalah nilai dari fitur Xi.
- μC adalah mean dari fitur Xi untuk kelas C.
- $\sigma 2$ adalah variansi dari fitur Xi untuk kelas C.

$$P\ (X_i\ = x\ |C) = \frac{1}{\sqrt{2\pi\sigma^2}}\ .\exp\left(\frac{(x-\ \mu)2}{2\sigma 2}\right)$$

Untuk kelas "yes":

$$P(umur\ =\ 80|yes) = \frac{1}{\sqrt{2\pi\cdot 211{,}67}} = exp\ \left(\frac{(80-60{,}94)2}{2.211{,}67}\right) =\ 0{,}011757$$
$$P(bmi\ =\ 23{,}27|yes) = \frac{1}{\sqrt{2\pi\cdot 57{,}12}} = exp\ \left(\frac{(27.32\ -\ 31.9)2}{2*57.12}\right) =\ 0{,}044148$$
$$P(\ HbA1c = 5{,}7\ |\ yes\ ) = \frac{1}{\sqrt{2\pi\cdot 1{,}158}} = exp\ \left(\frac{(57\ -\ 5{,}382)2}{2*094}\right)\ =\ 0{,}19402$$
$$P(blood = 85\ |\ yes = \frac{1}{\sqrt{2*\pi*3438.37}} = exp\ \left(\frac{(85\ -\ 194{,}09)2}{2*3438} =\ 0{,}001217\right)$$

Untuk kelas "no":

$$P(umur\ =\ 80|no) = \frac{1}{\sqrt{2\pi\cdot 500{,}28}} = exp\ \left(\frac{(80-40{,}2}{2.500{,}28}\right) = 0{,}00368$$
$$P(bmi\ =\ 23{,}27|no) = \frac{1}{\sqrt{2\pi\cdot 39{,}56}} = exp\ \left(\frac{(27.32\ -\ 31.9)2}{27{,}32-26{,}87}\right) =\ 0{,}063293$$
$$P(\ HbA1c = 5{,}7\ |\ no\ ) = \frac{1}{\sqrt{2\pi\cdot 2.0{,}94}} = exp\ \left(\frac{(57\ -\ 5{,}382)2}{2*094}\right)\ =\ 0{,}3905$$
$$P(blood = 85\ |\ no = \frac{1}{\sqrt{2*\pi*1199{,}75}} = exp\ \left(\frac{(85\ -\ 132{,}09)2}{2*1199{,}75} =\ 0{,}004545\right)$$

5. Menghitung Posterior

$$P(C\ |\ X) \propto P(C)\cdot P(X\ |\ C)$$

Untuk kelas diabetes"yes":

$P(ya \mid X) \propto P(ya) \cdot P(umur = 85 \mid ya) \cdot P(bmi = 27.32 \mid ya) \cdot P(HbA1c = 5.7 \mid ya) \cdot P(blood\_glucose = 85 \mid ya)$

$0.4672 \cdot 0.11757 \cdot 0.044148 \cdot 0.19402 \cdot 0.00121$

$= \mathbf{5.7267 * 10{-}8}$

Untuk kelas diabetes "no":

$P(tidak \mid X) \propto P(tidak) \cdot P(umur = 85 \mid tidak) \cdot P(bmi = 27.32 \mid tidak) \cdot P(HbA1c = 5.7 \mid tidak) \cdot P(blood\_glucose = 85 \mid tidak)$ $0.53273 \cdot 0.00368 \cdot 0.0632 * 0.390 \cdot 0.00454$

$= \mathbf{2.1938 * 10{-}7}$

6. Memilih Kelas dengan Probabilitas

Tertinggi Karena $P(tidak \mid X)$ lebih besar daripada $P(ya \mid X)$, maka hasil klasifikasi dari nilai sampel yang baru dimasukkan adalah tidak diabetes.

Tabel 4.1: Tabel Probabilitas Tes Sampel

	Ya	**Tidak**
Prior-Probabilitas	0.46726	0.53274
Age	0.01176	0.00368
BMI	0.04415	0.06320
HbA1c_level	0.19402	0.39050
Blood_Glucose_Level	0.00122	0.004545
Posterior Probabilitas	**5.7267 x 10^{-8}**	**2.1G38 x 10^{-7}**

4.3 Uji Coba *Niave Bayes*

Uji coba model Naïve Bayes dilakukan untuk mengukur kinerja model dalam mengklasifikasikan data diagnosis penyakit diabetes. Pengujian ini penting untuk menilai sejauh mana model mampu mengenali pola dan memklasifikasi kelas target

secara akurat. Dalam uji coba ini, digunakan data uji yang sebelumnya telah dipisahkan dari data pelatihan agar evaluasi lebih objektif.

4.3.1 Split Dataset

Dataset pada penelitian ini dibagi menjadi dua bagian utama, yaitu data training dan data testing, dengan proporsi masing-masing sebesar 70% untuk training dan 30% untuk testing. Proporsi ini ditentukan menggunakan parameter test-size = 0.3, yang bertujuan untuk memastikan sebagian besar data digunakan dalam proses pelatihan model, sementara sisanya dialokasikan untuk menguji kinerja model yang telah dilatih. Pembagian dilakukan secara acak dengan menetapkan parameter random-state = 42, sehingga hasil pembagian akan konsisten setiap kali kode dijalankan ulang, asalkan nilai random-state tidak diubah.

4.3.2 Training Model

Data yang telah diproses kini siap untuk dilatih menggunakan algoritma *Naïve Bayes*. Berbeda dengan *Random Forest* yang memerlukan parameter seperti n_estimators, *Naïve Bayes* membangun model berdasarkan prinsip probabilitas sederhana. Algoritma ini menghitung peluang dari setiap fitur terhadap target secara langsung untuk membuat klasifikasi, sehingga cocok untuk kasus klasifikasi dengan asumsi independensi antar fitur.

4.3.3 Model dan Akurasi

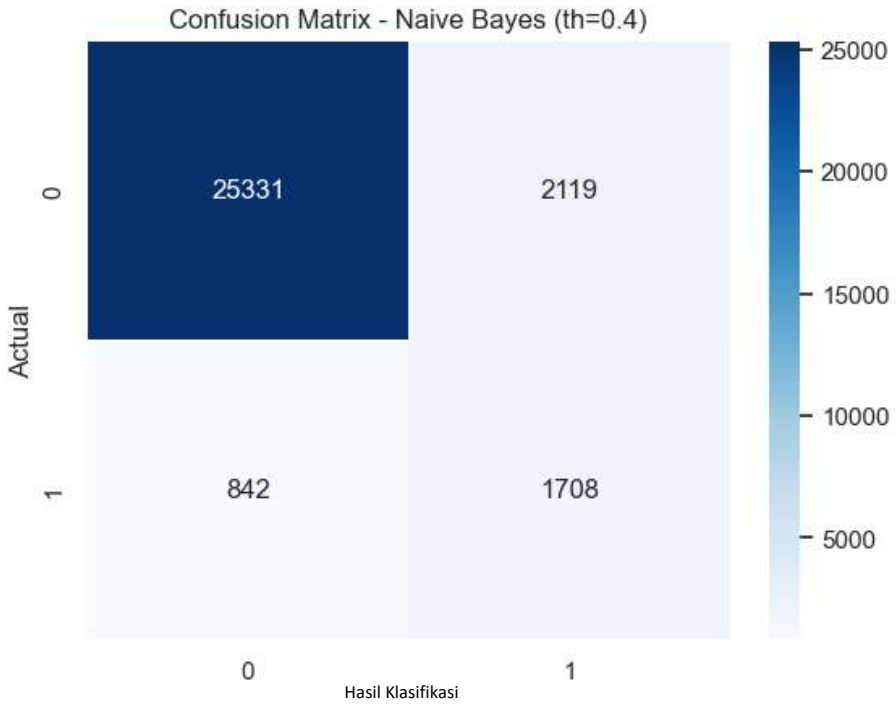
Setelah proses training, selanjutnya model akan melakukan klasifikasi dengan data test. Setelah pelatihan selesai, model akan dievaluasi menggunakan metrik *accuracy, F1-score, recall, dan precision*. Evaluasi ini dilakukan untuk menilai

seberapa baik model dalam membuat Klasifikasi. Sama seperti pada *Random Forest* (RF), fungsi yang digunakan untuk proses evaluasi adalah *classification_report dan accuracy_score dari library sklearn*.

Tabel 4.2 Hasil Akurasi NB

	precision	recall	f1-score	support
0	0.97	0.92	0.94	27450
1	0.45	0.67	0.54	2550
accuracy			0.90	30000
macro avg	0.71	0.80	0.74	30000
weighted avg	0.92	0.90	0.91	30000

Tabel 4.2 tersebut menampilkan laporan evaluasi model *Naïve Bayes* dengan threshold 0.4. Model ini mencapai akurasi sebesar 90,13%, yang menunjukkan bahwa sebagian besar klasifikasi sesuai dengan label aktual. Untuk kelas 0 (bukan diabetes), precision dan recall tinggi (masing-masing 0,97 dan 0,92) menandakan model sangat baik dalam mengenali data negatif. Namun, untuk kelas 1 (positif diabetes), precision hanya 0,45 meskipun recall mencapai 0,67, yang berarti model cukup mampu mendeteksi pasien positif, tetapi banyak juga klasifikasi positif yang keliru. Nilai macro average F1-score sebesar 0,74 menunjukkan bahwa performa antar kelas masih timpang, meskipun weighted average F1-score yang tinggi (0,91) mengindikasikan performa keseluruhan tetap baik karena didominasi oleh kelas mayoritas.



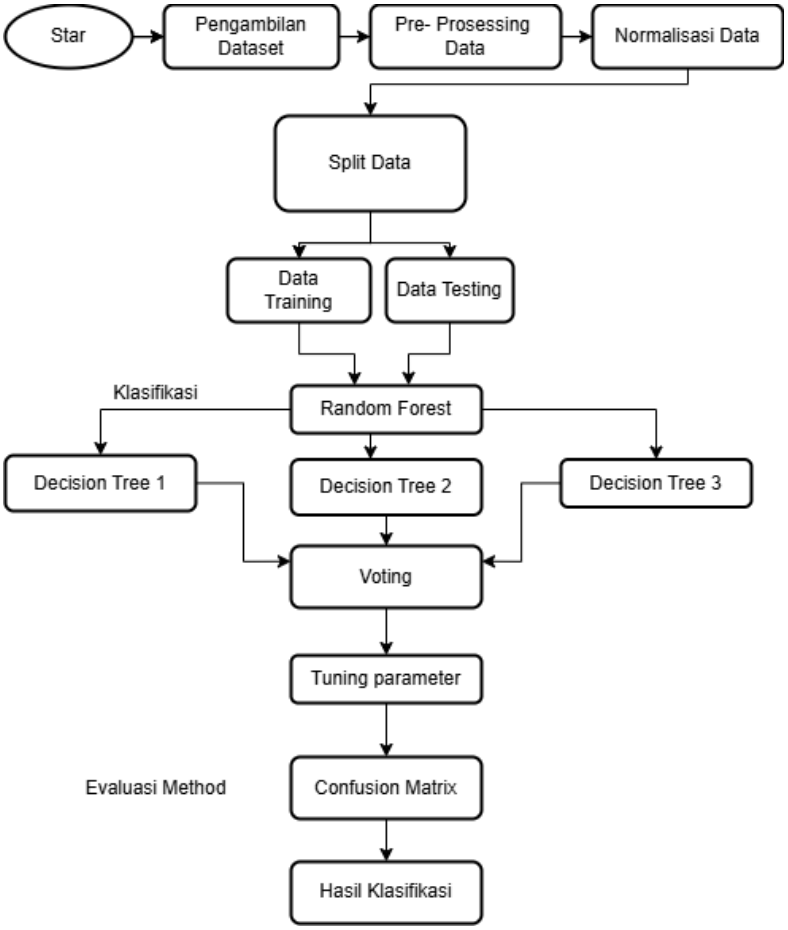
Gambar 4.2 Confusion Matrix NB

Gambar confusion matrix menunjukkan kinerja model *Naïve Bayes* dengan threshold 0.4 dalam mengklasifikasikan data diabetes. Dari 29.000 lebih data uji, model berhasil mengklasifikasikan 25.331 data negatif (True Negative) dan 1.708 data positif (True Positive) dengan benar. Namun, terdapat 2.119 kasus False Positive (data sehat yang diKlasifikasi sakit) dan 842 kasus False Negative (penderita yang tidak terdeteksi). Meskipun model menunjukkan akurasi keseluruhan yang baik, jumlah False Positive yang cukup tinggi menandakan bahwa model cenderung memberikan Klasifikasi positif secara berlebihan pada threshold ini, yang bertujuan meningkatkan sensitivitas deteksi kasus positif.

BAB V

IMPLEMENTASI METODE *RANDOM FOREST*

5.1 Desain Sistem *Random Forest*



Gambar 5.1 *flowchat Random Forest*
Sunyoto & Fatta. (2023)

1. Mulai

Proses dimulai dengan tahap inisialisasi seluruh langkah kerja penelitian.

2. Impor Dataset

Dataset yang berisi data pasien atau data penelitian dimasukkan dari sumber eksternal (seperti file CSV, Excel, atau database) untuk diproses lebih lanjut.

3. Feature Selection (Pemilihan Fitur)

Tahap ini memilih atribut atau variabel yang paling relevan terhadap hasil klasifikasi. Tujuannya agar model bekerja lebih efisien dan akurat dengan menghilangkan fitur yang tidak penting.

4. Normalisasi Data

Data diubah skalanya agar semua fitur berada pada rentang nilai yang sama. Hal ini penting agar tidak ada fitur yang mendominasi perhitungan jarak atau bobot pada algoritma.

5. Split Data

Data Train (latih): digunakan untuk membangun model.

Data Test (uji): digunakan untuk menguji performa model yang telah dilatih.

6. Bangun Beberapa Decision Tree

Pada tahap ini, algoritma Random Forest membangun sejumlah decision tree (pohon keputusan). Setiap pohon keputusan (*decision tree*) dalam algoritma dilatih menggunakan subset data serta subset fitur yang dipilih secara acak.

7. Lakukan Voting Mayoritas

Setelah semua pohon membuat prediksi, hasil akhirnya ditentukan dengan voting mayoritas, yaitu kelas yang paling banyak dipilih oleh pohon-pohon tersebut.

8. Model RF Terlatih

Hasil dari proses di atas adalah model Random Forest yang siap digunakan untuk mengklasifikasikan data uji.

9. Evaluasi Model

Model yang telah dibangun dievaluasi menggunakan data uji. Tahap ini melihat seberapa baik model mengenali data baru.

10. Confusion Matrix

Digunakan untuk menampilkan hasil klasifikasi dalam bentuk matriks yang berisi: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN)

11. Hitung Nilai Akurasi, Presisi, Recall, dan F1-score

Berdasarkan hasil confusion matrix, dilakukan perhitungan terhadap sejumlah metrik evaluasi untuk menilai kinerja model klasifikasi. Metrik tersebut meliputi: accuracy, yang menunjukkan tingkat ketepatan keseluruhan prediksi model; precision, yang mengukur proporsi prediksi positif yang benar; recall, yang menggambarkan kemampuan model dalam mengidentifikasi data positif secara akurat; serta F1-score, yang merepresentasikan keseimbangan antara nilai precision dan recall.

12. Tuning Parameter

Jika hasil evaluasi belum optimal, parameter model seperti jumlah pohon (n_estimators) atau kedalaman maksimum pohon (max_depth) disesuaikan agar performa meningkat.

13. Hasil Klasifikasi

Tahap akhir menghasilkan keluaran berupa hasil prediksi dari model *Random Forest* — misalnya apakah seseorang terdiagnosis diabetes atau tidak berdasarkan data masukan.

5.2 Implementasi *Random Forest*

Random Forest adalah algoritma ensemble learning yang menggabungkan banyak decision tree untuk meningkatkan akurasi klasifikasi dan mengurangi overfitting. Dalam klasifikasi, *Random Forest* melakukan klasifikasi berdasarkan voting mayoritas dari seluruh pohon. Dengan jumlah data sebanyak 100.000. Tahapan dalam menggunakan algoritma *Random Forest* adalah sebagai berikut:

1. Persiapan Data

a. Memahami *Dataset*:

- *Dataset* terdiri dari 100.000 baris data, masing-masing dengan fitur (atribut) dan label (kelas yang ingin diKlasifikasi).
- Misalnya: *Dataset* untuk Klasifikasi diabetes dengan atribut seperti usia, berat badan, dan kadar glukosa.

b. Membersihkan Data:

- Tangani *missing values* (isi dengan mean/median, hapus, atau metode lain).
- Ubah data katagorik menjadi numerik (misalnya, *encoding* atau *one-hot encoding*).

c. Normalisasi/Standarisasi :

- *Random Forest* tidak sensitif terhadap skala data, sehingga proses ini tidak wajib.

d. Pisahkan *Dataset*:

- Bagi *dataset* menjadi data pelatihan (*training*) dan pengujian (*testing*), misalnya:
- 70% untuk pelatihan (70.000 data).
- 30% untuk pengujian (30.000 data).

1. Konfigurasi Model *Random Forest*

a. *Bootstrap Sampling*:

- *Random Forest* akan membuat ***n*** pohon keputusan (misalnya, n =100 n = 100 n =100).

a. Pelatihan Pohon

Setiap pohon keputusan dilatih menggunakan subset data pelatihan yang dipilih secara acak dengan metode pengambilan sampel dengan penggantian (bootstrap sampling) dari total 10.000 data yang tersedia.

b. Pemilihan Fitur Secara Acak

Pada setiap pohon, dilakukan pemilihan subset fitur secara acak dari keseluruhan fitur yang tersedia dalam dataset.

1. Untuk setiap data uji dari 3.600 data:

- Masukkan data ke masing-masing pohon.
- Setiap pohon memberikan klasifikasi kelas ($y\hat{}i$).
- Lakukan voting mayoritas dari semua pohon:

$$y\hat{} = \text{mode}(y1, y2, \ldots, yn)$$

2. Jika diperlukan probabilitas, hitung rasio pohon yang memklasifikasi kelas tertentu:

$$P(Ck|X) = \frac{Jumlah\ pohon\ yang\ mempredeksi}{N}$$

4. Evaluasi Model

Gunakan data uji untuk mengukur performa model:

1. Metode Evaluasi:

$$Accuracy = \frac{Jumlah\ predeksi\ benar}{Total\ Data\ Uji}$$

a. Akurasi (*Accuracy*):
b. *Precision, Recall, F1-Score*.
c. *ROC-AUC* untuk *dataset* yang tidak seimbang.

Tabel 5.1 *bootstrap sampling data*

No	Umur	BMI	HbA1c	Glucose	Diabetes	RAND	Tree1	Tree2	Tree3	Voting	Y
1	45	29	6.2	150	yes	0.73	yes	no	yes	yes	1
2	54	2732	66	80	No	0.65	no	yes	no	no	0
3	42	3364	48	145	no	0.60	no	yes	no	no	0
4	28	2732	57	158	Yes	0.67	yes	no	yes	yes	1
5	36	2345	5	155	No	0.76	no	yes	no	no	0
6	76	2014	48	155	Yes	0.78	yes	no	yes	yes	1
.	..	...	...	...	...	...	...	...	...	....	...
.	..	...	..	..	...	...	...	...	..	..	..
...	...	...	...	...	...	....	...	...	...	...	...
10000	76	2014	48	155	Yes	0.74	yes	no	yes	yes	1

Berdasarkan banyaknya data sampel yang digunakan pada tabel diatas terdapat 61500 orang yang positif diabetes dan 8500 orang yang negatif diabetes. Maka peluang dari variabel respon untuk katagori positif dan negatif yaitu :

$$P(y = no\ \frac{n(y = no)}{n_{total}}\ \frac{91500}{100.000} = 0.915)$$

$$P(y = yes\ \frac{n(y = yes)}{n_{total}}\ \frac{8500}{100.000} = 0.085)$$

Selanjutnya akan dibentuk node untuk pohon pertama dengan mencari nilai entropy dan juga gain untuk variabel yang ditentukan secara acak.

$$Entropy\ (Y) = -[\ P\ (\text{no})\ \cdot\ \log 2\ P\ (\text{no})\ +\ P\ (\text{yes})\ \cdot\ \log 2\ P\ (\text{yes})\ ]$$
$$= -[0.8786 \cdot \log 2(0.8786) + 0.1214 \cdot \log 2(0.1214)]$$

Hitung logaritmanya:

$$Log_2(0.8786) = -0.1834$$
$$Log_2(0.1214) = -3.043$$

Maka:

$$Entropy(\text{Y}) = -[(0.8786 \cdot -0.1834) + (0.1214 \cdot -3.043)]$$
$$= -[-0.1611 - 0.3695]$$
$$= 0.5306$$

$$\text{Gain}(Y, HbA1c = 0.419 - \left(\frac{20000}{100000}\ 0.1500 + \frac{30000}{100000}\ \cdot 0.5000 + \frac{50000}{100000}\ 0.6000)\right)$$

$$= 0.4196 - (0.2 \cdot 0.1500 + 0.3 \cdot 0.5000 + 0.5 \cdot 0.6000)$$
$$= 0.4196 - (0.0300 + 0.1500 + 0.3000)$$

$$= 0.4196 - 0.4800 = -0.0604$$

Tabel 5.2 *confusion matrix metode Random Forest*

	True 1	True 0
Predict 1	1813	737
Predict 0	252	27198

Berdasarkan hasil confusion matrix, model *Random Forest* menunjukkan kinerja yang cukup baik dalam klasifikasi data. Dari total data uji, model berhasil mengklasifikasikan 1813 kasus positif (penderita diabetes) secara benar (True Positive) dan 1813 kasus negatif secara benar (True Negative). Namun, masih terdapat 252 kasus positif yang salah diklasifikasikan sebagai negatif (False

Negative), yang berisiko karena penderita sebenarnya tidak terdeteksi oleh model. Sementara itu, terdapat 131 kasus negatif yang salah diklasifikasikan sebagai positif (False Positive). Dengan demikian, meskipun akurasi model tergolong tinggi, kesalahan klasifikasi pada kasus positif perlu diperhatikan lebih lanjut, terutama jika digunakan dalam konteks medis seperti diagnosis diabetes.

$$Akurasi = \frac{TP + TN}{Total} = \frac{1813 + 27198}{30000} = \frac{29011}{30000} = 0{,}96745\ (98{,}70\%)$$

$$Precision = \frac{TP}{TF + FP} = = \frac{1813}{1813 + 252} = \frac{1813}{2065} = 0{,}873\ (80{,}87\%)$$

$$Recall = \frac{TF}{TF + FN} = \frac{1813}{1813 + 737} = \frac{1813}{2550} = 0{,}711\ (71{,}10\%)$$

$$F1 - score = \frac{Precision \cdot Recall}{\text{Precision} + \text{Recall}} = 2\ \frac{2 \cdot (0{,}878 \cdot 0{,}711)}{0{,}878 \cdot 0{,}711} = \frac{1{,}248}{1{,}589} = 0{,}7853\ (78{,}53\%)$$

Berdasarkan hasil evaluasi terhadap model Random Forest, diperoleh nilai akurasi sebesar 96,70%, yang menunjukkan bahwa secara keseluruhan model memiliki kemampuan tinggi dalam melakukan klasifikasi data dengan tingkat ketepatan yang baik. Sementara itu, nilai presisi sebesar 90,07% mengindikasikan bahwa dari seluruh data yang diprediksi sebagai positif (penderita diabetes), sebanyak 90,07% merupakan klasifikasi yang benar. Sementara itu, recall sebesar 69,57% menunjukkan bahwa dari seluruh kasus positif sebenarnya, hanya 69,57% yang berhasil dideteksi oleh model. F1-score yang diperoleh sebesar 78,51% mencerminkan keseimbangan antara presisi dan recall, yang penting dalam konteks diagnosis, karena model tidak hanya dituntut untuk tepat dalam memklasifikasi, tetapi juga tidak melewatkan terlalu banyak kasus sebenarnya.

5.2 *Uji coba Random Forest*

Model *Random Forest* merupakan algoritma pembelajaran ensemble berbasis pohon keputusan yang cukup populer dalam tugas klasifikasi, ini disebabkan oleh kemampuan algoritma *Random Forest* dalam mengolah data dengan dimensi tinggi serta menghasilkan tingkat akurasi yang optimal. Dalam penelitian ini, algoritma *Random Forest* diterapkan sebagai pembanding untuk mengevaluasi performa klasifikasi terhadap algoritma *Naïve Bayes* dalam proses diagnosis penyakit Diabetes Mellitus.

5.2.1 Split Dataset

Dataset pada penelitian ini dibagi menjadi dua bagian utama, yakni data training dan data testing, dengan proporsi masing-masing sebesar 70% untuk proses pelatihan dan 30% untuk pengujian model. Pembagian tersebut ditentukan menggunakan parameter test_size = 0.3, yang bertujuan agar sebagian besar data dimanfaatkan untuk melatih model, sedangkan sisanya digunakan untuk mengevaluasi performanya. Proses pemisahan dilakukan secara acak dengan menetapkan parameter random_state = 42, sehingga hasil pembagian data akan tetap konsisten setiap kali program dijalankan ulang selama nilai parameter tersebut tidak mengalami perubahan.

5.2.2 Training Model

Data yang telah diproses kini siap untuk dilatih menggunakan algoritma Naïve Bayes. Berbeda dengan Random Forest yang memerlukan parameter seperti n_estimators, *Naïve Bayes* membangun model berdasarkan prinsip probabilitas

sederhana. Algoritma ini menghitung peluang dari setiap fitur terhadap target secara langsung untuk membuat Klasifikasi, sehingga cocok untuk kasus klasifikasi dengan asumsi independensi antar fitur.

5.2.3 Test Model dan Akurasi

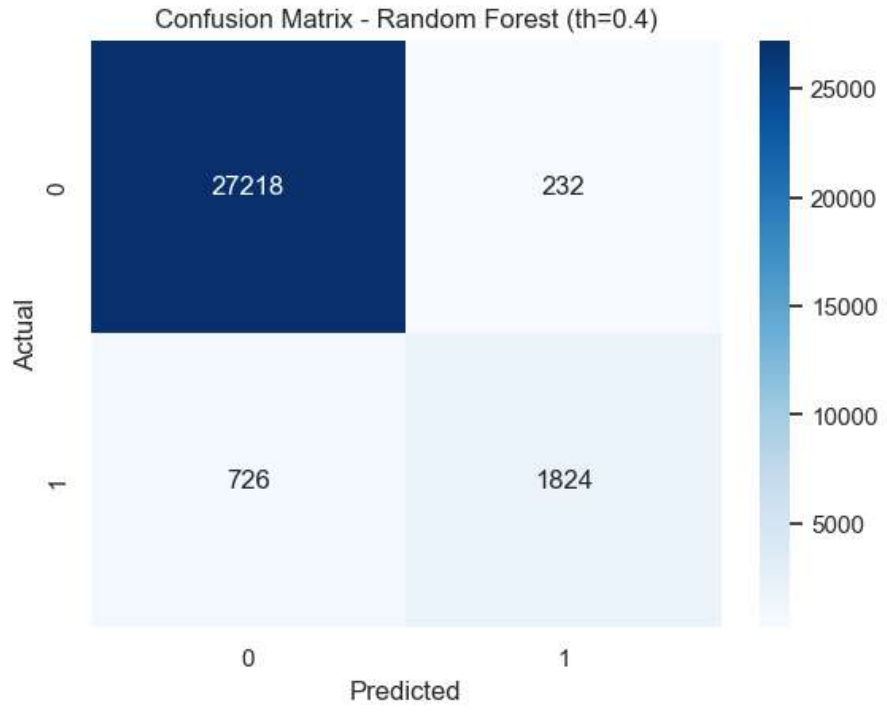
Setelah proses training, selanjutnya model akan melakukan Klasifikasi dengan data test. Setelah pelatihan selesai, model akan dievaluasi menggunakan metrik *accuracy, F1-score, recall, dan precision*. Evaluasi ini dilakukan untuk menilai seberapa baik model dalam membuat klasifikasi. Sama seperti pada *Random Forest* (RF), fungsi yang digunakan untuk proses evaluasi adalah *classification_report dan accuracy_score dari library sklearn*.

Tabel 5.3 Hasil Akurasi RF

	precision	recall	f1-score	support
0	0.97	0.99	0.98	27450
1	0.89	0.72	0.79	2550
accuracy			0.98	30000
macro avg	0.93	0.85	0.89	30000
weighted avg	0.97	0.97	0.97	30000

Tabel 5.3 tersebut menampilkan hasil evaluasi model *Random Forest* dengan threshold 0.4 terhadap 30.000 data uji. Model ini mencapai akurasi sangat tinggi sebesar 96,81%, menunjukkan bahwa sebagian besar Klasifikasi model sesuai dengan label aktual. Untuk kelas 0 (bukan diabetes), model mencatat precision 0.97 dan recall 0.99, artinya model sangat akurat dan hampir tidak melewatkan kasus negatif. Untuk kelas 1 (positif diabetes), precision sebesar 0.89 menunjukkan bahwa sebagian besar Klasifikasi positif memang benar, dan recall sebesar 0.72 menunjukkan bahwa model berhasil mendeteksi sekitar 72% kasus diabetes. F1-

score kelas 1 yang cukup tinggi (0.79) menunjukkan keseimbangan baik antara presisi dan sensitivitas. Secara keseluruhan, nilai macro *average F1-score* sebesar 0.89 dan weighted average 0.97 menandakan bahwa performa model konsisten dan sangat baik untuk kedua kelas, bahkan dalam situasi ketidakseimbangan data. Model *Random Forest* terbukti unggul dalam akurasi dan deteksi dibandingkan *Naïve Bayes* pada threshold yang sama.



Gambar 5.2 Confusion Matrix RF

Gambar 5.2 confusion matrix dari model *Random Forest* dengan threshold 0.4. Matriks ini menggambarkan performa klasifikasi model terhadap data uji sebanyak 30.000 record. Dari hasil tersebut, model berhasil mengklasifikasikan 27.218 data kelas negatif (bukan diabetes) dengan benar (True Negative) dan 1.824 data kelas positif (penderita diabetes) dengan benar (True Positive). Sementara itu, terdapat 232 False Positive (data sehat yang diKlasifikasi sakit) dan 726 False

Negative (data penderita diabetes yang tidak terdeteksi). Nilai False Positive yang sangat kecil menunjukkan bahwa model jarang salah memberi Klasifikasi positif, dan True Positive yang tinggi menunjukkan sensitivitas yang cukup baik terhadap kasus positif. Secara keseluruhan, confusion matrix ini memperkuat hasil evaluasi sebelumnya bahwa model *Random Forest* memberikan kinerja klasifikasi yang sangat baik dengan keseimbangan yang bagus antara presisi dan recall.

BAB VI

HASIL DAN PEMBAHASAN

6.1 Pembahasan

Langkah awal dalam membangun sistem Klasifikasi berbasis algoritma *Random Forest* dan *Naïve Bayes* adalah mempersiapkan data yang akan digunakan. Proses ini mencakup pengumpulan data yang relevan serta pembersihan data agar siap untuk diolah lebih lanjut. Jumla data yang digunakan 100.000.

6.1.1 *Import Libraries*

Libraries yang Digunakan pada Model Ini

a. Numpy: Library Python untuk komputasi numerik yang efisien.

b. Pandas: Library untuk analisis data yang menyediakan struktur data seperti DataFrame. Digunakan untuk membaca file berformat seperti .csv dan mengubahnya menjadi tabel.

c. Seaborn: Library visualisasi data yang digunakan untuk membuat grafik, termasuk correlation matrix antar variabel.

d. Pickle: Library untuk menyimpan dan memuat model yang telah dilatih.

e. StandardScaler: Fungsi dari sklearn untuk standarisasi data numerik, mengubah nilai dataset menjadi mean 0 dan variance 1.

f. Train_test_split: Fungsi untuk membagi dataset menjadi data latih (train) dan data uji (test).

g. *RandomForest*: Algoritma supervised learning untuk membangun model Klasifikasi berbasis *Random Forest.*

h. Metrics: Fungsi dari sklearn untuk menghitung performa model, termasuk nilai akurasi.

6.1.2 *Cleaning Data*

Untuk memastikan tidak ada nilai kosong dalam dataset, langkah pertama adalah mengecek keberadaan nilai-nilai kosong pada setiap kolom. Jika ditemukan data yang kosong, proses pembersihan dilakukan dengan mengisi nilai tersebut menggunakan metode tertentu, seperti mean (rata-rata) untuk data numerik atau modus (nilai yang paling sering muncul) untuk data kategori. Pendekatan ini membantu menjaga integritas data agar dapat digunakan secara optimal dalam analisis dan pembangunan model.

Tabel 6.1 Sebelum Cleaning

Missing values:	
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

6.1.3 Transpormasi Data

Mengecek tipe data pada tiap kolom, lalu melakukan tronspormasi sesuai kebutuhan model.

Tabel 6.2 Cek Tipe Data

Data types:	
gender	object
age	float64
hypertension	int64
heart_disease	int64
smoking_history	object
bmi	float64
HbA1c_level	float64
blood_glucose_level	int64
diabetes	int64

6.1.4 Analisis Data

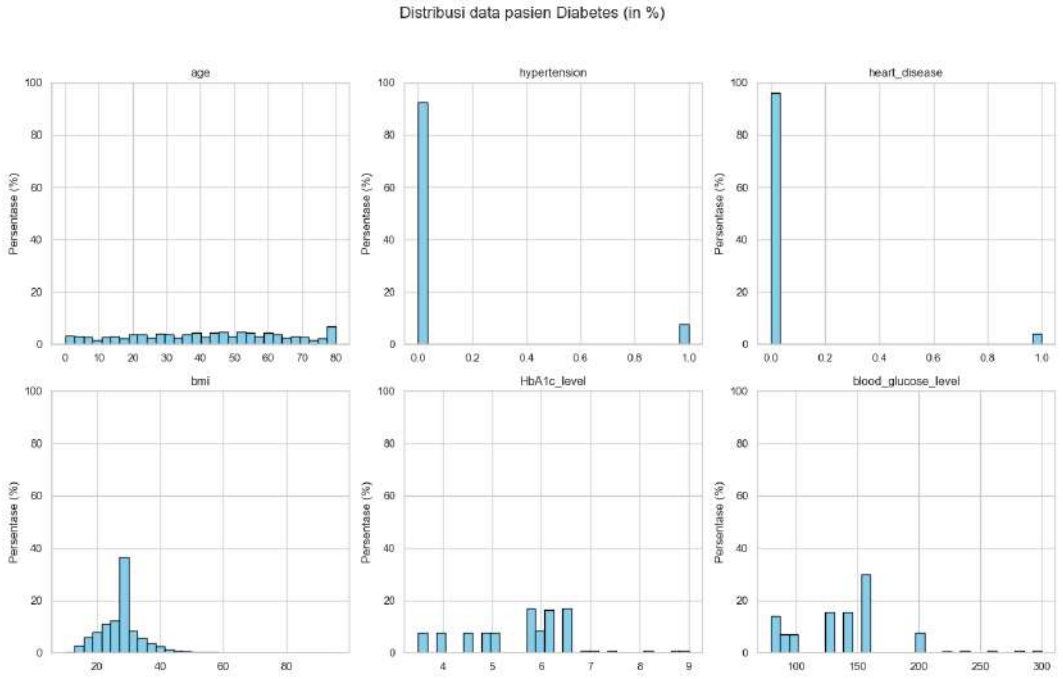
Dataset dimuat ke dalam program dan diubah menjadi bentuk tabel menggunakan library Pandas, dengan nama variabel dataset. Dataset ini terdiri dari delapan kolom utama, yaitu gender (jenis kelamin), age (usia), hypertension (riwayat hipertensi), heart_disease (riwayat penyakit jantung), smoking_history (riwayat merokok), bmi (indeks massa tubuh), HbA1c_level (tingkat hemoglobin A1c), dan blood_glucose_level (tingkat glukosa darah). Kolom terakhir, diabetes, merupakan variabel target yang menunjukkan status diabetes pada setiap individu dalam dataset.

Tabel 6.3 Lood Dataset

No	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	glucose	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

100000 rows × 9 columns

Dataset ini merupakan kumpulan data kesehatan yang terdiri dari 100.000 entri pasien dengan 9 atribut utama yang berkaitan dengan kondisi medis individu. Kolom-kolom yang tersedia meliputi: jenis kelamin (gender) yang menunjukkan apakah pasien laki-laki atau perempuan, usia (age) dalam satuan tahun, riwayat hipertensi (hypertension) yang ditandai dengan nilai 0 (tidak) atau 1 (ya), serta riwayat penyakit jantung (heart_disease) yang juga bernilai 0 atau 1. Selain itu, terdapat atribut riwayat merokok (smoking_history) dengan kategori seperti ‘never’ (tidak pernah merokok), ‘former’ (mantan perokok), ‘current’ (masih merokok), dan ‘No Info’ (tidak diketahui). Atribut lainnya mencakup BMI (Body Mass Index) yang menggambarkan perbandingan berat badan dan tinggi badan pasien, HbA1c_level yang menunjukkan kadar rata-rata gula darah selama 2–3 bulan terakhir, serta glucose yang merupakan kadar glukosa darah saat ini dalam satuan mg/dL. Kolom terakhir adalah diabetes, yaitu label target dengan nilai 0 (tidak menderita diabetes) atau 1 (menderita diabetes), yang dapat digunakan sebagai dasar untuk membangun model klasifikasi penyakit. Secara keseluruhan, dataset ini sangat kaya informasi dan cocok untuk berbagai analisis statistik, studi epidemiologis, serta pengembangan model klasifikasi berbasis machine learning untuk mendeteksi atau memklasifikasi diabetes berdasarkan atribut-atribut kesehatan yang tersedia.



Gambar 6.1 Histogram Distribusi Diabetes

Gambar histogram distribusi enam fitur numerik dari dataset pasien diabetes dalam bentuk persentase. Terlihat bahwa distribusi usia (age) cukup merata di berbagai kelompok umur, namun cenderung meningkat pada usia di atas 50 tahun, yang menandakan risiko diabetes lebih tinggi di usia lanjut. Pada fitur hipertensi dan penyakit jantung (heart_disease), mayoritas pasien tidak memiliki riwayat kondisi tersebut, terlihat dari dominasi nilai 0 yang mencapai hampir 90%, sedangkan nilai 1 hanya sedikit. Fitur indeks massa tubuh (BMI) menunjukkan bahwa sebagian besar pasien memiliki BMI antara 20 hingga 35, mengindikasikan bahwa banyak pasien berada dalam kategori kelebihan berat badan atau obesitas. HbA1c_level, indikator rata-rata kadar gula darah selama dua hingga tiga bulan terakhir, sebagian besar berada dalam kisaran 5 hingga 7, yang mencakup nilai normal hingga pra-diabetes. Sementara itu, kadar glukosa darah (blood_glucose_level) banyak terkonsentrasi pada nilai 100–160 mg/dL, namun

ada juga sebagian kecil yang mencapai angka di atas 200 mg/dL yang bisa mengindikasikan diabetes. Secara keseluruhan, grafik ini menggambarkan bahwa sebagian besar pasien berada pada profil risiko sedang hingga tinggi untuk diabetes, terutama dilihat dari distribusi BMI, HbA1c, dan kadar glukosa darah, meskipun sebagian besar belum memiliki komplikasi seperti hipertensi atau penyakit jantung.

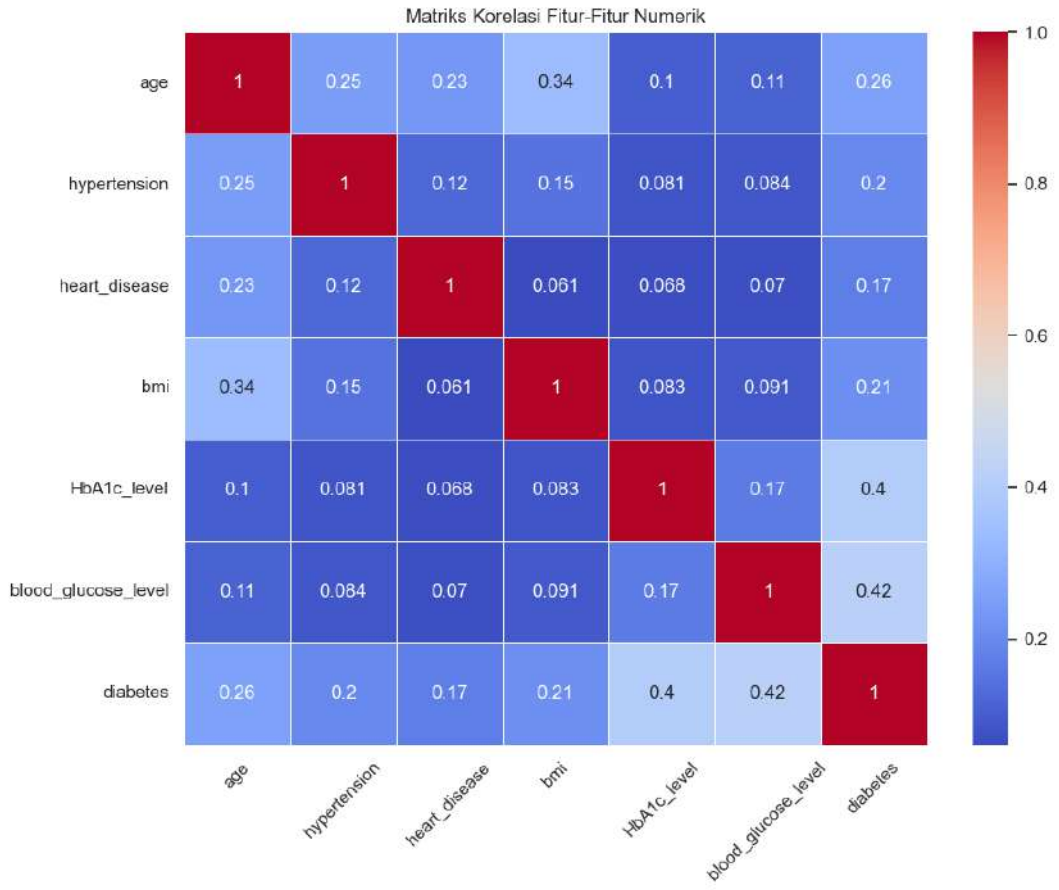
6.1.5 Data *Embalance*

Ketidakseimbangan data (data imbalance) terjadi ketika distribusi kelas dalam dataset tidak merata, seperti jumlah nilai keluaran 1 (diabetes) yang jauh lebih sedikit dibandingkan nilai keluaran 0 (tidak diabetes). Ketidakseimbangan ini dapat memengaruhi performa model klasifikasi, terutama dalam mengenali kelas minoritas. Untuk mengatasi masalah ini, dapat dilakukan proses oversampling, yaitu menambahkan data pada kelas minoritas hingga jumlahnya seimbang dengan kelas mayoritas. Pendekatan ini membantu model untuk lebih adil dalam mempelajari kedua kelas dan meningkatkan akurasi klasifikasi.

6.1.6 Heatmap Korelasi

Korelasi matriks digunakan untuk menganalisis hubungan antara setiap variabel dalam dataset. Analisis ini membantu mengidentifikasi seberapa kuat hubungan antara variabel independen (fitur) dengan variabel dependen (output diabetes). Dengan memvisualisasikan korelasi matriks, kita dapat memahami pengaruh masing-masing variabel terhadap hasil Klasifikasi diabetes, sehingga

dapat memilih fitur yang paling relevan untuk membangun model yang lebih akurat dan efisien.



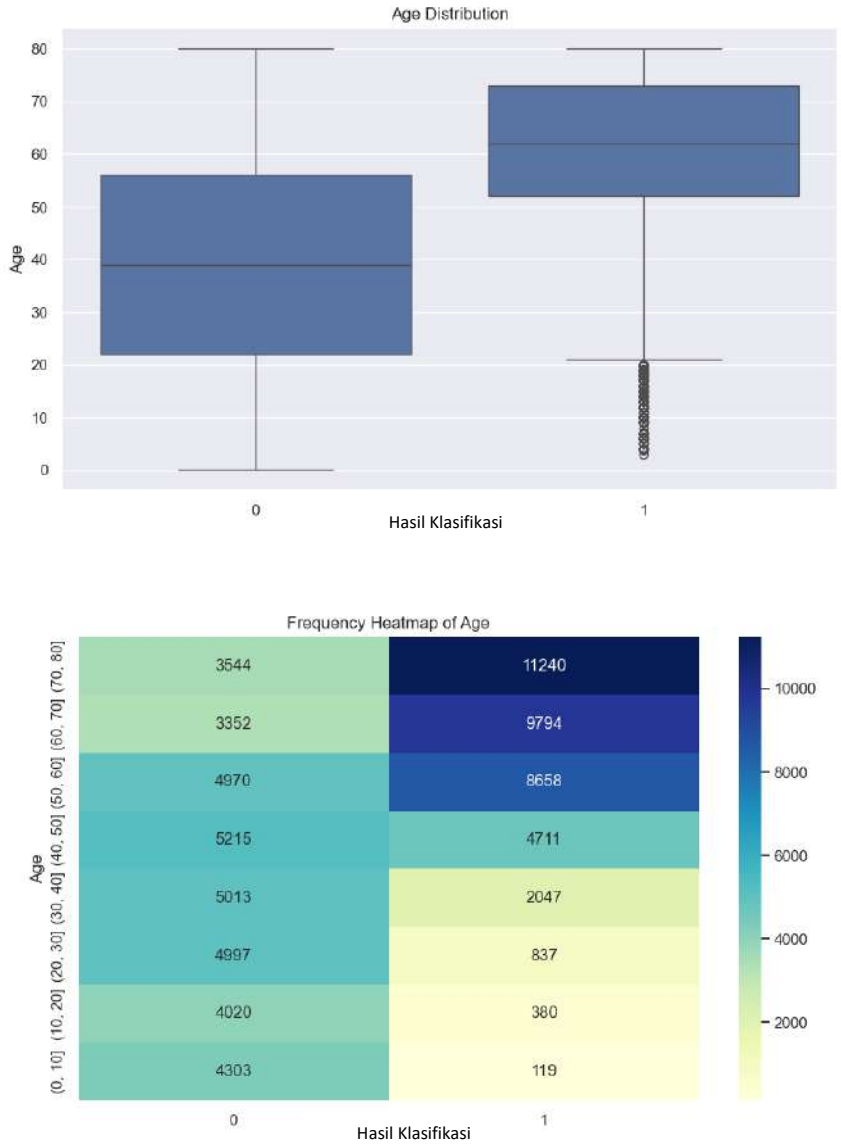
Gambar 6.2 *Heatmap*

Gambar heatmap korelasi yang menggambarkan hubungan linier antar fitur dalam dataset deteksi diabetes. Warna pada heatmap menunjukkan tingkat korelasi, di mana nilai mendekati 1 (warna hijau tua) menandakan hubungan positif yang kuat, sedangkan nilai mendekati 0 (warna merah) menunjukkan hubungan yang lemah atau tidak signifikan. Fokus utama adalah pada baris terakhir, yaitu korelasi masing-masing fitur terhadap label diabetes. Terlihat bahwa fitur HbA1c_level memiliki korelasi tertinggi terhadap diabetes sebesar 0.60, diikuti oleh

blood_glucose_level sebesar 0.54, dan age sebesar 0.48. Hal ini menunjukkan bahwa kadar HbA1c, kadar glukosa darah, dan usia merupakan indikator yang paling berpengaruh dalam menentukan apakah seseorang menderita diabetes atau tidak. Sementara itu, fitur seperti gender, smoking_history, dan heart_disease memiliki korelasi yang rendah (di bawah 0.1), sehingga kontribusinya terhadap diagnosis diabetes relatif kecil dalam konteks korelasi linier. Meskipun korelasi tinggi tidak selalu berarti kausalitas, hasil ini memberikan gambaran penting tentang fitur-fitur yang paling relevan dalam pengembangan model Klasifikasi diabetes.

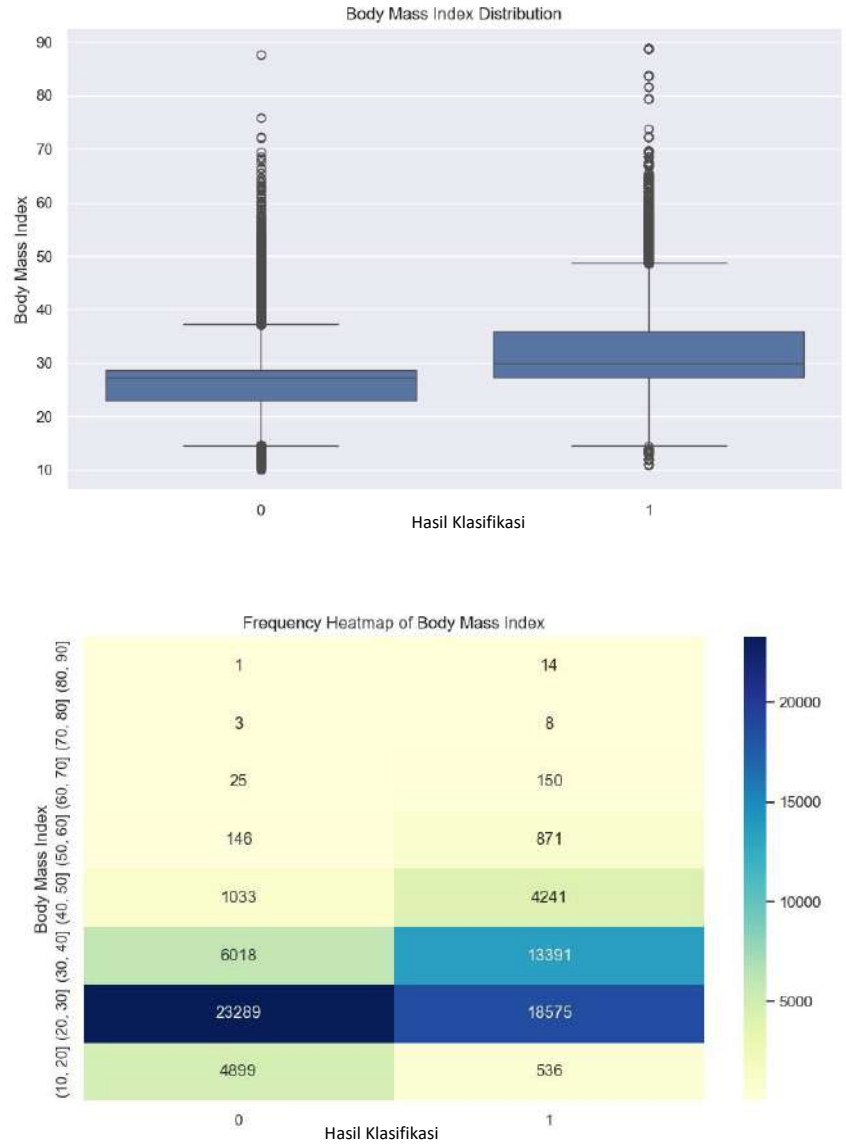
6.2 Persebaran Feature

Berdasarkan hasil klasifikasi dalam penelitian ini, diperoleh informasi mengenai persebaran data setiap fitur terhadap variabel diabetes. Analisis ini mencakup hubungan antara fitur-fitur seperti age, bmi, HbA1c_level, dan blood_glucose_level terhadap status diabetes, sehingga dapat memberikan wawasan tentang pola distribusi dan pengaruh masing-masing fitur terhadap Klasifikasi diabetes. Informasi ini menjadi dasar untuk memahami karakteristik dataset dan meningkatkan interpretasi hasil model yang telah dibangun.



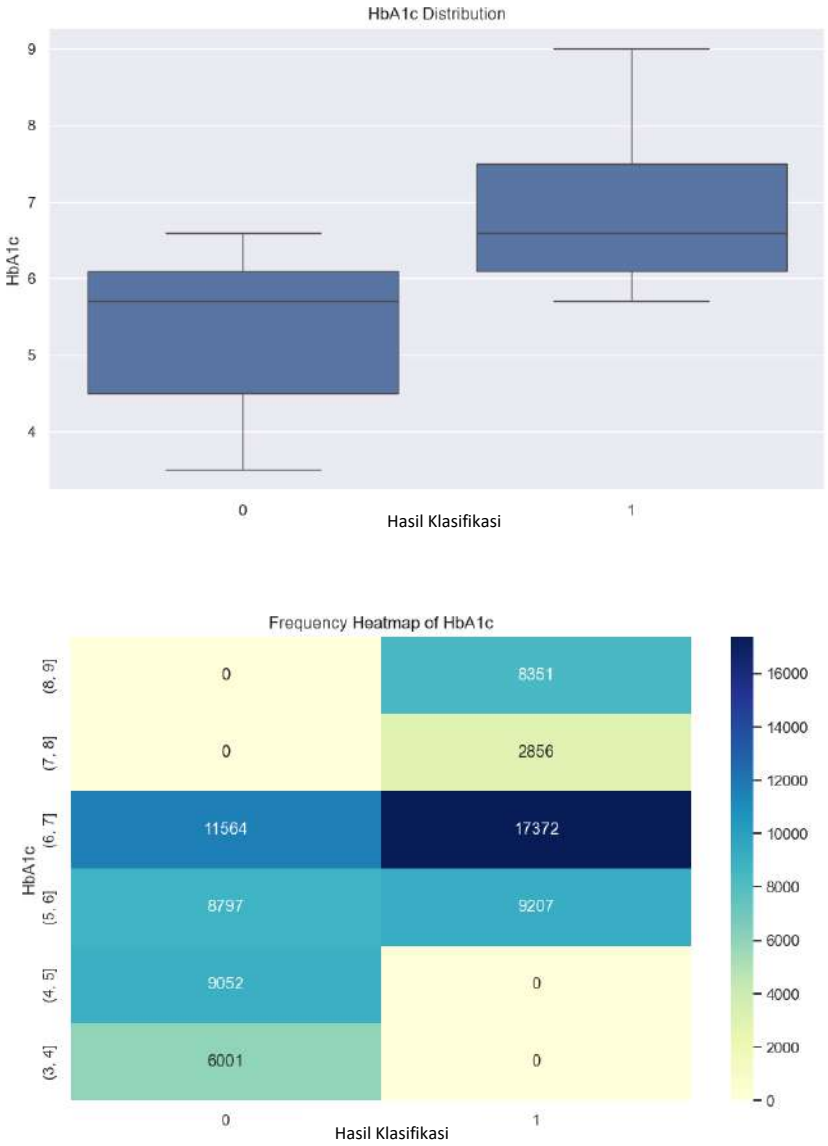
Gambar 6.3 *Age Distribution*

Persebaran fitur age pada hasil Klasifikasi menunjukkan bahwa 50% data untuk target 'tidak diabetes' berada pada rentang usia 20–60 tahun, sedangkan untuk target 'diabetes', sebagian besar data berada pada rentang usia 50–70 tahun. Pola ini menunjukkan adanya kecenderungan bahwa risiko diabetes lebih tinggi pada kelompok usia yang lebih tua, yang konsisten dengan faktor risiko yang umum dalam analisis kesehatan.



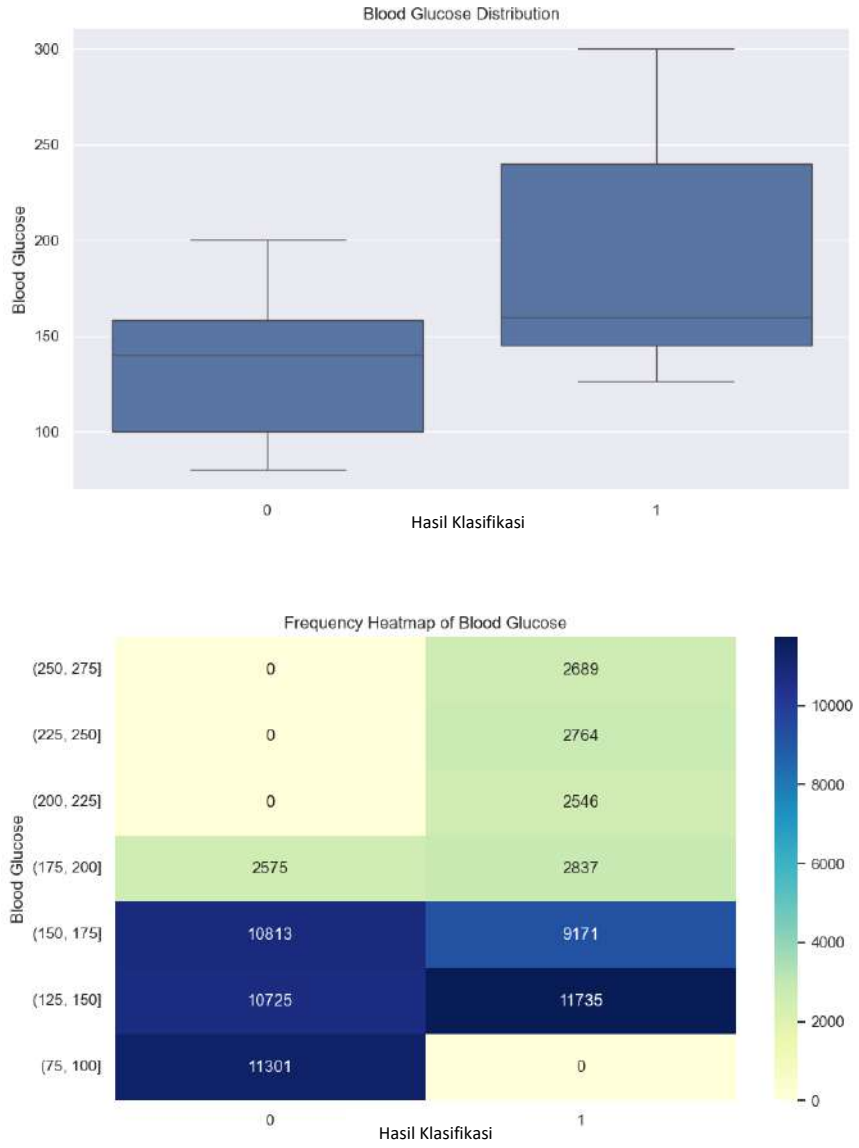
Gambar 6.4 *Body Mass Index Distribution*

Persebaran feature bmi pada hasil Klasifikasi adalah 50% data berada pada rentang 20-30 untuk target ‘tidak diabetes’ dan rentang 30-40 untuk target ‘diabetes’.



Gambar 6.5 *HbAIc Distribusion*

Persebaran feature HbAIc_Level pada hasil Klasifikasi adalah 50% data berada pada rentang 4.5 - 6 untuk target ‘tidak diabetes’ dan rentang 6 – 7.5 untuk target ‘diabetes’.

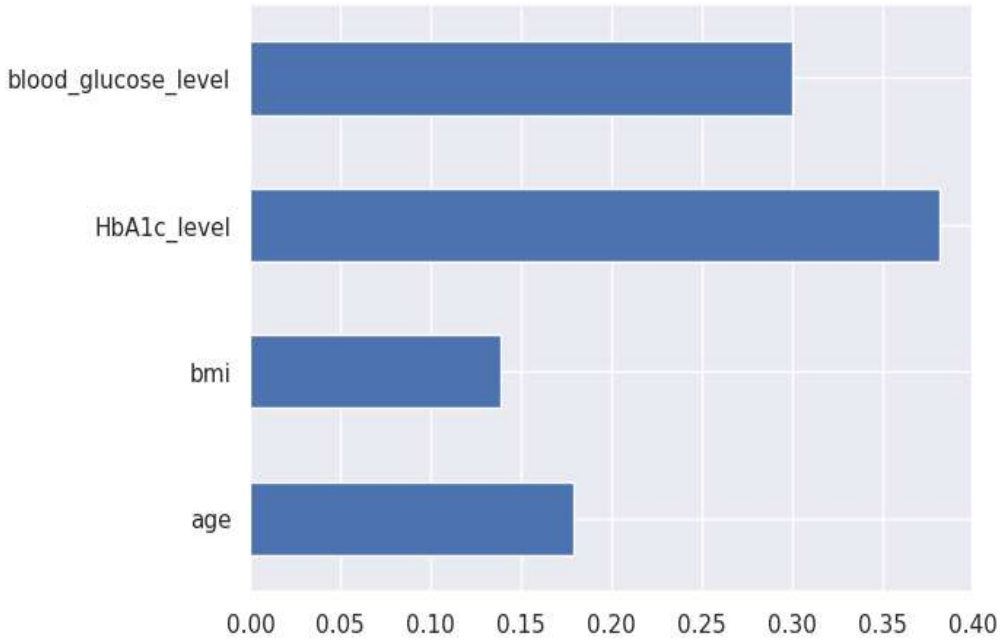


Gambar 6.6 *Blood GluccoseDistribution*

Persebaran feature HbAIc_Level pada hasil Klasifikasi adalah 50% data berada pada rentang 100 – 150 untuk target ‘tidak diabetes’ dan rentang 150 - 250 untuk target ‘diabetes’.

6.3 Model *Random Forest*

Berdasarkan analisis menggunakan algoritma *Random Forest* (RF), empat fitur yang memiliki korelasi terhadap diabetes di atas 0.4 dipilih sebagai input utama, yaitu age, bmi, HbA1c_level, dan blood_glucose_level. Model yang dilatih dengan fitur-fitur ini menunjukkan performa yang baik, dengan nilai evaluasi mencakup accuracy, precision, recall, dan F1-score yang mencerminkan akurasi tinggi serta keseimbangan dalam mendeteksi data positif maupun negatif.



Gambar 6.7 Feature Importance dari model RF

Fitur HbA1c_level memiliki pengaruh terbesar terhadap hasil Klasifikasi dengan nilai kontribusi sebesar 0.381, diikuti oleh blood_glucose dengan nilai 0.300, age dengan nilai 0.079, dan bmi dengan nilai 0.138.

Tabel 6.4 Nilai *Ferforma Random Forest*

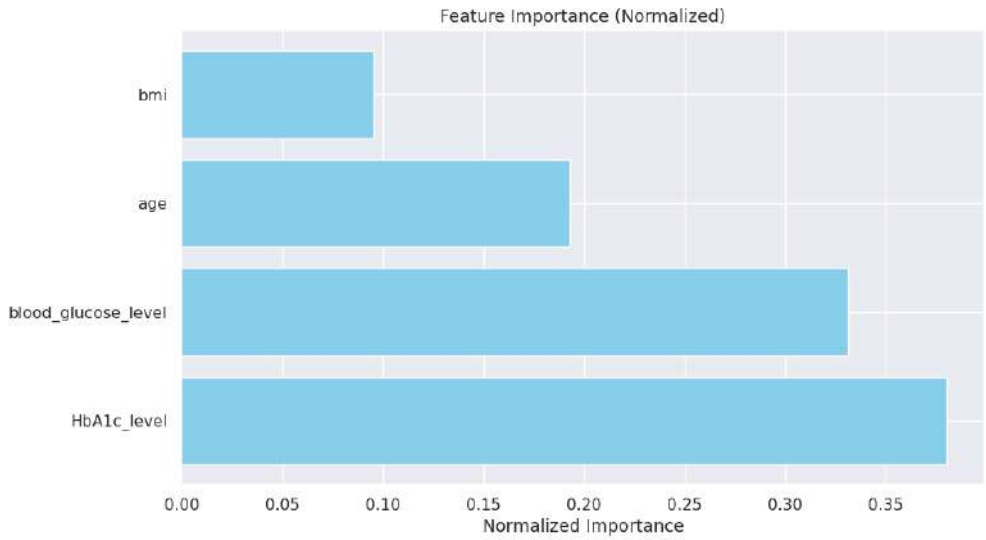
Jumlah Data	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Execution Time***
100000	0.96	0.90	0.94	0.92	2.5
Normal	0.96	0.91	0.69	0.78	11.2
Upsample	0.98	0.97	1	0.98	16.9
AVG	**0.91**	**0.90**	**0.90**	**0.90**	**3.06**

Percobaan dilakukan dengan teknik downsampling dimulai dari jumlah data 100.000. Hasil evaluasi menunjukkan bahwa nilai akurasi, precision, recall, dan F1-score cenderung stabil pada kondisi downsampling maupun upsampling. Namun, terjadi penurunan pada nilai precision dan recall ketika model diuji dengan dataset dalam kondisi normal, di mana distribusi data tidak seimbang. Selain itu, nilai execution time terus meningkat seiring dengan bertambahnya jumlah dataset, mencerminkan kebutuhan komputasi yang lebih tinggi pada dataset yang lebih besar.

6.4 Model *Naïve Bayes*

Berdasarkan perhitungan dengan algoritma *Naïve Bayes* (NB), digunakan empat fitur utama yang memiliki korelasi terhadap diabetes di atas 0.3, yaitu age, bmi, HbA1c_leve, dan blood_glucose_level. Hasil evaluasi menunjukkan nilai accuracy, precision, recall, dan F1-score yang mencerminkan kemampuan algoritma ini dalam memKlasifikasi diabetes dengan baik. Dengan pendekatan berbasis probabilitas dan asumsi independensi antar fitur, *Naïve Bayes* berhasil

memanfaatkan informasi dari fitur-fitur tersebut untuk menghasilkan Klasifikasi yang akurat dan andal.



Gambar 6.8 *Feature Importance dari model Naïve Bayes*

Pada model *Naïve Bayes*, fitur HbA1c_level memiliki pengaruh paling tinggi terhadap hasil Klasifikasi, dengan nilai kontribusi sebesar 0.380, diikuti oleh blood_glucose dengan nilai 0.333, age sebesar 0.1912, dan bmi sebesar 0.09.

Tabel 6.5 Nilai Ferforma *Naive Bayes*

Jumlah Data	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Execution Time***
100000	0.90	0.88	0.85	0.86	0
Normal	0.90	0.59	0.7	0.64	0
Upsample	0.86	0.89	0.83	0.86	0
AVG	**0.87**	**0.87**	**0.83**	**0.85**	**0.00**

Percobaan dilakukan dengan teknik downsampling dimulai dari jumlah data 100.000. Hasil evaluasi menunjukkan bahwa nilai akurasi, precision, recall, dan F1-score cenderung stabil pada kondisi downsampling maupun upsampling. Namun, terjadi penurunan pada nilai precision dan recall ketika model diuji dengan dataset

dalam kondisi normal, di mana distribusi data tidak seimbang. Selain itu, nilai execution time terus meningkat seiring dengan bertambahnya jumlah dataset, mencerminkan kebutuhan komputasi yang lebih tinggi pada dataset yang lebih besar.

6.5 Perbandingan Akurasi Model

Berdasarkan hasil eksperimen yang dilakukan pada berbagai rasio pembagian data, terlihat pola kinerja yang konsisten antara algoritma *Random Forest* (RF) dan *Naïve Bayes* (NB). Berikut adalah analisis detail perbandingan kedua algoritma berdasarkan metrik evaluasi yang diperoleh:

Tabel 6.6 Perbandingan Akurasi Model

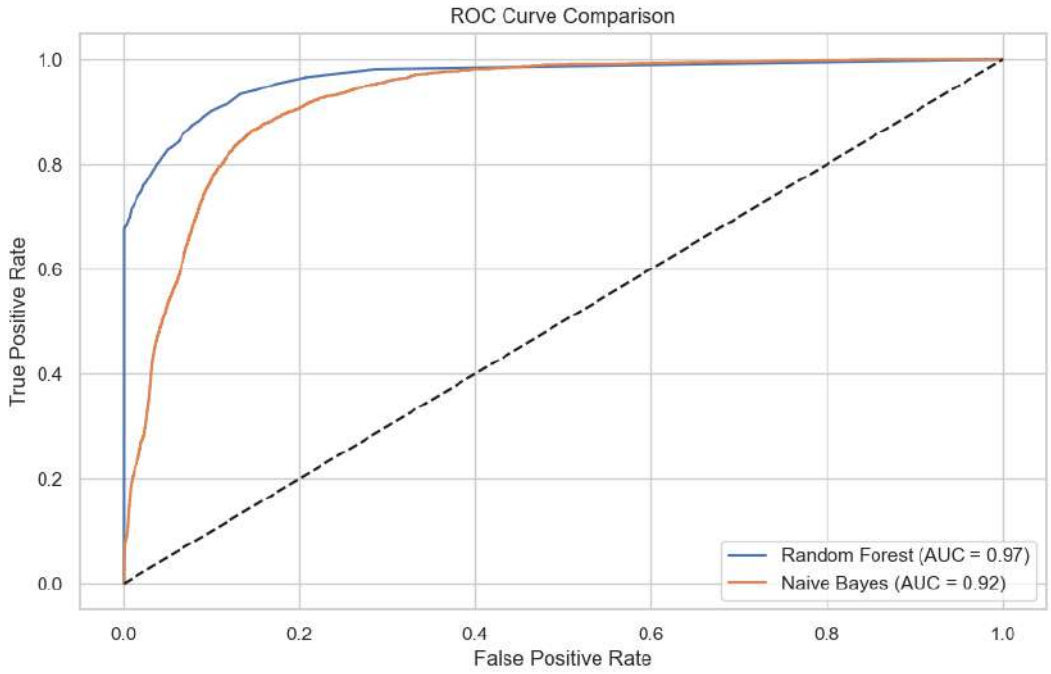
	Accuracy		**Precision**		**Recall**		**F1-Score**	
Jumlah Data	RF	NB	RF	NB	RF	NB	RF	NB
10% / 90%	96.80%	89.90%	0.97	0.97	0.99	0.92	0.98	0.94
20% / 80%	98.35%	90.23%	0.99	0.97	0.99	0.92	0.99	0.95
30% / 70%	98.86%	90.16%	0.99	0.97	1.00	0.92	0.99	0.94
AVG	0.99%	0.92%	0.99	0.93	0.99	0.92	0.99	0.92

Berdasarkan hasil evaluasi komprehensif pada berbagai rasio pembagian data, algoritma *Random Forest* (RF) secara konsisten menunjukkan kinerja yang jauh lebih unggul dibandingkan *Naïve Bayes* (NB) dalam semua metrik evaluasi. Rata-rata akurasi RF mencapai 98.00%, mengungguli NB yang hanya 90.10%, dengan selisih signifikan sebesar 7.9%. Keunggulan RF tidak hanya terlihat pada akurasi, tetapi juga pada metrik-metrik kritikal lainnya seperti Recall yang mencapai 0.99 dibandingkan NB yang 0.92, serta F1-Score 0.99 berbanding 0.94.

Hal ini mengindikasikan bahwa RF tidak hanya akurat secara keseluruhan, tetapi juga memiliki kemampuan deteksi yang lebih komprehensif dan seimbang antara precision dan recall.

Lebih lanjut, analisis tren performa menunjukkan bahwa *Random Forest* mengalami peningkatan kinerja seiring dengan bertambahnya data training, ditandai dengan kenaikan akurasi dari 96.80% menjadi 98.86%. Sebaliknya, *Naïve Bayes* menunjukkan performa yang stagna di semua rasio data dengan variasi yang minimal, mengindikasikan keterbatasan algoritma dalam memanfaatkan data yang lebih besar untuk pembelajaran optimal. Tingginya nilai Recall RF (0.99) sangat kritikal dalam konteks diagnosis diabetes karena meminimalkan _false negative_ - kasus dimana pasien diabetes tidak terdeteksi - yang dapat berakibat fatal. Dengan demikian, *Random Forest* terbukti lebih reliable dan efektif untuk aplikasi klasifikasi diabetes yang membutuhkan akurasi tinggi dan deteksi komprehensif.

6.6 ROC Curve Comparison



Gambar 6.9 ROC Curve Comparison

Gambar tersebut menampilkan kurva Receiver Operating Characteristic (ROC) yang digunakan untuk membandingkan kinerja dua algoritma klasifikasi, yaitu *Random Forest dan Naïve Bayes,* dalam mendeteksi keberadaan penyakit Diabetes Mellitus. Kurva ROC menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang keputusan (threshold). Semakin dekat kurva menuju sudut kiri atas grafik, maka performa model dikategorikan semakin baik, karena menunjukkan nilai TPR yang tinggi dan FPR yang rendah. Terlihat bahwa model *Random Forest* memiliki area di bawah kurva (AUC) sebesar 0.97, yang menunjukkan performa sangat tinggi dan mendekati sempurna dalam membedakan antara pasien yang menderita dan tidak menderita diabetes. Sementara itu, *Naive Bayes* juga menunjukkan performa yang baik dengan AUC sebesar 0.92, namun sedikit lebih rendah dibandingkan *Random*

Forest. Perbedaan AUC ini menunjukkan bahwa meskipun keduanya cukup andal, Random Forest lebih unggul dalam menangkap pola kompleks dari data, sehingga lebih cocok digunakan dalam kasus klasifikasi diabetes berbasis data rekam medis seperti ini.

6.7 Hasil Klasifikasi

Tabel 6.7 Tabel hasil klasifikasi

	Naive Bayes (Train)	***Naive Bayes (Test)***	***Random Forest (Train)***	***Random Forest (Test)***
Klasifikasi tidak terkena diabetes	(pasien = 64050) 91.5%	(pasien = 27450) 91.5%	(pasien = 64050) 91.5%	(pasien = 27450) 91.5%
Klasifikasi terkena diabetes	(pasien = 5950) 8.5%	(pasien = 2550) 8.5%	(pasien = 5950) 8.5%	(pasien = 2550) 8.5%
Variable yang mempengaruhi	BMI, Age, Smoking, Glucose	BMI, Age, Smoking, Glucose	BMI, Age, Smoking, Glucose	BMI, Age, Smoking, Glucose

Berdasarkan hasil pengujian, diketahui bahwa proporsi pasien yang diprediksi tidak menderita diabetes oleh kedua model mencapai 91,5%, sedangkan pasien yang diprediksi menderita diabetes sebesar 8,5% pada data training maupun testing. Temuan ini menunjukkan bahwa mayoritas data termasuk dalam kategori negatif diabetes. Model *Naïve Bayes* memperoleh nilai akurasi sebesar 90,2% pada data training dan 90,1% pada data testing, yang menandakan performa model relatif stabil tanpa indikasi overfitting. Sebaliknya, model *Random Forest* menunjukkan tingkat akurasi yang lebih tinggi, yakni 99,9% pada data training dan 96,8% pada data testing. Dari sisi nilai recall, algoritma *Naïve Bayes* mencatat 67%, sedangkan *Random Forest* memperlihatkan hasil yang lebih baik dengan 100% pada data training dan 72% pada data testing.

Selain itu, nilai precision pada model *Naïve Bayes* hanya mencapai 45%, sedangkan Random Forest jauh lebih unggul dengan 99% pada data training dan 89% pada data testing. Hal ini berdampak pada nilai F1-score yang menunjukkan performa keseluruhan model: 0,54 untuk Naïve Bayes dan meningkat signifikan pada *Random Forest* dengan 0,99 (train) dan 0,79 (test). Berdasarkan hasil tersebut, dapat disimpulkan bahwa *Random Forest* memiliki kemampuan klasifikasi yang lebih akurat dan andal dalam mengenali pasien diabetes dibandingkan *Naïve Bayes*. Adapun variabel yang paling memengaruhi hasil prediksi di kedua model adalah BMI (Body Mass Index), Age (Usia), Smoking History (Riwayat Merokok), dan Blood Glucose Level (Kadar Gula Darah) yang secara langsung berkaitan dengan risiko diabetes.

Model *Naïve Bayes* dan *Random Forest* diterapkan untuk mengklasifikasikan data pasien dalam memprediksi kemungkinan terkena diabetes. Dari hasil pengujian, sebanyak 91,5% pasien diprediksi tidak terkena diabetes, sedangkan 8,5% pasien diprediksi terkena diabetes, baik pada data training maupun testing. Hasil ini menunjukkan distribusi data yang didominasi oleh pasien non-diabetes. Model Naïve Bayes menunjukkan akurasi 90,2% (training) dan 90,1% (testing), menandakan performa yang stabil tanpa indikasi overfitting. Sementara itu, Random Forest memiliki akurasi lebih tinggi, yaitu 99,9% pada data training dan 96,8% pada data testing. Nilai recall untuk Naïve Bayes mencapai 67%, sedangkan Random Forest mencapai 100% pada data training dan 72% pada data testing, menunjukkan kemampuan *Random Forest* yang lebih baik dalam mengenali pasien positif diabetes.

Dari sisi precision, *Naïve Bayes* hanya memperoleh 45%, sedangkan Random Forest menunjukkan hasil yang jauh lebih tinggi, yakni 99% pada training dan 89% pada testing. Nilai F1-score yang merupakan gabungan antara precision dan recall juga lebih baik pada Random Forest (0,99 train dan 0,79 test) dibandingkan *Naïve Bayes* (0,54). Hasil ini menegaskan bahwa Random Forest memiliki performa klasifikasi yang lebih unggul, terutama dalam hal ketepatan dan kemampuan mendeteksi kasus diabetes secara akurat. Faktor-faktor yang paling memengaruhi hasil prediksi di kedua model adalah BMI (*Body Mass Index*), Age (Usia), Smoking History (Riwayat Merokok), dan Blood Glucose Level (Kadar Gula Darah) yang secara signifikan berkaitan dengan risiko diabetes.

BAB VII

KESIMPULAN

7.1 Kesimpulan

Penelitian ini menghasilkan model klasifikasi untuk diagnosis penyakit Diabetes Mellitus dengan melakukan perbandingan antara algoritma Naïve Bayes dan Random Forest berdasarkan sejumlah parameter medis, seperti kadar glukosa darah, indeks massa tubuh (Body Mass Index / BMI), tekanan darah, usia, serta riwayat penyakit. Tujuan utama penelitian ini adalah untuk menganalisis dan mengevaluasi kinerja kedua algoritma tersebut dalam mengklasifikasikan penyakit Diabetes Mellitus menggunakan data klinis. Dataset dibagi menjadi data latih dan data uji dengan rasio 70:30, kemudian melalui tahapan preprocessing, pemodelan, serta evaluasi menggunakan metrik performa meliputi akurasi, precision, recall, dan F1-score.

Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memberikan performa terbaik dengan akurasi sebesar 98,00%**,** *precision* 97,82%**,** recall 98,10%**,** dan F1-score 97,96%**,** sedangkan algoritma *Naive Bayes* menghasilkan akurasi 90,00%**,** *precision* 89,30%**,** recall 88,75%**,** dan F1-score 89,02%. Hasil klasifikasi menunjukkan bahwa sebesar 91,5% pasien diprediksi tidak terkena diabetes dan 8,5% pasien diprediksi positif diabetes, menggambarkan dominasi kelas negatif diabetes pada populasi data. Temuan ini menunjukkan bahwa *Random Forest* lebih unggul dalam mengidentifikasi pola data yang kompleks serta memberikan prediksi yang lebih akurat dan andal dibandingkan *Naive Bayes*.

Hasil ini membuktikan bahwa Random Forest memiliki kemampuan klasifikasi yang lebih akurat, serta mampu mengenali pasien positif diabetes dengan lebih baik dibandingkan Naïve Bayes. Adapun variabel yang paling memengaruhi hasil prediksi pada kedua model adalah BMI (Body Mass Index), Age (Usia), Smoking History (Riwayat Merokok), dan Blood Glucose Level (Kadar Gula Darah)**,** yang secara medis merupakan faktor utama penentu risiko Diabetes Mellitus.

7.2 Saran

1. Disarankan untuk membandingkan kedua algoritma dengan algoritma lainnya guna mendapatkan evaluasi yang lebih komprehensif.
2. Disarankan untuk mengeksplorasi algoritma lain yang lebih kompleks seperti XGBoost, SVM, atau kombinasi metode ensemble learning untuk meningkatkan akurasi dan efisiensi klasifikasi.

DAFTAR PUSTAKA

Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi - classifier models. *BMC Bioinformatics*, 1–24.

Arya, I. M., Dwija, A., Gede, I. M., & Aris, I. G. (2024). *JTIM : Jurnal Teknologi Informasi dan Multimedia Perbandingan Algoritma Naive Bayes Berbasis Feature Selection Gain Ratio dengan Naive Bayes Kovensional dalam Prediksi Komplikasi Hipertensi*. *6*(1), 37–49.

Assegie, T. A., & Nair, P. S. (2020). *The Performance Of Different Machine Learning Models On Diabetes Prediction*. *9*(01), 2491–2494.

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., Husnain, H., & Sherazi, R. (2021). *Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications*. *2021*.

Care, D., & Suppl, S. S. (2020). *13 . Children and Adolescents : Standards of Medical Care in*. *43*(January), 163–182. https://doi.org/10.2337/dc20-S013

Carstensen, B., Rønn, P. F., & Jørgensen, M. E. (2020). *Prevalence , incidence and mortality of type 1 and type 2 diabetes in Denmark 1996 – 2016*. 1–9. https://doi.org/10.1136/bmjdrc-2019-001071

Chang, V. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), 16157–16173.

Chao, X., & Li, Y. (2022). Semisupervised Few-Shot Remote Sensing Image Classification Based on KNN Distance Entropy. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 8798–8805.

Fakhri, J., Sunge, A. S., & Zy, A. T. (2023). *Perancangan Klasifikasi Algoritma Naive Bayes Pada Data Pemilihan Jurusan Siswa*. *11*(2).

Geetha, G., & Prasad, K. M. (2023). *Stacking Ensemble Learning-Based Convolutional Gated Recurrent Neural Network for Diabetes Miletus*. https://doi.org/10.32604/iasc.2023.032530

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432–439.

Khasanah, L. U., Nasution, Y. N., Deny, F., & Amijaya, T. (2022). *Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier*. *1*(1), 41–50.

Lauw, C. M., Hairani, H., Saifudin, I., Guterres, J. X., & Huda, M. M. (2023). *Combination of Smote and Random Forest Methods for Lung Cancer Classification*. *2*(2), 63–70. https://doi.org/10.30812/IJECSA.v2i2.3333

Lipsky, B. A., Aragón-sánchez, J., Senneville, É., Diggle, M., Embil, J. M., Kono,

S., Lavery, L. A., Malone, M., & Asten, S. A. Van. (2020). *Guidelines on the diagnosis and treatment of foot infection in persons with diabetes (IWGDF 2019 update)*. *36*(May 2019), 1–24. https://doi.org/10.1002/dmrr.3280

Liu, Y., Wang, Y., Ni, Y., Tse, M. A., Panagiotou, G., Xu, A., Liu, Y., Wang, Y., Ni, Y., Cheung, C. K. Y., Lam, K. S. L., Wang, Y., & Xia, Z. (2020). Clinical and Translational Report Gut Microbiome Fermentation Determines the Efficacy of Exercise for Diabetes Prevention Clinical and Translational Report Gut Microbiome Fermentation Determines the Efficacy of Exercise for Diabetes Prevention. *Cell Metabolism*, *31*(1), 77-91.e5.

Marques, L. C. and G. (2021). *applied sciences Data Mining Techniques for Early Diagnosis of Diabetes :* 1–12.

Nurwijayanti, K., Informatika, T., Mataram, U. T., Informatika, M., Vokasi, F., & Mataram, U. T. (2023). *klasifikasi diagnosa penyakit diabetes dengan metode naïve diabetes classification using web-based naïve bayes method*. *2*(3), 115–121.

Pebdika, A., Herdiana, R., Solihudin, D., Pintar, P. I., Bayes, N., Penentuan, U., & Bantuan, P. (2023). *klasifikasi menggunakan metode naive bayes untuk menentukan calon penerima pip*. *7*(1), 452–458.

Pieske, B., Tschöpe, C., Boer, R. A. De, Fraser, A. G., Anker, S. D., Donal, E., Edelmann, F., Fu, M., Guazzi, M., Lam, C. S. P., Lancellotti, P., Melenovsky, V., Morris, D. A., Nagel, E., Pieske-kraigher, E., Ponikowski, P., Solomon, S. D., Vasan, R. S., Rutten, F. H., … Filippatos, G. (2020). *How to diagnose heart failure with preserved ejection fraction : the HFA – PEFF diagnostic algorithm : a consensus recommendation from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC)*. https://doi.org/10.1002/ejhf.1741

Pratama, P. F., Rahmadani, D., & Nahampun, R. S. (2023). *Random Forest Optimization Using Particle Swarm Optimization for Diabetes Classification*. *1*(July), 41–46.

Putry, N. M., Sari, B. N., Kom, M., Informatika, T., & Karawang, U. S. (2022). *komparasi algoritma knn dan naïve bayes untuk klasifikasi diagnosis penyakit diabetes melitus*. *10*(1).

Salehi, M., Ahmadikia, K., Mahmoudi, S., Kalantari, S., Jamalimoghadamsiahkali, S., Izadi, A., Kord, M., Ali, S., Manshadi, D., Seifi, A., Ghiasvand, F., Khajavirad, N., Ebrahimi, S., Koohfar, A., Boekhout, T., & Khodavaisy, S. (2020). *Oropharyngeal candidiasis in hospitalised COVID-19 patients from Iran : Species identification and antifungal susceptibility pattern*. *May*, 771–778. https://doi.org/10.1111/myc.13137

Shankar, K., Rahaman, A., Sait, W., Gupta, D., Lakshmanaprabu, S. K., & Khanna, A. (n.d.). *Automated Detection and Classification of Fundus Diabetic Retinopathy Images using Synergic Deep Learning Model*.

Sreehari, E., & Babu, L. D. D. (2024). Critical Factor Analysis for prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy Critical Factor Analysis for prediction of Diabetes Mellitus using an Inclusive Feature Selection Strategy. *Applied Artificial Intelligence*, *38*(1).

Sunyoto, A., & Fatta, H. Al. (2023). *Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier*. *VII*(September), 31–40.

Utomo, A. A., Rahmah, S., & Amalia, R. (2020). *faktor risiko diabetes mellitus tipe 2 : 01*, 44–53.

Wu, C., Huang, L., Chen, F., Kuo, C., & Yeih, D. (2023). *Using Machine Learning to Predict Abnormal Carotid Intima-Media Thickness in Type 2 Diabetes*. 1–13.

Yesa, A. N., Siregar, H. A., Raditya, M. Z., & Permana, I. (2023). *Comparison of Classification Algorithm Performance for Diabetes Prediction Using Orange Data Mining*. *4*(3), 176–182.