TESIS

ANALISIS PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK IDENTIFIKASI DIABETES DENGAN MENGGUNAKAN METODE *RANDOM FOREST* DAN *NAÏVE BAYES*

Disusun oleh:

Nama : **Muhammad Rafli Zuhri**
NIM : **22.55.1241**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

TESIS

ANALISIS PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK IDENTIFIKASI DIABETES DENGAN MENGGUNAKAN METODE *RANDOM FOREST* DAN *NAÏVE BAYES*

COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHM FOR IDENTIFICATION OF DIABETES USING RANDOM FOREST AND NAÏVE BAYES METHODS

Diajukan untuk memenuhi salah satu syarat mencapai derajat Pascasarjana
Program Studi S2 Teknik Informatika

Disusun oleh:

MUHAMMAD RAFLI ZUHRI
22.55.1241
Konsentrasi : Business Intelligence

FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

HALAMAN PERSETUJUAN

ANALISIS PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK IDENTIFIKASI DIABETES DENGAN MENGGUNAKAN METODE RANDOM FOREST DAN NAÏVE BAYES

COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHM FOR IDENTIFICATION OF DIABETES USING RANDOM FOREST AND NAÏVE BAYES METHODS

yang disusun dan diajukan oleh

Muhammad Rafli Zuhri

22.55.1241

telah disetujui oleh Dosen Pembimbing Tesis
pada tanggal 16 Desember 2025

Dosen Pembimbing,

Prof. Dr. Kusrini, M.Kom
NIK. 190302106

HALAMAN PENGESAHAN

ANALISIS PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK IDENTIFIKASI DIABETES DENGAN MENGGUNAKAN METODE RANDOM FOREST DAN NAÏVE BAYES

COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHM FOR IDENTIFICATION OF DIABETES USING RANDOM FOREST AND NAÏVE BAYES METHODS

yang disusun dan diajukan oleh

Muhamamd Rafli Zuhri

22.55.1241

Telah dipertahankan di depan Dewan Penguji
pada tanggal 16 Desember 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Hanafi, S.Kom., M.Eng., P.hD **NIK. 190302024**	
Emha Taufiq Luthfi, S.T., S.Kom., P.hD **NIK. 190302125**	
Prof. Dr. Kusrini, M.Kom **NIK. 190302106**	

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer
Tanggal 27 Februari 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Rafli Zuhri
NIM : 22.55.1241
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Analisis Perbandingan Algoritma Klasifikasi Untuk Identifikasi Diabetes Dengan Menggunakan Metode Random Forest Dan Naïve Bayes

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom.
Dosen Pembimbing Pendamping : Dhani Ariatmanto, M.Kom,m Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 16 Desember 2025
Yang Menyatakan,

Muhammad Rafli Zuhri

HALAMAN PERSEMBAHAN

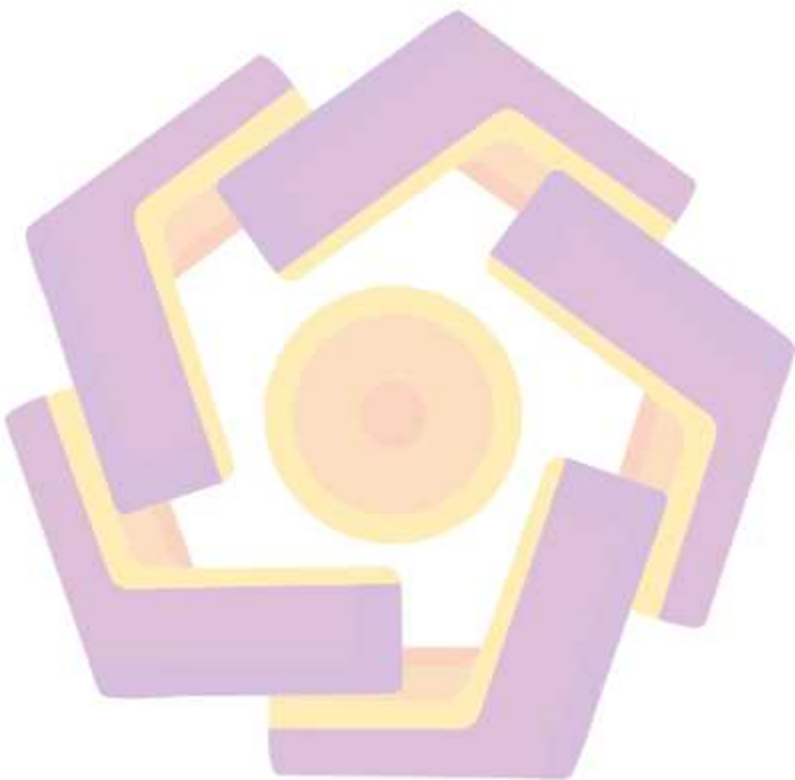
Dengan segala kerendahan hati dan rasa syukur kepada Tuhan Yang Maha Esa, tesis ini saya persembahkan kepada:

1. Kedua Orang Tua yang selalu memberikan cinta, dukungan, dan do'a yang tiada henti. Terima kasih atas segala pengorbanan, kesabaran, dan kasih sayang yang telah membentuk saya menjadi seperti sekarang;
2. Dosen Pembimbing dan Dosen Penguji, yang telah memberikan bimbingan, ilmu, dan arahan selama proses penyusunan tesis ini. Terima kasih atas kesabaran dan dedikasinya dalam membantu saya mencapai tujuan ini;
3. Teman-Teman Seperjuangan dan Rekan Kerja, yang telah menjadi sumber inspirasi dan motivasi. Terima kasih atas kebersamaan, canda tawa, dan kerja sama selama masa studi;
4. Almamater Tercinta, yang telah menjadi tempat saya menimba ilmu dan mengembangkan diri. Terima kasih atas semua pengalaman berharga yang saya dapatkan selama di sini.

Semoga karya ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan dan menjadi inspirasi bagi mereka yang akan melanjutkan perjuangan di bidang ini.

HALAMAN MOTTO

Jangan pernah menyerah, karena setiap kegagalan adalah langkah menuju kesuksesan.



KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa, atas segala rahmat, karunia, dan petunjuk-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul "**ANALISIS PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK IDENTIFIKASI DIABETES DENGAN MENGGUNAKAN METODE RANDOM FOREST DAN NAÏVE BAYES**". Tesis ini disusun untuk memenuhi salah satu syarat dalam memperoleh gelar Magister Komputer (M.Kom.) di Program Studi S2 PJJ Informatika, Universitas Amikom Yogyakarta.

Dalam penyusunan tesis ini, penulis telah menerima banyak bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan penuh rasa hormat dan terima kasih, penulis ingin menyampaikan penghargaan yang sebesar-besarnya kepada:

1. **Prof. Dr. Kusrini, M.Kom.**, selaku dosen pembimbing utama sekaligus dosen penguji ketiga, yang dengan penuh kesabaran telah memberikan motivasi, bimbingan, arahan, dan masukan yang sangat berharga selama proses penyusunan tesis ini;
2. **Dhani Ariatmanto, M.Kom,m Ph.D.**, selaku dosen pembimbing kedua, yang telah memberikan motivasi, masukan dan koreksi yang sangat berharga untuk kesempurnaan tesis ini;
3. **Hanafi, S.Kom., M.Eng., Ph.D.** selaku dosen penguji pertama, yang telah memberikan kritik, masukan, saran, dan motivasi yang sangat berarti dalam penyempurnaan dan pengujian tesis ini;

4. **Emha Taufiq Luthfi, S.T., M.Kom., Ph.D.** selaku dosen penguji kedua, yang telah memberikan kritik, masukan, saran, dan motivasi yang sangat berarti dalam penyempurnaan dan pengujian tesis ini;
5. **Prof. Dr. M. Suyanto, M.M.**, selaku Rektor Universitas Amikom Yogyakarta, yang telah memberikan kesempatan dan fasilitas kepada penulis untuk menyelesaikan studi dan penelitian ini;
6. **Seluruh Bapak dan Ibu Dosen** di Program Studi S2 PJJ Informatika, Universitas Amikom Yogyakarta, yang telah memberikan ilmu, pengalaman, dan dukungan akademik selama masa studi;
7. **Seluruh Admisi, Staff dan Karyawan** di Program Studi S2 PJJ Informatika, Universitas Amikom Yogyakarta, yang telah memberikan bantuan, support dan pelayanan yang begitu ramah dan sabar, sehingga sangat membantu selama masa studi;
8. **Keluarga Tercinta**, terutama orang tua, adik, yang selalu memberikan do'a, dukungan, dan semangat yang tiada henti, serta pengorbanan yang tak ternilai dalam perjalanan pendidikan penulis;
9. **Sahabat dan Teman-Teman**, yang selalu memberikan semangat, motivasi, dan kebersamaan selama masa studi dan penyusunan tesis ini;

Penulis menyadari bahwa tesis ini masih jauh dari sempurna. Oleh karena itu, penulis dengan tangan terbuka menerima segala bentuk saran dan kritik yang membangun demi perbaikan dan pengembangan lebih lanjut.

Akhir kata, penulis berharap semoga tesis ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan dan menjadi referensi yang berguna bagi para pembaca.

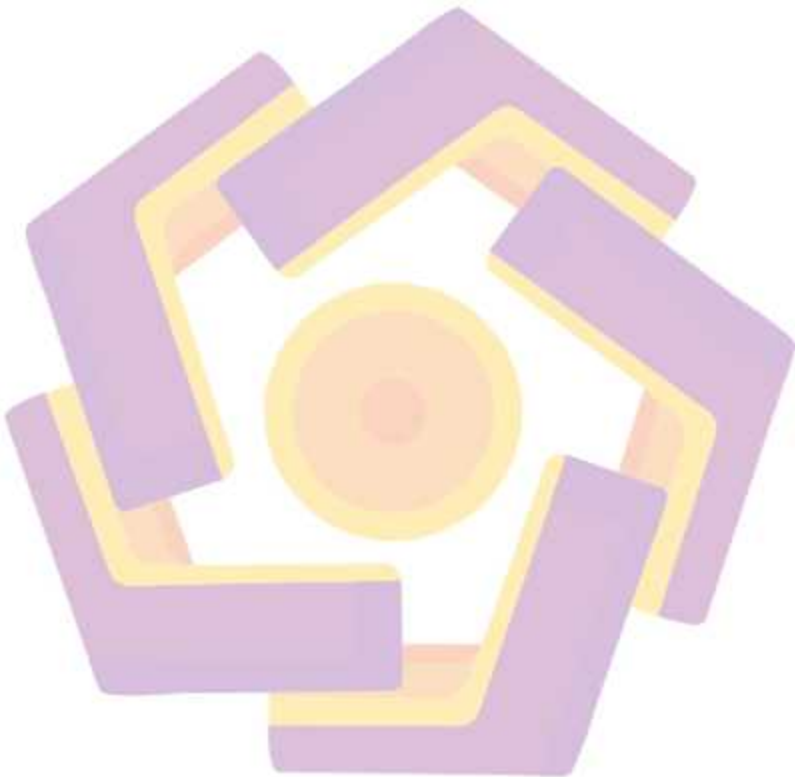
Yogyakarta, 25 Februari 2026

Muhammad Rafli Zuhri

DAFTAR ISI

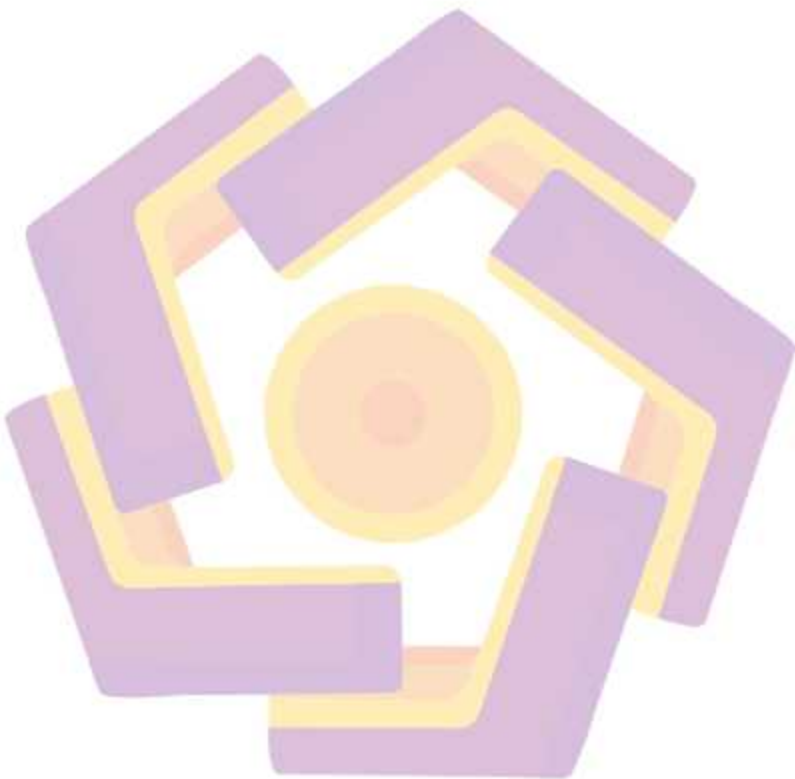
DAFTAR TABEL

DAFTAR GAMBAR



DAFTAR ISTILAH

(jika ada)



INTISARI

Penyakit diabetes kini menyerang manusia tanpa mengenal usia. Bahkan lebih dari 1,2 juta anak-anak dan remaja di dunia terkena penyakit diabetes. Penyakit diabetes pun masih masuk ke daftar penyakit paling mematikan di dunia. Penanganan penyakit diabetes menjadi penting karena komplikasi yang dapat terjadi jika tak ditanggulangi dengan benar. Oleh karena itu, pengembangan metode yang efektif dalam mendiagnosis penyakit diabetes pada perempuan menjadi sangat penting. Banyak faktor yang mempengaruhi orang menderita diabetes, beberapa diantaranya yaitu tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, usia, jumlah kehamilan seseorang, ketebalan lipatan kulit, dan jumlah kadar insulin dalam tubuh. Berbagai metode telah digunakan untuk mengidentifikasi diabetes, termasuk pemeriksaan fisik, tes darah, dan tes urine, namun hasil pemeriksaan membutuhkan waktu yang lama untuk mengetahui hasilnya sehingga dibutuhkan teknologi untuk memberikan kemudahan setiap orang dalam menerima informasi secara cepat.

Random Forest dan *Naïve Bayes* merupakan dua algoritma klasifikasi yang populer. Random Forest adalah metode kompleks yang didasarkan pada penggabungan beberapa pohon keputusan untuk mendapatkan prediksi yang lebih akurat. Sedangkan Naïve Bayes merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan denegan asumsi antar variabel penjelas saling bebas (independen). Tujuan dari penelitian ini untuk menganalisis perbandingan antara algoritma *Naïve Bayes* dan *Random Forest* dalam mengidentifikasi penyakit diabetes.

Berdasarkan hasil penelitian, *Naïve Bayes* dapat dinyatakan sebagai algoritma dengan kinerja paling stabil dalam penelitian ini karena unggul pada metrik akurasi dan presisi. Namun, *Random Forest* lebih baik dalam menjaga keseimbangan performa antara presisi dan recall sebagaimana tercermin pada nilai F1-score. Perbedaan kinerja kedua algoritma dipengaruhi oleh karakteristik dataset, keberadaan imbalance data, serta nilai ekstrem pada beberapa atribut medis. Dengan demikian, pemilihan algoritma terbaik bergantung pada tujuan penggunaan model. Jika fokus utama adalah ketepatan prediksi secara umum, maka *Naïve Bayes* lebih direkomendasikan. Namun, jika fokus diarahkan pada keseimbangan deteksi kasus diabetes, maka *Random Forest* menjadi alternatif yang lebih sesuai.

Kata kunci: diabetes, diagnosa, klasifikasi, naïve bayes, random forest

ABSTRACT

Diabetes is now attacking people regardless of age. Even more than 1.2 million children and adolescents in the world are affected by diabetes. Diabetes is still on the list of the most deadly diseases in the world. Handling diabetes is important because of the complications that can occur if not treated properly. Therefore, developing an effective method for diagnosing diabetes in women is very important. Many factors affect people suffering from diabetes, some of which are high blood pressure, excess sugar levels, weight, family history of diabetes, age, number of pregnancies, thickness of skin folds, and the amount of insulin in the body. Various methods have been used to identify diabetes, including physical examinations, blood tests, and urine tests, but the results of the examination take a long time to know the results so that technology is needed to make it easier for everyone to receive information quickly.

Random Forest and Naïve Bayes are two popular classification algorithms. Random Forest is a complex method based on combining several decision trees to obtain more accurate predictions. While Naïve Bayes is a classification method based on simple probability and is designed to be used with the assumption that the explanatory variables are independent. The purpose of this study is to analyze the comparison between the Naïve Bayes and Random Foret algorithms in identifying diabetes.

Based on the research results, Naïve Bayes can be stated as the algorithm with the most stable performance in this study because it excels in accuracy and precision metrics. However, Random Forest is better at maintaining a performance balance between precision and recall, as reflected in the F1-score value. The difference in performance between the two algorithms is influenced by dataset characteristics, the presence of data imbalance, and extreme values in several medical attributes. Therefore, choosing the best algorithm depends on the model's intended use. If the primary focus is general prediction accuracy, then Naïve Bayes is more recommended. However, if the focus is directed at balancing diabetes case detection, then Random Forest is a more suitable alternative.

Keyword: diabetes, diagnosis, classification, naïve bayes, random forest

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Diabetes adalah suatu penyakit metabolik yang diakibatkan oleh meningkatnya kadar glukosa atau gula darah. Gula darah sangat vital bagi kesehatan karena merupakan sumber energi yang penting bagi sel-sel dan jaringan. Jika tidak dikelola dengan baik, diabetes dapat menyebabkan terjadinya berbagai komplikasi, seperti penyakit jantung koroner, stroke, obesitas, serta gangguan pada mata, ginjal, dan saraf (Argina, 2020).

Diabetes mellitus (DM), menurut definisi *World Health Organization* (WHO), Karena gejalanya yang mirip dengan kondisi sakit biasa, banyak orang yang tidak menyadari bahwa mereka mengidap penyakit diabetes dan bahkan sudah mengarah pada komplikasi (Apriliah et al., 2021).

Penyakit diabetes kini menyerang manusia tanpa mengenal usia. Bahkan lebih dari 1,2 juta anak-anak dan remaja di dunia terkena penyakit diabetes. Penyakit diabetes pun masih masuk ke daftar penyakit paling mematikan di dunia(Putry, 2022). Berdasarkan data *International Diabetes Federation* (IDF), Indonesia berada dalam status waspada diabetes dan Indonesia sendiri berada di urutan ke-7 dari 10 negara dengan jumlah penderita diabetes terbanyak di dunia (Nainggolan & Sinaga, 2023).

Penanganan penyakit diabetes menjadi penting karena komplikasi yang dapat terjadi jika tak ditanggulangi dengan benar. Oleh karena itu, pengembangan metode

yang efektif dalam mendiagnosis penyakit diabetes pada perempuan menjadi sangat penting. Banyak faktor yang mempengaruhi orang menderita diabetes, beberapa diantaranya yaitu tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, usia, jumlah kehamilan seseorang, ketebalan lipatan kulit, dan jumlah kadar insulin dalam tubuh (Cahyani et al., 2022). Berbagai metode telah digunakan untuk mengidentifikasi diabetes, termasuk pemeriksaan fisik, tes darah, dan tes urine, namun hasil pemeriksaan membutuhkan waktu yang lama untuk mengetahui hasilnya sehingga dibutuhkan teknologi untuk memberikan kemudahan setiap orang dalam menerima informasi secara cepat

Salah satu contohnya yaitu penerapan teknologi dalam bidang kesehatan karena membutuhkan peralatan yang mampu memberikan diagnosa suatu penyakit dengan beberapa pertimbangan sehingga teknik data mining bisa digunakan untuk memberikan prediksi ataupun mengklasifikasikan suatu penyakit berdasarkan himpunan data yang terdapat pada rumah sakit maupun layanan kesehatan lain (Khasanah et al., 2022)

Klasifikasi merupakan salah satu metode yang dapat digunakan untuk mengidentifikasi diabetes. Algoritma klasifikasi ini dapat menganalisis data pasien, seperti usia, jenis kelamin, riwayat kesehatan, dan hasil tes, untuk memprediksi apakah pasien tersebut memiliki diabetes atau tidak(Nurussakinah & Faisal, 2023). Algoritma klasifikasi terdiri dari 5 yaitu *neural network*, K-*Nearest Neighbors*, *Decision Tree*, *Random Forest*. *Random Forest* dan *Naïve Bayes* merupakan dua algoritma klasifikasi yang populer. *Random Forest* adalah metode kompleks yang didasarkan pada penggabungan beberapa pohon keputusan untuk mendapatkan

prediksi yang lebih akurat(Supriyadi et al., 2020). Sedangkan *Naïve Bayes* merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. Pemilihan metode *Naïve Bayes* dan *Random Forest* yaitu kedua metode tersebut memiliki kinerja yang stabil untuk dataset kecil hingga sedang dan mampu menangani fitur tipe data yang campuran (numerik dan kategorical) yang sering dijumpai pada data kesehatan / rekam medis.

Penelitian ini tidak menggunakan algoritma yang lebih modern seperti *XGBoost* atau *Deep Learning* karena tujuan utama penelitian bukan untuk mengejar akurasi tertinggi, melainkan untuk menganalisis perbedaan karakteristik pendekatan klasifikasi antara model sederhana dan model *ensemble* pada data medis dengan ukuran terbatas. *Naïve Bayes* digunakan sebagai baseline probabilistik yang interpretatif, sedangkan *Random Forest* merepresentasikan model *non-linear* yang lebih kompleks dalam domain kesehatan.

Penelitian yang dilakukan oleh (Afif, 2020) dengan proses klasifikasi menghasilkan class precision = 82.35% dan class recall = 87.50%. dan hasil evaluasi klasifikasi menghasilkan *accuracy* yaitu 90,20% artinya algoritma *Naïve Bayes* terhadap dataset diabetes sudah bagus akurasinya. Pada penelitian (Apriliah et al., 2021) Hasil percobaan menentukan kecukupan sistem yang dirancang dengan akurasi yang dicapai sebesar 97,88%. Kami menemukan bahwa algoritma *Random Forest* telah bekerja dengan akurasi terbaik. Meskipun *Naïve Bayes* dan *Random Forest* telah menunjukkan kinerja yang baik dalam berbagai bidang, namun masih

diperlukan penelitian lanjutan untuk mengidentifikasi algoritma mana yang lebih unggul dalam klasifikasi penyakit diabetes Perbandingan antara *Naïve Bayes* dan *Random Forest* akan memberikan wawasan yang lebih jelas tentang kekuatan dan kelemahan masing-masing algoritma dalam konteks Penyakit diabetes. Penelitian ini mengklasifikasikan penyakit diabetes Mellitus yang termasuk suatu penyakit atau gangguan metabolism dengan kadar gula yang tinggi sertai dengan gangguan metabolism karbohidrat, lemak, dan protein sebagai akibat gangguan fungsi insulin dan penyakit diabetes gestasional penyakit diabetes yang menyerang ibu hamil.

Tujuan dari penelitian ini untuk menganalisis perbandingan antara algoritma *Naïve Bayes* dan *Random Foret* dalam mengidentifikasi penyakit diabetes. Alasan penelitian ini menggunakan dua metode tersebut karena dua metode ini merupakan metode yang paling umum digunakan dalam melakukan, kedua metode tersebut digunakan pada klasifikasi data ke dalam suatu kelas sehingga dapat membandingkan kedua metode tersebut dari segi perbedaan karakteristik pendekatan klasifikasi antara model

Penelitian ini diharapkan dapat memberikan informasi yang bermanfaat tentang kinerja *Random Forest* dan *Naïve Bayes* dalam mengidentifikasi diabetes. Hasil penelitian ini dapat membantu para dokter dan peneliti dalam mengembangkan metode yang lebih efektif untuk mengidentifikasi diabetes.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat diketahui bahwa rumusan masalah dalam penelitian ini yaitu.

a. Bagaimana hasil pengujian data menggunakan persentase perbandingan data training dan testing didapatkan pada algoritma *Random Forest* dan *Naive Bayes* dalam mengidentifikasi diabetes?

b. Bagaimana performa metode Naïve Bayes dan Random Forest dalam mengklasifikasi prediksi penyakit diabetes ?

1.3. Batasan Masalah

Batasan masalah digunakan untuk memfokuskan penelitian. Dalam penelitian ini terdapat batasan masalah sebagai berikut:

a. Penelitian ini difokuskan pada perbadingan kinerja dua algoritma *Machine Learning*, yaitu *Random Forest* dan *Naïve Bayes* dalam konteks klasifikasi penyakit diabetes.

b. Implementasi algoritma *Random Forest* dan *Naïve Bayes* menggunakan bahasa pemrograman *Pyton* dengan menggunakan *tools Google Colab*.

c. Dataset yang digunakan dalam peneltian ini bersumber dari Dinas Kesehatan atau Rumah Sakit kota Makassar.

d. Klasifikasi yang dilakukan hanya untuk penyakit diabetes.

e. Parameter yang akan dianalisis dalam penelitian ini yaitu usia, indeks massa tubuh, tekanan darah, dan hasil tes laboratorium terkait gula darah yang akan menghasilkan kategori pasien termasuk ke terkena diabetes atau tidak

1.4. Tujuan Penelitian

Bagian ini memuat penjelasan secara spesifik:

a. Mengetahui sejauh mana algoritma *Naïve Bayes* dan *Random Forest* dalam memprediksi penyakit diabetes berdasarkan data yang ada.

b. Menentukan algoritma mana yang lebih baik dalam mengidentifikasi diabetes.

c. Membandingkan kinerja algoritma Random Forest dan Naive Bayes dalam mengidentifikasi diabetes.

1.5. Manfaat Penelitian

Dalam penelitian yang dilakukan, terdapat manfaat penelitian yaitu sebagai berikut:

a. Penelitian ini akan memberikan nformasi tentang kinerja *Random Forest* dan *Naive Bayes* dalam mengidentifikasi diabetes.

b. Hasil penelitian ini dapat membantu dokter kesehatan dalam mendiagnosis diabetes dengan lebih akurat.

c. Memberikan wawasan kepada pembaca tentang penyakit diabetes.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Tinjauan Pustaka merupakan sarana untuk menunjukkan keaslian penelitian. Tinjauan pustaka merupakan pembahasan terkait penelitian terdahulu yang dijadikan referensi dalam penelitian. Tinjauan pustaka yang digunakan sebagai berikut:

Penelitian yang dilakukan oleh Ary Prandika Siregar, Dwi Priyadi Purba , Jojor Putri Pasaribu , Khairul Reza Bakara tentang klasifikasi diagonosis penyakit Stroke menggunakan algoritma *Random Forest*. Penelitian ini bertujuan untuk dengan menggunakan Metode *Random Forest* menjadi pilihan tepat dalam melakukan preprosessing data dalam mengidentifikasi gejala awal. Hasil model penyesuaian menghasilkan 96% skor pelatihan dan dari tabel hasil *precision, recall*, F1-score, dan *accuracy* Yang mendapatkan hasil akurasi sebesar 0.95 atau 95%, serta hasil akhir dari AUC sebesar 0.80 yang menunjukan hasil model tersebut termasuk ke dalam klasifikasi baik (Prandika Siregar et al., 2023).

Penelitian berikutnya Nur Farisah Patimah, Muhamad Abdurrohman, Ade Rizki Rinaldi, Arif rinaldi Dikananda tentang klasifikasi penyakit diabetes menggunakan *Naïve Bayes*. Tujuan penelitian ini adalah agar mempermudah dunia medis khususnya dokter ahli menentukan suatu klasifikasi Diabetes Mellitus kepada pasien. Hasil Penelitian *Performance vector*: *Accuracy*: 35.00% *Confusion matrix: True*: Dirawat Pulang Dirawat: 3 9 Pulang: 4 4 *Precision:* 50.00% (*Positive*

Class: Pulang) *Confusionmatrix: True*: Dirawat Pulang Dirawat : 3 9 Pulang: 4 4 *Recall*: 30.77% (*Positive Class:* Pulang) *Confusionmatrix: True:* Dirawat Pulang Dirawat : 3 9 Pulang : 4 4 (Patimah et al., 2021).

Penelitian berikutnya Mursyid Ardiansyah, Andi Sunyoto, Emha Taufiq Luthfi tentang analisis perbandingan akurasi untuk klasifikasi diabetes dengan menggunakan algoritma *Naïve Bayes* dan C4.5. Tujuan penelitian ini dilakukan untuk mengetahui hasil perbandingan nilai performa algoritma *Naïve Bayes* dan C4.5 dengan 7 skenario berbeda pada klasifikasi penyakit diabetes yang akan diuji performa *accuracy, precision*, dan *recall*. Hasil penelitian kami menunjukkan bahwa algoritma C4.5 (skenario 4) memiliki hasil yang baik dalam klasifikasi penyakit diabetes dibandingkan algoritma *Naïve Bayes* (skenario 2) dimana performa algoritma C4.5 memiliki accuracy 99.03%, precision 100%, dan recall 98.18%.

Penelitian berikutnya Fitriana Sholekhah, Adinda Dwi Putri, Rahmaddeni, Lusiana Efrizoni tentang perbandingan algoritma klasifikasi Metabolik Sindrom dengan menggunakan algoritma *naïve Bayes* dan *K-Nearest Neighbors*. Tujuan penelitian ini untuk menentukan algoritma model mana yang memiliki nilai akurasi, presisi dan recall yang lebih tinggi. Penelitian ini juga melakukan mengevaluasi tingkat akurasi dari tiga spliting data. Hasil penelitian ini menunjukkan bahwa penggunaan algoritma Naïve Bayes menghasilkan akurasi sebesar 79%, sedangkan akurasi tertinggi dari algoritma *K-Nearest Neighbors* (KNN) adalah 82%. Kesimpulannya, dari hasil penelitian ini menunjukkan bahwa algoritma K-NN

dengan pembagian data 50:50 lebih efektif dalam memprediksi dan mengklasifikasikan sindrom metabolic (Sholekhah et al., 2024).

Penelitian berikutnya Dewi Nasien, Ricalvin Darwin, Alexander Cia, Andrean Leo Winata, Jerry Go, Richard M.C, Ryan Charles Wijaya, Kevin Charles Lo, tentang perbandingan implementasi *Machine Learning* untuk klasifikasi penyakit diabetes dengan menggunakan metode *KNN, Naïve Bayes,* dan *Logistik Regression.* Tujuan penelian ini memberikan pemahaman mendalam tentang potensi dan kecocokan metode klasifikasi tertentu dalam menangani permasalahan klasifikasi data diabetes melalui pemilihan metode ekstraksi fitur yang tepat, pembagian data yang representatif, dan evaluasi kinerja yang cermat.Tujuan penelitian ini terfokus pada menganalisis kinerja tiga metode klasifikasi utama, yaitu *K-Nearest Neighbor* (KNN), *Naive Bayes*, dan *Logistic Regression*, dalam konteks pengklasifikasian data diabetes.

2.2. Keaslian Penelitian

Tabel 2. 1 Matriks literatur review dan posisi penelitian
Analisis Perbandingan Algoritma Klasifikasi Untuk Identifikasi Diabetes Dengan Menggunakan Metode *Random Forest* Dan *Naïve Bayes*

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression	Nama peneliti, Muhamad Ichsan Gunawan, Dedy Sugiarto, Is Mardianto Publikasi : Jurnal Edukasi dan Penelitian Informatika Tahun : 2020	Bertujuan untuk meningkatkan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression.	Hasil prediksi Logistic Regression tanpa Grid Search yaitu 13 data benar dan 5 data salah dari 18 data cek, dengan akurasi 72,22%. Sedangkan Prediksi Logistic Regression dengan Grid Search yaitu 15 data benar dan 3 data salah dari 18 data cek, dengan akurasi 83,33%. Sehingga terjadi peningkatan akurasi sebesar 11,11%	-Dari pemodelan Prediksi Logistic Regression denganGrid Search penelitian yang dilakukan ini, dapat digunakan dalam pembuatan aplikasi berbasis web untuk mendeteksi penyakit diabetes	Pada penelitian tersebut dapat mengklasifikan data ke dalam dua atau lebih kategori sedangkan penelitian penulis mengklasifikan ke dalam dua kategori saja sehingga proses pengolahan datanya lebih sedikit
2	Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest	Nama Peneliti : Widya Apriliah, Ilham Kurniawan, Muhamad	Merancang model yang dapat memprakirakan kemungkinan terjadinya diabetes pada pasien dengan	Hasil yang diperoleh menunjukkan Random Forest mengungguli dengan nilai akurasi tertinggi	Kedepannya, sistem yang dirancang dengan algoritma klasifikasi machine learning dapat digunakan untuk	Pada penelitian tersebut hanya menggunakan satu metode yaitu random forest sedangkan penelitian penulis menggunakan dua metode yaitu random forest dan naïve bayes untuk

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Baydhowi, Tri Haryati Publikasi :Jurnal Sistem Informasi Tahun : 2021	ketelitian yang maksimal	97,88% dibandingkan algoritma lain. Hasil ini diverifikasi menggunakan kurva Receiver Operating Characteristic (ROC) secara tepat dan sistematis.	memprediksi atau mendiagnosis penyakit lain. Penelitian dapat diperpanjang dan ditingkatkan untuk otomatisasi analisis diabetes termasuk beberapa algoritma machine learning lainnya.	membandingkan tingkat akurasi kedua metode
3	Komparasi algoritma knn dan naïve bayes untuk klasifikasi diagnosis penyakit diabetes melitus	Nama Peneliti : (Putry, 2022) Publikasi : Jurnal Sains dan Manajemen Tahun : 2022	Untuk membandingkan antara kedua algoritma tersebut yang memiliki tingkat akurasi yang terbaik	Didapatkan hasil bahwa nilai akurasi dari Naïve Bayes lebih tinggi dibandingkan KNN. Dimana nilai akurasi paling tinggi yang didapatkan dari algoritma Naïve Bayes yaitu sebesar 80%. Sedangkan algoritma KNN nilai akurasi tertinggi yaitu sebesar 75%	Dapat memberikan mengenai penyakit diabetes dan algoritma yang unggul.	Pada penelitian tersebut menggunakan metode KNN dan naïve bayes dalam mengklasifikan penyakit diabetes ini menggunakan lima pembagian data training dan data testing yang dilakukan sedangkan penelitian penulis menggunakan tiga pembagian data dari data training dan data testing
4	Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier	Nama Peneliti : Hidayat, Andi Sunyoto, Hanif Al Fatta	Untuk meningkatkan peforma model saat melakukan	Algoritma Random Forest memiliki keberhasilan tinggi untuk mengklasifikasikan	Dapat melakukan klasifikasi dengan baik dan melakukan penambahan metode algoritma	Pada penelitian tersebut menggunakan studi kasus pada penyakit jantung sedangkan penelitian penulis mengangkat studi kasus pada penyakit diabetes

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Publikasi : Jurnal Sistem Komputer dan Kecerdasan Buatan	klasifikasi peyakit jantung	sebuah kasus klasifikasi penyakit jantung. Kelebihan dari Algoritma Random *missing value* dan dapat mengatasi data dalam jumlah besar. Walaupun kekurangan dari Algoritma Random Forest adalah interpretasi yang sulit serta membutuhkan tuning model yang tepat untuk datanya.	lagi untuk melakukan proses klasifikasi penyakit jantung	
5	Implementasi Algoritma Naïve Bayes dalam Klasifikati Penyakit Diabetes	Nama Peneliti : Nur Farisah Patimah, Muhamad Abdurohman, Ade rizki Rinaldi, Arif rinaldi Dikananda Publikasi : Jurnal Data Science & Informatika Tahun : 2021	Mendapatkan klasifikasi penyakit diabetes militus menggunakan algoritma *Naïve Bayes*.	Accuracy : 35.00%, Confusion Matrix : True : dirawat Pulang Dirawat : 39 Pulang : 44 Precision : 50.00% (Positive Class: Pulang) Confusionmatrix : True : Dirawat Pulang Dirawat : 39 Pulang : 44 Recall : 30.77% (Positive Class : Pulang) Confusionmatrix : True: Dirawat Pulang Dirawat : 39 Pulang : 44.		Pada penelitian tersebut mengklasifikan data pasien ke dalam ke class pulang atau dirawat sedangkan penelitian penulis mengklasifikan ke dalam class terdiagnosa penyakit diabetes atau tidak

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6	Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke	Nama Peneliti : Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, Khairul Reza Bakara Publikasi : Jurnal Penelitian Rumpun Ilmu Teknik Tahun : 2023	Dapat membantu kalangan medis untuk dengan mudah mendiagnosa seseorang terkena penyakit stroke. Penanganannya lebih cepat jika penyakit terdeteksi lebih awal.	96% skor penelitian dan dari tabel hasil precision, recall, F1-score, dan accuracy yang mendapatkan hasil akurasi sebesar 0.95 atau 95% serta hasil AUC sebesar 0.80 menunjukan hasil klasifikasi terbaik.	Menyajikan proses uji coba dengan akurat agar dapat di implementasi dengan cukup baik.	Pada penelitian tersebut menggunakan persentase 80% data latih dan 20% data uji, sedangkan penelitian penulis menggunakan 3 pemisahan data training dan data testing yaitu 80%:20% ,70%:30% dan 50%:50% untuk mengetahui tingkat akurasi kedua metode yang terbaik
7	A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes	(Ratna Patil, 2018) Publikasi : Internasional Journal Of Electrical and Computer Engineering Tahun : 2018	Karya ini menyajikan sebuah eksperimen mempelajari beberapa algoritma yang mengklasifikasikan data Diabetes Melitus secara efektif. Algoritma yang ada dianalisis secara menyeluruh untuk mengidentifikasinya kelebihan dan keterbatasan.	Berdasarkan hasil penelitian menyimpulkan bahwa Regresi Logistik dan Peningkatan Gradien pengklasifikasi mencapai akurasi pengujian yang lebih tinggi sebesar 79% dibandingkan pengklasifikasi lainnya	Dapat menggunakan metode yang lain untuk menentukan akurasi metode yang lain	Pada penelitian ini mengklasifikan penyakit diabetes dengan menggunakan regresi logistik dengan tingkat pengujian 79% sedangkan penelitian penulis membandingkan kedua metode random forest dan naïve buyes dalam mengklasifikan penyakit diabetes

2.3. Landasan Teori

Landasan teori merupakan definisi dari teori-teori yang digunakan, berikut merupakan teori yang digunakan yaitu:

2.3.1 Penyakit

Penyakit yaitu suatu keadaan yang terjadi akibat adanya gangguan atau kegagalan mekanisme terhadap keseimbangan fungsi tubuh atau sistem tubuh sehingga tubuh tidak dalam keadaan yang normal. Penyakit merupakan suatu tekanan sehingga dapat menimbulkan gangguan pada fungsi sistem tubuh, penyakit tidak hanya dilihat dari luar saja tetapi juga adanya ketidakteraturan fungsi- fungsi dalam tubuh (Putry, 2022).

2.3.2 Diabetes

Diabetes merupakan suatu penyakit yang mengganggu metabolisme yang ditandai dengan tingginya glukosa darah. Orang yang menderita penyakit diabetes memiliki peningkatan risiko masalah kesehatan hingga mengancam jiwa yang mengakibatkan biaya perawatan medis, penurunan kualitas hidup dan peningkatan kematian (Nam Han Cho et al., 2017). Penyebabnya adalah kekurangan hormon insulin. Hormon adalah unsur kimia yang dibuat oleh tubuh. dalam hal ini pankreas) dan dilepas ke dalam aliran darah untuk digunakan oleh bagian tubuh yang membutuhkannya. Ada orang yang sama sekali tak dapat menghasilkan insulin seperti pada diabetes tipe 1. Namun pada tipe 2, mungkin insulin hanya diproduksi sedikit, dan respon tubuh terhadap hormon itu menurun (Informatika et al., 2008). Penyakit diabetes memberikan kontribusi sebagai salah satu penyebab kematian

seseorang pada penderita penyakit jantung (Adnan et al., 2013). Penyakit diabetes dikelompokkan beberapa bagian yaitu (Punthakee et al., 2018):

1. Diabetes melitus tipe 1 meliputi diabetes disebabkan oleh penghancuran sel β pankreas baik oleh proses autoimun maupun idiopatik sehingga produksi insulin berkurang bahkan berhenti. Biasanya diabetes tipe ini terjadi pada usia kurang dari 20 tahun
2. Diabetes melitus tipe 2 merupakan diabetes dengan kelainan metabolik yang ditandai dengan kadar glukosa darah yang tinggi dalam konteks resistensi insulin dan defisiensi insulin relatif. Diabetes tipe ini biasanya di derita pada usia lebih dari 20 tahun
3. Diabetes melitus gestasional mengacu pada intoleransi glukosa pada masa pengenalan pertama selama kehamilan.
4. Jenis spesifik lainnya mencakup berbagai macam kondisi tidak umum, terutama bentuk diabetes yang ditentukan secara genetik atau diabetes yang terkait dengan penyakit lain atau penggunaan narkoba.

Berikut adalah gejala gejala apabila seseorang menderita diabetes (Maulidah et al., 2021)

1. Cepat haus
2. Penurunan berat badan
3. Rasa lelah yang tak biasa
4. Pandangan kabur
5. Pemulihan luka yang lama
6. Warna kulit gelap

2.3.3 Data Mining

Data *mining* adalah proses untuk mengidentifikasi dan mengekstraksi informasi yang berguna dan pengetahuan terkait bersumber dari basis data besar menggunakan pendekatan matematika, statistik, kecerdasan buatan, dan pembelajaran mesin. Data mining merupakan sekumpulan prosedur yang digunakan untuk menemukan nilai tambah dari sumber data berupa pengetahuan yang sebelumnya tidak diketahui. Data mining melakukan pencarian informasi dari kumpulan data yang besar. Pencarian tersebut dilakukan dengan cara analisis, pengumpulan data, menemukan pola serta hubungan yang ada didalamnya.

Dalam penggunaan data mining diperlukan pemahaman dalam penyelesaian masalah dalam melakukan pemilihan pengolahan data yang masuk ke dalam tahap klasifikasi, regresi, clustering dan sebagainya. Banyaknya teknik data mining dan jenis informasi maka diperlukan penerapan dan relevansi metode sesuai data yang disiapkan dan tujuan yang ingin dicapai (Vadim, 2018)

Data mining memiliki dua tujuan utama yaitu prediksi dan deskripsi. Prediksi melibatkan penggunaan beberapa atribut dalam dataset untuk memprediksi nilai yang tidak diketahui dari atribut lain yang relevan. Sementara itu, deskripsi melibatkan penemuan pola dan tren dalam suatu data. Dengan bantuan teknik data mining dapat diketahui pola suatu penyakit berdasarkan data yang sudah ada seperti nama pasien, usia, jenis kelamin, dan lainnya. Dengan demikian, jika sudah diketahui faktor-faktor yang mempengaruhi suatu diagnosa penyakit, maka memudahkan untuk klasifikasi penyakit (Ente et al., 2020).

2.3.4 *Machine Learning*

Machine Learning adalah merupakan salah satu komponen utama dalam kecerdasan buatan yang memungkinkan sistem untuk belajar dari data dan mengambil keputusan atau membuat prediksi. Esensi dari pembelajaran mesin terletak pada kemampuannya dalam mengenali pola serta hubungan dalam data, yang selanjutnya dimanfaatkan untuk menghasilkan keputusan yang lebih akurat di masa mendatang. Teknologi ini sangat berguna di berbagai bidang, termasuk medis, di mana dapat mengelola data kompleks dan besar untuk menciptakan model prediksi yang akurat (Diana et al., 2023). Dengan kemampuannya untuk memproses data yang beragam dan rumit, Machine Learning dapat mengidentifikasi pola yang tidak terlihat dengan metode tradisional, membantu dalam pencegahan dan manajemen penyakit. Sebagai contoh, teknik ini bisa mengungkap faktor risiko untuk penyakit kronis berdasarkan data medis yang luas, mendukung pengembangan strategi pencegahan yang lebih tepat (Fangatulo Dodo Telaumbanua et al., 2019).

Terdapat beberapa macam dari *machine learning* yang diantaranya yaitu *supervised learning, unsupervised leaning*, dan *reinforcement learning*. *Supervised learning* membangun fungsi input-output berdasar pada data yang ada untuk dipelajari algoritma berdasarkan data training yang diberi label dengan tujuan untuk generalisasi data input. Sedangkan *unsupervised learning* tidak dilakukan pemberian sebuah label dari kumpulan data serta tidak membutuhkan data training. Sedangkan reinforcement learning berada diantara *supervised learning* dan *unsupervised learning*. *Reinforcement learning* digunakan pada sejumlah data

dengan ukuran besar yang dibagi ke dalam dua bagian yang tidak diberi label dan diberi label dimana konsepnya menyelesaikan suatu tujuan dengan tanpa ada pemberitahuan perangkat komputer dengan secara eksplisit apabila tujuan itu sudah tercapai (Roihan et al., 2019)

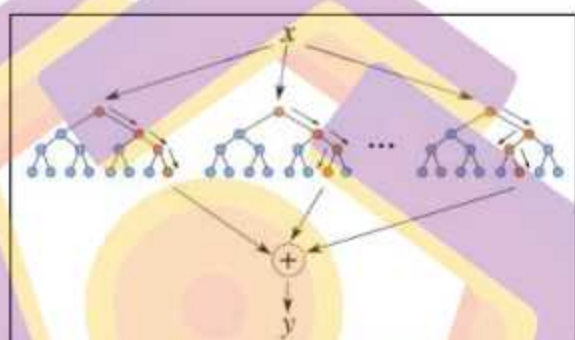
2.3.5 Klasifikasi

Klasifikasi merupakan pengelompokan record data baru ke salah satu dari beberapa kelas yang sebelumnya telah didefinisikan. Pengelompokan data didasarkan sepenuhnya pada ciri-ciri yang dimiliki sesuai dengan kelasnya. Dalam prosesnya, klasifikasi dapat diselesaikan dengan banyak cara baik secara manual maupun dengan bantuan teknologi. Klasifikasi yang diselesaikan secara manual adalah klasifikasi yang diselesaikan dengan bantuan manusia tanpa bantuan algoritma komputer (Amari, 2023). Klasifikasi diberikan sejumlah record yang dinamakan data pelatihan, yang terdiri dari beberapa atribut, salah satu atribut menunjukkan kelas untuk record yang dapat digunakan untuk menemukan model dari data pelatihan sehingga dari hasil tersebut kita dapat membedakan record ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan record yang kelasnya belum diketahui sebelumnya (Sunjana, 2010). Ada beberapa jenis klasifikasi yaitu *Decision Tree, rule-based classifiers, Bayesian classifier Support Vector Machines, Artificial Neural Networks, Lazy Learners,* dan *ensemble methods* (Azhari et al., 2021)

2.3.6 *Random Forest*

Random Forest merupakan hasil pengembangan dari metode *Classification* dan *Regression Tree* (CART), yang menggunakan metode *bag or bootstrap*

aggregation dan *random feature selection*. *Bagging* merupakan salah satu teknik yang dapat digunakan untuk memperbaiki hasil dari suatu algoritma klasifikasi. Metode *bagging* ini didasarkan pada metode *ensemble*. Metode algoritma hutan acak dapat dibagi menjadi dua tahap, tahap pertama melibatkan pembuatan "k" pohon untuk membentuk hutan acak, sedangkan tahap kedua menggunakan hutan acak yang telah dibentuk untuk membuat prediksi.



Gambar 2. 1 *Random Forest*

Dalam konstruksi pohon keputusan menggunakan metode *CART*, komputasi melibatkan verifikasi informasi yang menjelaskan seberapa penting atribut dalam mengklasifikasikan setiap simpul pohon. Secara khusus, jika kita menganggap N sebagai simpul yang memisahkan kelas data D berdasarkan atribut-atributnya, maka komputasi ini membantu mengukur seberapa relevan atau informatif setiap atribut tersebut dalam proses pemisahan kelas data (Hidayat et al., 2023). Random Forest merupakan salah satu metode yang dapat meningkatkan hasil akurasi dalam membangkitkan atribut untuk setiap node yang dilakukan secara acak. RF terdiri dari sekumpulan *decision tree*, dimana kumpulan pohon keputusan ini digunakan untuk mengklasifikasi data ke suatu kelas. Pohon keputusan dibuat dengan

menentukan node akar dan berakhir dengan beberapa node daun untuk mendapatkan hasil akhir (Alita & Rahman, 2020). Karakteristik dari metode *random forest* yaitu kemampuan menangani data yang komples yang menggabungkan sejumlah *decision tree* untuk meningkatkan akurasi dan mengurangi risiko *overfitting* (Anisya et al., 2020) .

Rumus yang digunakan untuk menghitung tingkat informasi validasi adalah sebagai berikut :

$$Gain\ (A) = Info\ (D) - Info\ (D) \quad (1)$$

Untuk mendapatkan nilai info(D), kita dapat menghitungnya menggunakan rumus 2 dan 3, yang akan menghasilkan nilai info A(D) :

$$Info\ (D) = -\sum_{i=1}^{n} p_i \log 2(p_i) \quad (2)$$

Keterangan :

n = jumlah kelas target

pi = proporsi kelas i terhadap partisi D

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ info(D_j) \quad (3)$$

Keterangan :

v = jumlah partisi.

Dj = total partisi ke j.

D = total baris pada semua partisi.

2.3.7 *Supervised Learning*

Supervised Learning merupakan salah satu tipe dari *Machine Learning*. Model ini digunakan ketika dataset yang dimiliki memiliki label setiap input atau outputnya. Hasil dari training data digunakan untuk memprediksi output yang

benar. Pada penerapannya, proses supervised learning yaitu pengumpulan data, split data, pemilihan model, evaluasi dan uji program dengan prediksi data baru (Nicholas, 2024). Algoritma klasifikasi pada *supervised learning* ada beberapa macam antara lain *logistic regression, support vector machine, decision tree*, dan *random forest* yang masing-masing memiliki karakteristik dan tujuan yang berbeda sesuai dengan studi kasus (Abdi, 2018).

a. *Logistic Regression*

   Algoritma ini digunakan untuk klasifikasi biner yang dapat diklasifikasikan ke multi-class. Contoh penerapan algoritma ini adalah pengenalan spam email atau prediksi kehilangan pelanggan

b. *Decision Tree*

   Dalam algoritma ini, model membuat serangkaian keputusan dalam mencapai final decision. Algoritma ini tidak memerlukan normalisasi data dan dapat menangani data numerik maupun kategori. Contoh penerapan algoritma ini adalah prediksi keputusan atau klasifikasi penyakit

c. *Random Forest*

   Algoritma *random forest* berkaitan dengan *decision tree*. Algoritma ini akan menggabungkan beberapa *decision tree* untuk meningkatkan kinerja dan menghindari overfitting data. Contoh penerapannya adalah klasifikasi obyek dan klasifikasi penyakit

d. *Naïve Bayes*

   Naïve Bayes adalah sebuah metode klasifikasi yang berdasarkan pada *teorema bayes* dengan asumsi bahwa fitur-fitur yang digunakan untuk klasifikasi saling

independen. Metode ini dinamakan naïve (sederhana) karena mengasumsikan bahwa setiap pasangan dalam data pelatihan adalah independen, meskipun dalam hal ini mungkin tidak selalu terjadi. Contoh penerapannya yaitu klasifikasi teks dan diagnosa penyakit medis.

2.3.8 *Naïve Bayes*

Naive Bayes Classifier (NBC) ialah salah satu metode klasifikasi yang sering digunakan. *Naive Bayes Classifier* (NBC) adalah salah satu metode klasifikasi dan statistik pengklasifikasi yang dapat memprediksi peluang untuk menjadi anggota kelas (Yudha Prawira et al., 2024). *Naive Bayes* merupakan suatu proses klasifikasi probabilistik dengan teorema Bayes, menganggap bahwa tiap feature mempunyai kinerja sama pada kelas target. Klasifikasi *Naive Bayes*, yang termasuk dalam kelas tertentu yang berkontribusi pada pengambilan sampel probabilistik, mudah diterapkan, cepat dihitung, dan cocok untuk kumpulan data berdimensi tinggi yang besar. Cocok untuk aplikasi *real-time* dan tahan terhadap gangguan. Klasifikasi *naïve bayes* mengolah kumpulan data trainer untuk menjumlahkan probabilitas kelas dan probabilitas bersyarat yang dapat menentukan berbagai frekuensi setiap nilai feature untuk nilai class tertentu. Pengklasifikasi *Naive Bayes* bekerja paling baik ketika mereka berkorelasi. Dikarenakan fitur korelasi dipakai dua kali dalam model, fitur tersebut disembunyikan, yang terlalu menekankan dibutuhkannya fitur yang berkorelasi. Karakteristik metode naïve bayes menggunakan pendekatan probabilistik berbasis *Teorema Bayes* dengan asumsi independensi antar fitur. Hal ini membuat algoritma ini sangat efisien untuk dataset dengan jumlah atribut yang tidak terlalu besar dan tidak memerlukan proses

komputasi yang kompleks.(Zainal Macfud et al., 2023). Pendekatan Nave Bayes memiliki keuntungan karena hanya menggunakan sedikit data training untuk mendapatkan estimasi parameter yang diperlukan dalam proses klasifikasi. Oleh karena itu, terlepas dari variabel independen yang dipertimbangkan, hanya varian dari variabel kelas yang diperlukan untuk menentukan kategorisasinya, bukan seluruh variabel kelas (Widya Ningsih, 2020)

$$P(c|x) = \frac{p(X|C)\,p(c)}{p(x)} \quad (4)$$

$$P(c|x) = P(x_1|\,c)\;x\;P(x_2|\,c)\;x\;P(x_3|\,c)\;x\;\ldots\ldots x\;(P(x_n|\,c)\;x\;P(c) \quad (5)$$

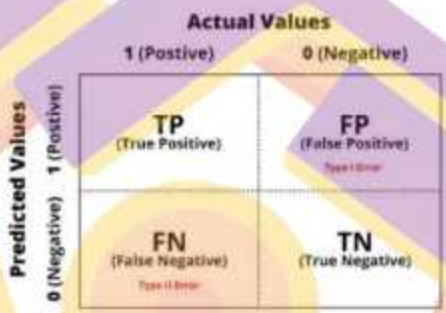
Dalam persamaan (1) dan (2), teorema *Bayes* memudahkan untuk menghitung probabilitas posterior dari P(c|x) dari P(c), P(x), dan P(x|c). klasifikasi dari *naive bayes* menganggap efek nilai prediktor (x) pada class tertentu (c) tidak bergantung prediktor yang lain. Anggapan ini dikenal dengan independensi class bersyarat. dimana P(c) merupakan probabilitas kelas sebelumnya, P(x) merupakan probabilitas sebelumnya dari predictor, P(c|x) merupakan probabilitas posterior kelas (target)diberikan prediktor (atribut), dan P(x|c) merupakan peluang yang merupakan peluang kelas yang diberikan oleh prediktor (Fathurahman et al., 2023).

2.3.9 *Confusion Matrix*

Confusion matrix yaitu tabel yang mengkategorikan prediksi berdasarkan seberapa dekat kesesuaiannya dengan nilai aktual data. Salah satu dimensi tabel menunjukkan kemungkinan kategori nilai prediksi sementara dimensi lainnya menunjukkan hal yang sama untuk nilai sebenarnya. Meskipun sampai saat ini matrix yang sering digunakan yaitu *confusion matrix* berukuran 2 x 2, sejumlah kelas dapat diprediksi oleh model menggunakan matriks. Matriks konfusi 3 x 3

model tiga kelas ditunjukkan pada gambar terlampir bersama dengan matriks konfusi yang terkenal untuk model biner dua kelas.

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi performa model klasifikasi *machine learning*. Matriks ini menyajikan ringkasan hasil prediksi model pada sebuah dataset, memungkinkan kita mengevaluasi seberapa akurat atau salah model dalam mengklasifikasikan kumpulan data.



Gambar 2. 2 *Confusion Matrix*

Berdasarkan *Confusion Matrix*, Berikut Rumus untuk menghitung matrik evaluasi klasifikasi:

- *Accuracy* (Akurasi)

  Akurasi merupakan metrik yang menggambarkan seberapa akurasi model dapat mengklasifikasikan dengan benar. Dalam hal ini merupakan rasio prediksi benar dengan keseluruhan data atau tingkat kedekatan nilai prediksi dengan nilai aktual. Adapun persamaan untuk memperoleh nilai accuracy sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

- *Precision* (presisi)

  Presisi merupakan metrik yang menggambarkan tingkat keakuratan antara data yang diminta dengan jasil prediksi yang diberikan oleh model klasifikasi dengan kata lain *precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Adapun persamaan untuk memperoleh nilai *precision* sebagai berikut:

$$Precision : \frac{TP}{TP+FP} \quad (7)$$

- *Recall*

  Recall merupakan metrik yang menggambarkan keberhasilan dari mode klasifikasi dalam menemukan informasi, *recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Adapun persamaan untuk memperoleh recall sebagai berikut:

$$Recall : \frac{TP}{TP+FN} \quad (8)$$

- F1-Score

  F1-Score merupakan metrik yang menggambarkan rata-rata harmonis antara nilai *precision* dengan *recall*. Semakin besar nilai F1 yang dihasilkan maka semakin baik performasinya. Adapun persamaan untuk memperoleh nilai *F1-Score* sebagai berikut:

$$F1\text{-}Score : 2x\frac{Precesion\ X\ Recall}{Precesion+Recall} \quad (9)$$

2.3.10 *Python*

Python merupakan salah satu bahasa pemrograman yang banyak digunakan oleh perusahaaan besar maupun para developer untuk mengembangkan berbagai

macam aplikasi berbasis desktop, web dan mobile. *Python* adalah bahasa pemrograman yang semakin populer dan gratis bagi para ilmuwan. Ini menggabungkan sintaks sederhana dan sumber daya online yang melimpah. Sebagai bahasa pemrograman tujuan umum, bahasa ini tidak memiliki dukungan khusus untuk struktur data atau algoritma ilmiah, tidak seperti platform ilmiah seperti *Matlab* atau R. Namun, hal ini menyediakan ekosistem yang kaya akan perangkat yang berfokus pada sains dengan dukungan komunitas yang kuat. Salah satu fitur *Python* yang paling menarik bagi para ilmuwan adalah kemungkinan untuk langsung menjalankan perpustakaan yang dikodekan dalam bahasa seperti *Fortran* dan C, tanpa harus khawatir tentang pemrograman dalam bahasa tingkat rendah tersebut. Hal ini memungkinkan penggunaan kembali berbagai paket ilmiah yang diterapkan dalam bahasa tersebut, seperti BLAS atau LAPACK. Selain itu, karena *Python* adalah bahasa yang ditafsirkan, dan oleh karena itu lambat dibandingkan dengan bahasa yang dikompilasi, hal ini secara dramatis meningkatkan efisiensi, yang merupakan aspek kunci dalam sains. Inilah alasan utama mengapa sebagian besar paket ilmiah *Python* tidak ditulis dengan *Python* itu sendiri, atau setidaknya di bagian terpentingnya (Magaña et al., 2020).

2.3.11 *Google Colaboratory*

Google Colaboratory, sering disebut Google Colab, merupakan sebuah platform berbasis web yang memberikan akses gratis ke sumber daya komputasi berbasis cloud untuk keperluan analisis data dan pelatihan model dalam bidang pembelajaran mesin. Dengan menggunakan *Google Collaboratory*, pengguna dapat menulis dan mengeksekusi kode *Python*, membuat dan berbagi buku catatan, serta

berkolaborasi dengan orang lain secara real-time. Platform tersebut menyediakan beragam perpustakaan terintegrasi, seperti *TensorFlow* dan *PyTorch*, yang sangat sesuai untuk melakukan berbagai tugas seperti manipulasi data, visualisasi, pengembangan model pembelajaran mesin, dan proses pelatihan. *Google Colab* menyediakan lingkungan yang nyaman dan efisien bagi pengembang dan peneliti untuk mengerjakan proyek pembelajaran mesin mereka (Mashudi et al., 2022).

Platform ini sangat menguntungkan bagi pemula di bidang pemrograman komputer karena antarmukanya yang ramah pengguna dan dokumentasi yang ekstensif. *Google Colab* adalah layanan sumber terbuka yang disediakan oleh *Google* untuk semua pemilik akun Gmail, menawarkan GPU untuk penelitian terkait pembelajaran mesin. Selain itu, *Google Colab* menawarkan opsi untuk mengupgrade ke *Google Colab Pro*, yang memberikan performa lebih cepat dan waktu proses lebih lama untuk tugas yang lebih intensif. Singkatnya, *Google Collaboratory*, adalah platform berbasis *cloud* yang memungkinkan pengguna menulis dan mengeksekusi kode *Python* untuk tugas analisis data dan pembelajaran mesin, menyediakan akses ke sumber daya komputasi yang kuat dan berbagai perpustakaan. Ini adalah alat yang nyaman dan efisien bagi pengembang dan peneliti, terutama pemula, karena antarmukanya yang ramah pengguna dan dokumentasi yang ekstensif (Rosita et al., 2022).

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Penelitian ini menggunakan jenis penelitian eksperimen yaitu: meneliti perbandingan tingkat akurasi, Presisi, dan Recall pada metode algoritma *Random Forest* dan *Naive bayes* untuk klasifikasi penyakit diabetes. Penelitian yang dilakukan bersifat deskriptif karena menggambarkan atau menganalisis secara rinci karakteristik suatu fenomena tanpa mengambil keputusan sebab-akibat. Penelitian deskriptif dimana data dijelaskan dalam bentuk angka dan tabel atau diagram. Setelah data diolah akan dilakukan analisis dengan pendekatan kuantitatif yang dijelaskan dengan hasil perhitungan angka dan tabel atau diagram. Pada penelitian ini menggunakan pendekatan kuantitatif yang nantinya hasil dari penelitian ini merupakan informasi-informasi berupa angka dan diagram hasil dari eksperimen penggabungan dua metode yang dilakukan. Pengumpulan data dilakukan melalui hasil eksperimen yang kemudian data tersebut dilakukan penggabungan dan analisis seperti dibuatkannya tabel dan diagram untuk melihat penggabungan metode mana yang paling baik dalam klasifikasi penyakit diabetes.

3.2. Metode Pengumpulan Data

Metode penumpulan data merupakan cara yang digunakan untuk mengumpulkan data yang diperlukan dalam penelitian. Berikut metode penelitian yang digunakan yaitu :

1. Pengumpulan Data :

   Penelitian ini mengumpulkan data pasien di *kaagle* yang akan dijadikan sebagai sampel data training dan data testing yang akan mengelompokkan dari dua kelompok pasien : pasien yang telah terdiagnosa dengan diabetes dan pasien yang tidak terdiagnosa dengan diabetes. Data ini mencakup beberapa atribut yang relavan dalam analisis penyakit diabetes, seperti usia, indeks massa tubuh, tekanan darah (data privat) dan hasil tes laboratorium terkait gula darah (public).

2. Pengelompokkan Data

   Setelah data terkumpul, selanjutnya data dikelompokan ke dalam data training dan data testing. Adapun pembagian data training dan data testing ini 80%:20%, 70%:30% dan 50%:50% dari total keseluruhan data untuk mengetahui tingkat dari akurasi dari metode yang digunakan.

3.3. Metode Analisis Data

Analisis data yang dilakukan dalam penelitian ini menggunakan teknik data mining dengan menerapkan algoritma *Naïve Bayes* dan *Random Forest* untuk dilakukan komparasi akurasi dalam klasifikasi penyakit diabetes.

1. *Preprocessing* Data

   Pada tahap ini dilakukan pembersihan data dengan menghilangkan data membersihkan data, menghilangkan noise dan nilai yang hilang sebelum melakukan langkah pemodelan. Pembersihan data menghapus data dari nilai yang hilang dan outlier. Pemisahan data training dan data testing

menggunakan pemisahan data, yaitu 80% untuk data train saat melatih model dan 20% untuk data uji saat menguji model.

2. Implementasi Algoritma

   Pada tahap ini mengimplementasikan metode yang akan digunakan yaitu metode *naïve bayes* dan *random forest* untuk mengklasifikan data pasien ke dalam class non diabetes dan diabetes.

3. Evaluasi kinerja model

   Mengevaluasi kinerja kedua algoritma pada dataset testing menggunakan confusion matrix untuk mencari nilai *true positif* (TN), *true negatif* (TN), *false positif* (FN) dan *false negatif* (FN) serta menggunakan metrik seperti akurasi, presisi, *recall* dan F1-*score* dan membandingkan kinerja kedua algoritma berdasarkan metrik yang telah dihitung. Presisi mengukur seberapa banyak pasien diabetes yang benar-benar terklasifikasikan dengan benar. *Recall* mengukur seberapa banyak pasien diabetes yang terklasifikasikan dengan benar dari total pasien diabetes yang sebenarnya. F1-*score* adalah ukuran yang menggabungkan presisi dan *recall*. untuk memberikan gambaran yang lebih komprehensif tentang kinerja model.

4. Analisis Hasil

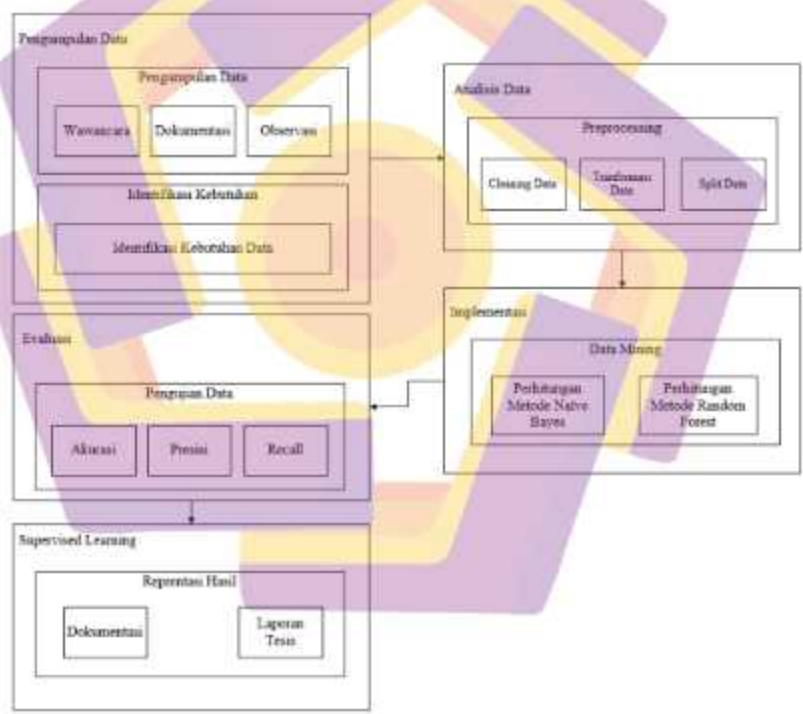
   Menarik kesimpulan dari evalusi kinerja model untuk mengetahui metode yang memiliki kinerja yang baik dalam mengindentifikasi pasien yang terkena diabetes

5. Pelaporan

Menyajikan temuan dan rekomendasi dalam laporan yang berguna untuk pengambilan keputusan di bidang kesehatan

3.4. Alur Penelitian

Bagian ini berisi diagram alur langkah penelitian secara lengkap dan terinci termasuk di dalamnya tercermin algoritma, rute, pemodelan-pemodelan, desain, yang terkait dengan aspek perancangan sistem.



Gambar 3. 1 Alur Penelitian

1. Pengumpulan Data :

   Pada tahap pengumpulan data ini melakukan pengumpulan data dengan mengambil data sampel pasien berupa adalah data usia, indeks massa tubuh (BMI), tekanan darah, dan hasil tes laboratorium terkait gula darah yang kemudian dikelompokkan ke dalam pasien yang terdiagnosa diabetes dan tidak terdiagnosa untuk kemudian dilakukan identifikasi kebutuhan data yang sesuai dengan sistem yang akan dibangun.

2. Analisis Data :

   a. Cleaning Data, Untuk tahapan cleaning data dilakukan proses pembersihan data dengan menghilangkan *missing value* atau bisa disebut data tidak berisi atau kosong (*null*), serta data yang tidak lengkap. Kemudian data yang akan dibersihkan akan melalui proses *cleaning* dengan melalui cara *Replace Missing Values* untuk mengisi nilai rata-rata atribut tertentu disetiap daerah yang kosong yang mengacu pada atributnya.

   b. Transformation Data, Tahap ini dilakukan proses transformasi atau normalisasi data kedalam format yang dapat dikelola oleh sistem. Dengan menggunakan *Google Colab* dengan mengubah format data awal sesuai dengan kategori dikarenakan analisis asosiasi hanya bisa menerima input data kategorikal.

   c. Split Data. Tahap ini membagi data ke dalam data training dan data testing, data training yaitu data yang digunakan sebagai data pelatihan dan data testing yang akan sebagai data uji dalam pemisahan data

dilakukan persentase 70%;30% dari total data keseluruhan hasil dari *pre-processing*.

3. Implementasi

   Melakukan perhitungan *random forest* dan perhitungan *naïve bayes* untuk mencari akurasi perbandingan dari kedua metode tersebut.

4. Evaluating

   Melakukan pengujian hasil dari implementasi *naïve bayes* dan *random forest* dengan menggunakan metode *Confusion Matrix* untuk mencari nilai akurasi, presisi, *recall* dan F1-score untuk melihat tingkat persentase dari kedua metode

5. *Specifying Learning*

   Pada tahap ini dilakukan proses dokumentasi dan publikasi thesis berisi hasil peneltian yang sudah diterapkan.

6. Pelaporan

   Menyusun laporan Thesis

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil Penelitian

Pada bagian ini penulis akan menjelaskan hasil dari penelitian yang telah dilakukan berdasarkan alur penelitian yang penulis buat sebelumnya

4.1.1 Pengumpulan Data

Pada tahap ini dilakukan pengambilan sampel data pasien seperti pada gambar 4 dibawah ini yang di ambil dari kaagle yang berisi 8 variabel dan 1 class hasil yang terdiri dari 768 data. Detail atribut pada penelitian ini seperti pada tabel 4.1 dibawah ini

Tabel 4. 1 Fitur Data Set

No	Fitur	
	Atribut	Deskripsi
1	Kehamilan	Menunjukkan data jumlah kehamilan
2	Glukosa	Menunjukkan kadar glukosa pada darah
3	Tekanan darah	Menunjukkan pengukuran dari tekanan darah
4	Ketebalan Kulit	Menunjukkan ketebalan kulit
5	Insulin	Menunjukkan nilai insulin pada darah
6	BMI	Menunjukkan hasil pengukuran berat badan
7	Riwayat diabetes	Menunjukkan persentase dari riwayat diabetes
8	Umur	Menunjukkan umur dari pasien

Pada tabel 4.1 merupakan variabel yang digunakan dalam penelitian yang terdiri dari attribut kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes dan umur, variabel tersebut nantinya sebagai bahan perhitungan dan pertimbangan dalam mengidentifikasi penyakit diabetes seseorang

Tabel 4. 2. Kelas Dataset

No	Fitur	
	Atribut	Deskripsi
1	Hasil	Menunjukkan variabel klasifikasi ketika 0 artinya tidak menderita diabetes dan 1 berarti menderita diabetes

Pada tabel 4.2 merupakan pengelompokkan kelas yang digunakan yang terdiri dari angka 1 jika pasien menderita diabetes dan 0 jika tidak menderita diabetes

Data penelitian secara keseluruhan dapat dilihat pada tabel 4.3 dibawah ini

Tabel 4. 3 Data Penelitian

No	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil
1	6	148	72	35	0	33,6	0,627	50	1
2	1	85	66	29	0	26,6	0,351	31	0
3	8	183	64	0	0	23,3	0,672	32	1
4	0	89	66	23	94	28,1	0,167	21	0
5	0	137	40	35	168	43,1	2,288	33	1
...	.....	....	...	....	....	...	.....	...	...
764	10	101	76	48	180	32,9	0,171	63	0
765	2	122	70	27	0	36,8	0,340	27	0
766	5	121	72	23	112	26,2	0,245	30	0
767	1	126	60	0	0	30,1	0,349	47	1
768	1	93	70	31	0	30,4	0,315	23	0

Pada gambar tabel 4.3 merupakan data yang peneliti gunakan dalam penelitian yang terdiri dari 768 data yang terdiri 8 variabel yaitu kehamilan: data jumlah kehamilan pasien, glukosa : data kadar glukosa dalam darah, tekanan darah: pengukuran tekanan darah, ketebalan kulit: menyatakan ketebalan kulit, insulin : pengukuran kadar insulin dalam darah, BMI : indeks massa tubuh, riwayat diabetes : persentase

penyakit diabetes , umur : menyatakan umur pasien, sedangkan hasil yaitu variabel klasifikasi dimana 0 jika tidak mengidap diabetes dan 1 jika mengidap diabetes

Tabel 4. 4 Keterangan Dataset

Variabel	Keterangan
Kehamilan	Kehamilan, berapa kali pasien hamil
Glukosa	Jumlah konsentrasi glukosa plasma selama 2 jam dalam tes toleransi glukosa oral
Tekanan Darah	Tekanan darah diastolik (mmHg)
Ketebalan Kulit	Ketebalan lipatan kulit trisep (mm)
Insulin	Insulin serum dua jam (µ1U/mL)
BMI	Indeks massa tubuh (kg/m2)
Riwayat Diabetes	Menilai kemungkinan diabetes berdasarkan riwayat keluarga
Umur	Usia tahun ini
Hasil	kelas 0 jika non-diabetes dan kelas 1 jika diabetes

4.1.2 Analisis Data

a. *Pre-processing* Data

1. *Cleaning* Data

Pada tahap ini dilakukan pembersihan data dengan menghilangkan *missing value* atau bisa disebut data tidak berisi atau kosong (*null*), serta data yang duplikat

```
Jumlah missing value per kolom:
 Kehamilan           0
Glukosa              0
Tekanan Darah        0
Ketebalan Kulit      0
Insulin              0
BMI                  0
Riwayat Diabetes     0
Umur                 0
Hasil                0
```

Gambar 4. 1 Hasil *Missing Values*

Pada gambar 4.1 merupakan hasil *missing values* atau data yang kosong pada gambar diatas terlihat bahwa hasilnya menunjukkan dataset yang akan dianalisis tidak memiliki data yang kosong sehingga tidak ada data yang dibuang/dihapus. Sehingga struktur dataset yang akan digunakan yaitu :

- Total Data : 768 data
- Total Fitur : 9 kolom

Tidak ada *missing values* (kosong) disemua kolom

2. *Outlier* (Nilai Menyimpang)

Outlier pada dataset merupakan nilai data yang sangat berbeda dibandingkan dengan sebagian besar data lainnya. *Outlier* biasanya muncul dikarenakan adanya kesalahan data atau pengukuran data. Berikut adalah outlier pada penelitian ini

Tabel 4. 5 Data *Outlier*

No	Fitur	*Outlier*
1	Kehamilan	4
2	Glukosa	5
3	Tekanan Darah	45
4	Ketebalan Kulit	1
5	Insulin	34

6	BMI	19
7	Riwayat Diabetes	29
8	Umur	9

Fitur Tekanan Darah, Insulin, BMI dan Riwayat Diabetes memiliki nilai outlier yang tinggi, Outlier dianalisis berdasarkan sebaran data pada setiap atribut. Nilai ekstrem tidak dihapus karena dapat merepresentasikan kondisi medis tertentu, melainkan ditangani melalui normalisasi untuk mengurangi pengaruhnya terhadap model klasifikasi.

3. Bias (Distribusi Data)

   Bias ini ditandai dengan kemunculan nilai 0 pada atribut-atribut tersebut. Secara medis, nilai 0 pada atribut tersebut tidak mungkin terjadi dan menunjukkan bahwa data tersebut tidak tercatat atau tidak dilakukan pengukuran. Untuk mengurangi bias data maka akan diterapkan penanganan yaitu :

   a. Identifikasi nilai tidak valid

      Nilai 0 pada atribut medis diidentifikasi sebagai sumber utama biasa

   b. Transformasi dan normalisasi data

      Seluruh atribut numerik dinormalisasi sebagai metode min-max agar nilai berada pada rentang 0-1 agar dapat menyeimbangkan distribusi data

   c. Tidak menghapus data

      Data tidak dihapus untuk menjaga sampel tetap utuh dan menghindari kehilangan informasi penting

4. Tranformasi Data

Pada tahap ini dilakukan mengubah data mentah menjadi data desimal untuk mempermudah dan meningkatkan kualitas data untuk analisis yang lebih akurat. Hasil data transformasi data dapat dilihat pada tabel 4.6 dibawah ini

Tabel 4. 6. Data Sebelum Normalisasi

No	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil
1	6	148	72	35	0	33,6	0,627	50	1
2	1	85	66	29	0	26,6	0,351	31	0
3	8	183	64	0	0	23,3	0,672	32	1
4	0	89	66	23	94	28,1	0,167	21	0
5	0	137	40	35	168	43,1	2,268	33	1
..	.....	....	...	....	....	...	.....	...	...
764	10	101	76	48	180	32,9	0,171	63	0
765	2	122	70	27	0	36,8	0,340	27	0
766	5	121	72	23	112	26,2	0,245	30	0
767	1	126	60	0	0	30,1	0,349	47	1
768	1	93	70	31	0	30,4	0,315	23	0

Pada tabel 4.5 adalah data sebelum dilakukan normalisasi dan tabel 4.6 merupakan data setelah dilakukan normalisasi data, tahap ini dilakukan dengan mengubah angka numerik ke angka desimal. Pada data sebelum normalisasi, data masih dalam bentuk asli dari dataset diabetes. Nilai tersebut menyebabkan bias dan distribusi tidak normal yang bisa menurunkan performa klasifikasi. Terdapat

beberapa fitur medis memiliki nilai 0 yang tidak logis misalnya glukosa, tekanan darah, insulin, dan BMI, data sebelum normalisasi belum seragam untuk algoritma sensitif seperti *naïve bayes* dan *random forest* tanpa normalisasi sehingga perlu dilakukan proses normalisasi data agar kualitas data merata untuk menghasilkan analisis data yang akurat. Berikut adalah rumus dari normalisasi yang dilakukan pada tabel 4.6

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Keterangan :

X_{max} : data tertinggi dalam suatu sampel

X_{min} : data terendah dalam suatu sampel

Contoh : $\frac{X - X_{min}}{X_{max} - X_{min}} = \frac{6 - 0}{17 - 0} = \frac{6}{17} = 0.352$

Tabel 4. 7 Data Sesudah Normalisasi

No	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil
1	0,352	0,743	0,590	0,353	0	0,500	0,234	0,483	1
2	0,058	0,427	0,540	0,292	0	0,396	0,116	0,116	0
3	0,470	0,919	0,524	0	0	0,347	0,253	0,183	1
4	0,058	0,447	0,540	0,232	0,111	0,418	0,038	0	0
5	0	0,688	0,327	0,353	0,198	0,642	0,943	0,200	1
..	.....	....	...	....	....	...	.....	...	...
764	0,588	0,507	0,622	0,484	0,212	0,490	0,039	0,700	0
765	0,117	0,613	0,537	0,272	0	0,548	0,111	0,100	0
766	0,294	0,608	0,590	0,232	0,132	0,390	0,071	0,150	0

767	0,058	0,633	0,491	0	0	0,448	0,115	0,433	1
768	0,058	0,467	0,573	0,313	0	0,453	0,101	0,033	0

Pada tabel 4.6 merupakan data setelah dilakukan normalisasi data, pada tabel tersebut data semua fitur diubah ke dalam rentang nilai 0–1 (desimal) sehingga distribusi data menjadi lebih merata dan seimbang antar fitur dan mengurangi pengaruh dominasi variabel dengan rentang nilai besar (glukosa dan umur)

Tabel 4. 8 Perbandingan Sebelum dan Setelah Normalisasi

No	Aspek	Sebelum Normalisasi	Setelah Normalisasi
1	Rentang Nilai	Beragam	Seragam (0-1)
2	Nilai Tidak Logis	Masih ada pada fitur medis	Diubah ke skala proporsional
3	Outlier	Terlihat jelas	Teredam skala proporsional
4	Skala Fitur	Tidak seimbang antar kolom	Sudah seimbang antar kolom
5	Kesiapan Data Model	Kurang baik, perlu preprocessing	Siap digunakan untuk pelatihan model
6	Dampak Akurasi Model	Bisa menurunkan akurasi karena bias nilai	Meningkatkan stabilitas dan akurasi prediksi

5. *Imbalance* Data

Imbalance data merupakan kondisi ketika jumlah data pada masing-masing kelas tidak seimbang, sehingga satu kelas memiliki jumlah data yang jauh lebih banyak dibandingkan kelas lainnya. Kondisi ini

dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas dan kurang mampu mengenali kelas minoritas secara akurat.

Kolom Hasil

- 0 (Tidak Diabetes) : 65,1%
- 1 (Diabetes) : 34,9%

Distribusi tersebut menunjukkan bahwa dataset tidak seimbang,dengan kelas negatif sebagai kelas mayoritas dan kelas positif sebagai kelas minoritas. Perbandingan jumlah data antar kelas adalah sekitar 65% : 35%, sehingga dapat dikategorikan sebagai moderate *class imbalance* . Dalam penelitian ini, imbalance data tidak ditangani menggunakan teknik resampling seperti *SMOTE* atau *undersampling*. Dataset digunakan dalam kondisi asli dengan pertimbangan:

1. Menjaga keaslian distribusi data medis,
2. Menghindari potensi *overfitting* akibat penambahan data sintetis
3. Fokus penelitian adalah pada perbandingan performa algoritma *Naïve Bayes* dan *Random Forest*, bukan pada optimasi dataset.

d. Split Data

Pada penelitian ini akan dilakukan proses split data menjadi 2 bagian yaitu data training dan data testing. Variasi pembagian data akan dilakukan menjadi beberapa bagian antara lain dengan rasio 70% data training dan 30% data testing, 80% data training dan 20% data

testing, 90% data training dan 10% data testing. Variasi pembagian data ini merupakan upaya eksperimen untuk mengetahui teknik split data yang memiliki perfoma terbaik. Jumlah komposisi pembagian data seperti pada tabel 4.9 dibawah ini

Tabel 4. 9 Tabel Pembagian Data

No	Training	Testing	Jumlah Data Training	Jumlah Data Testing
1	70%	30%	537	231
2	80%	20%	614	154
3	90%	10%	691	77

Pada tabel 4.7 merupakan tabel pembagian data training dan data testing yang digunakan, dalam penelitian ini digunakan perbandingan 70% untuk data training dan 30% untuk data testing dan 80% untuk data training dan 20% untuk data testing, serta 90% untuk data training dan 10% untuk data testing

4.1.3 Pemodelan Klasifikasi

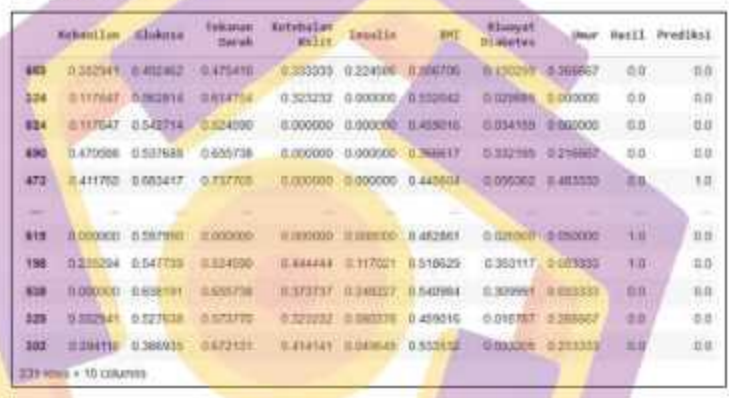
Hasil prediksi ini adalah hasil yang didapat dari split data menggunakan *Naïve Bayes Classifier* dan *Random Forest*. Pada Tabel 4.7 menampilkan hasil split data menggunakan *Naïve Bayes Classifier* dan pada Tabel 4.8 menampilkan split data menggunakan *Random Forest*. Dari hasil prediksi kita dapat memperoleh nilai akurasi, F1, dan presisi seperti yang terlihat di Tabel 4.7 dan 4.8. Nilai akurasi diperoleh dari persentase jumlah prediksi benar (Diabetes dan Tidak diabetes) dibanding dengan jumlah data test secara keseluruhan. Untuk F1 didapatkan dari nilai rata rata antara nilai dari presisi dan nilai recall dimana nilai *recall* merupakan perbandingan antara jumlah diprediksikan bernilai positif dengan banyak data

positif yang memang benar positif. Terakhir, untuk nilai presisi diperoleh dari jumlah prediksi benar positif dibandingkan dengan jumlah hasil yang diprediksi positif.

a. *Naïve Bayes*

Pada tahap ini akan dibuatkan hasil analisis data testing menggunakan *naïve bayes* sesuai dengan hasil perbandingan split data pada tabel 4.7

1. Perbandingan 70:30



	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
603	0.352941	0.452462	0.475410	0.333333	0.224586	0.358706	0.130299	0.366667	0.0	0.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.532042	0.029889	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.455016	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.396617	0.332195	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.445604	0.095362	0.483333	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
619	0.000000	0.587940	0.000000	0.000000	0.000000	0.482861	0.020069	0.050000	1.0	0.0
198	0.235294	0.547739	0.524590	0.444444	0.117021	0.516625	0.353117	0.033333	1.0	0.0
638	0.000000	0.638191	0.655738	0.373737	0.248227	0.540984	0.309991	0.033333	0.0	0.0
329	0.352941	0.527638	0.573770	0.323232	0.000378	0.459016	0.016787	0.366667	0.0	0.0
302	0.294118	0.386935	0.672131	0.414141	0.040649	0.533532	0.000000	0.233333	0.0	0.0

239 rows × 10 columns

Gambar 4. 2 Hasil Prediksi Data Testing Naive Bayes 70:30

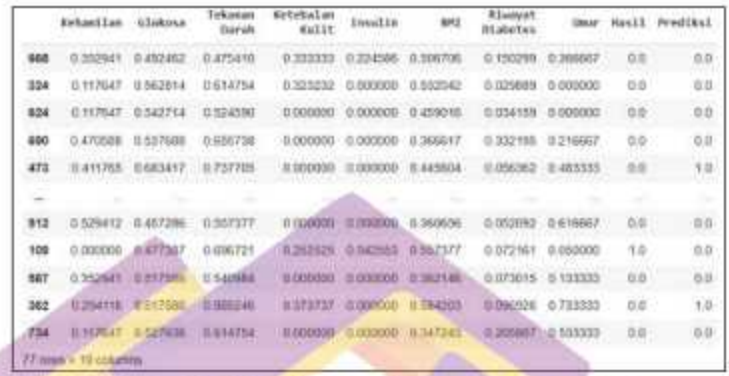
2. Perbandingan 80:20

	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
603	0.352941	0.452462	0.475410	0.333333	0.224586	0.358706	0.130299	0.366667	0.0	0.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.532042	0.029889	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.455016	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.396617	0.332195	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.445604	0.095362	0.483333	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
355	0.529412	0.829146	0.721311	0.000000	0.000000	0.453055	0.095648	0.466667	1.0	1.0
534	0.058824	0.386935	0.459016	0.303030	0.068194	0.496274	0.503854	0.050000	0.0	0.0
344	0.470588	0.477387	0.590164	0.000000	0.000000	0.546405	0.173783	0.600000	0.0	0.0
296	0.117647	0.733668	0.573770	0.383838	0.425532	0.417288	0.110589	0.133333	1.0	1.0
492	0.470588	0.371859	0.573770	0.404040	0.057920	0.526680	0.267720	0.300000	0.0	0.0

154 rows × 10 columns

Gambar 4. 3 Hasil Prediksi Data Testing *Naïve Bayes* 80:20

3. Perbandingan 90:10



	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
668	0.352941	0.492462	0.475410	0.333333	0.224586	0.356706	0.150299	0.366667	0.0	0.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.552342	0.025689	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.459016	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.366617	0.332155	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.445604	0.056362	0.483333	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
512	0.529412	0.467286	0.557377	0.000000	0.000000	0.360656	0.052092	0.616667	0.0	0.0
109	0.000000	0.477387	0.606721	0.252525	0.042553	0.367377	0.072161	0.050000	1.0	0.0
587	0.352941	0.517588	0.540984	0.000000	0.000000	0.382148	0.073015	0.133333	0.0	0.0
362	0.294118	0.512588	0.655246	0.373737	0.000000	0.584203	0.096926	0.733333	0.0	1.0
734	0.117647	0.527638	0.614754	0.000000	0.000000	0.347243	0.205857	0.533333	0.0	0.0

77 rows × 10 columns

Gambar 4. 4 Hasil Prediksi Data Testing *Naïve Bayes* 90:10

Pada gambar 4.3, 4.4 dan 4.5 merupakan hasil perbandingan data fakta dan data prediksi menggunakan *naïve bayes* berdasarkan pembagian data training dan data testing adapun perbandingan data yang digunakan yaitu 70:30, 80:20 dan 90:10

Tabel 4. 10 Hasil *Naive Bayes Classifier*

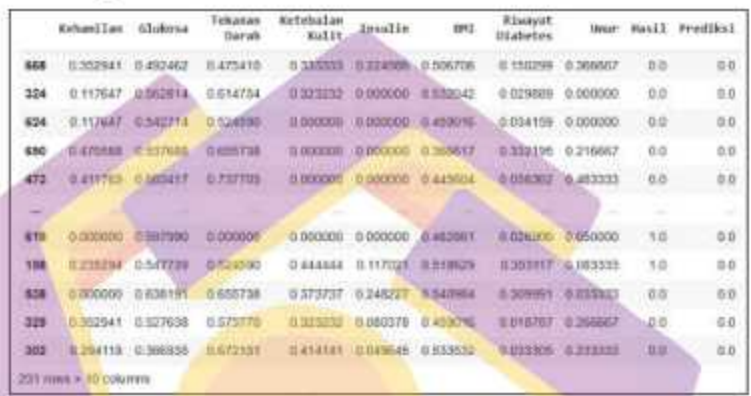
No	Split Data	*Naïve Bayes Classifier*		
		Akurasi	F1	Presisi
1	70% : 30%	74%	64%	62%
2	80% : 20%	77%	68%	66%
3	90% : 10%	70%	62%	56%

Pada tabel 4.8 merupakan hasil analisa dari klasifikasi *naïve bayes* dengan menampilkan akurasi, *f1-score* dan presisi dari masing-masing split data yang digunakan. Dari tabel tersebut didapatkan persentase perbandingan 80%:20% memiliki hasil tertinggi dengan hasil akurasi 77%, f1 score 68% dan presisi 66% .

b. *Random Forest*

Pada tahap ini akan dibuatkan hasil prediksi data testing menggunakan *random forest* sesuai dengan hasil perbandingan split data pada tabel 4.9

1. Perbandingan 70:30

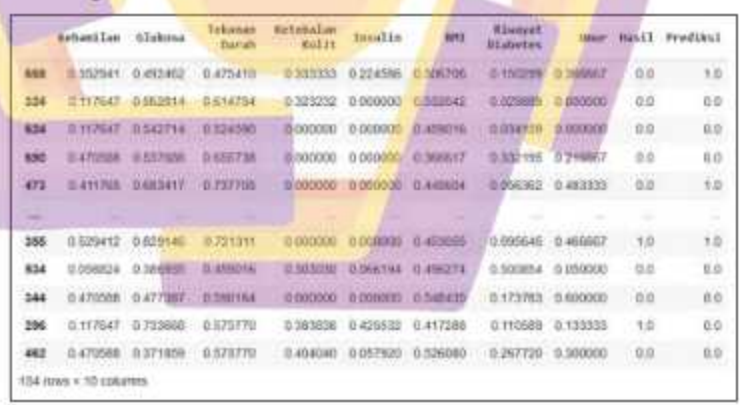


	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
668	0.352941	0.492462	0.475410	0.333333	0.224586	0.506706	0.150299	0.366667	0.0	0.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.532042	0.029889	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.459016	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.366617	0.332195	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.445604	0.056362	0.483333	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
619	0.000000	0.597990	0.000000	0.000000	0.000000	0.462001	0.024765	0.650000	1.0	0.0
198	0.235294	0.547739	0.524590	0.444444	0.117021	0.518629	0.353117	0.183333	1.0	0.0
538	0.000000	0.638191	0.655738	0.373737	0.248227	0.540984	0.309991	0.033333	0.0	0.0
329	0.352941	0.527638	0.573770	0.323232	0.080378	0.493876	0.016787	0.266667	0.0	0.0
302	0.294118	0.366834	0.672131	0.414141	0.049645	0.533532	0.033305	0.233333	0.0	0.0

231 rows × 10 columns

Gambar 4. 5 Hasil Prediksi Data Testing *Random Forest* 70:30

2. Perbandingan 80:20



	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
668	0.352941	0.492462	0.475410	0.333333	0.224586	0.506706	0.150299	0.366667	0.0	1.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.532042	0.029889	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.459016	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.366617	0.332195	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.445604	0.056362	0.483333	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
355	0.529412	0.629146	0.721311	0.000000	0.000000	0.453055	0.055646	0.466667	1.0	1.0
634	0.058824	0.366834	0.459016	0.505050	0.064194	0.496274	0.500854	0.050000	0.0	0.0
344	0.470588	0.477387	0.590164	0.000000	0.000000	0.548435	0.173783	0.600000	0.0	0.0
296	0.117647	0.733668	0.573770	0.383838	0.429532	0.417288	0.110589	0.133333	1.0	0.0
462	0.470588	0.371859	0.573770	0.404040	0.057920	0.526080	0.297720	0.500000	0.0	0.0

154 rows × 10 columns

Gambar 4. 6 Hasil Prediksi Data Testing *Random Forest* 80:20

3. Perbandingan 90:10

	Kehamilan	Glukosa	Tekanan Darah	Ketebalan Kulit	Insulin	BMI	Riwayat Diabetes	Umur	Hasil	Prediksi
668	0.352941	0.493462	0.475410	0.323333	0.234586	0.395706	0.150299	0.366667	0.0	1.0
324	0.117647	0.562814	0.614754	0.323232	0.000000	0.532042	0.029889	0.000000	0.0	0.0
624	0.117647	0.542714	0.524590	0.000000	0.000000	0.450075	0.034159	0.000000	0.0	0.0
690	0.470588	0.537688	0.655738	0.000000	0.000000	0.366617	0.332195	0.216667	0.0	0.0
473	0.411765	0.683417	0.737705	0.000000	0.000000	0.449664	0.056362	0.483333	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...
512	0.529412	0.457286	0.557377	0.000000	0.000000	0.360656	0.052092	0.616667	0.0	0.0
109	0.000000	0.477387	0.606721	0.252525	0.042553	0.559177	0.072161	0.050000	1.0	0.0
587	0.352941	0.517588	0.540984	0.000000	0.000000	0.362146	0.073013	0.133333	0.0	0.0
362	0.294118	0.517588	0.885246	0.373737	0.000000	0.554604	0.096526	0.733333	0.0	1.0
734	0.117647	0.522636	0.614754	0.000000	0.000000	0.347243	0.205807	0.233333	0.0	0.0

77 rows × 10 columns

Gambar 4. 7 Hasil Prediksi Data Testing *Random Forest* 90:10

Pada gambar 4.6, 4.7 dan 4.8 merupakan hasil perbandingan data fakta dan data prediksi menggunakan *random forest* berdasarkan pembagian data training dan data testing adapun perbandingan data yang digunakan yaitu 70:30, 80:20 dan 90:10

Tabel 4. 11 Hasil *Random Forest*

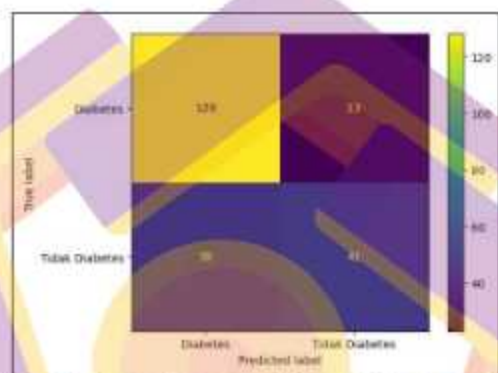
No	Split Data	*Random Forest*		
		Akurasi	F1	Presisi
1	70% : 30%	73%	57%	64%
2	80% : 20%	73%	62%	63%
3	90% : 10%	70%	58%	57%

Pada tabel 4.10 merupakan hasil dari klasifikasi *random forest* dengan menampilkan akurasi, f1-score dan presisi dari masing-masing split data yang digunakan. Dari tabel tersebut didapatkan persentase perbandingan 80%:20% memiliki hasil tertinggi dengan hasil akurasi 73%, *f1 score* 62% dan presisi 63%

4.1.4 Evaluasi

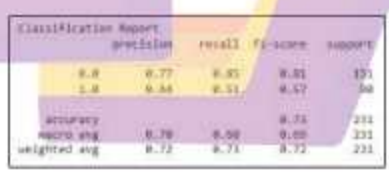
Pada tahap ini dilakukan penilaian akurasi dari kedua metode untuk mengetahui metode terbaik dalam mengidentifikasi diabetes. Adapun hasil evaluasi dari pemodelan *Naïve Bayes Classifier* dan *Random Forest*

a. *Naïve Bayes*

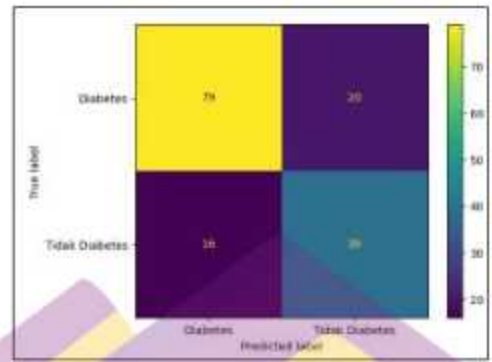


Gambar 4. 8 *Confusion Matrix Naïve Bayes* 70:30

Pada gambar 4.9 merupakan hasil dari *confusion matrix* dari *naïve bayes* dengan perbandingan data training dan data testing sebesar 70:30. Sehingga didapatkan hasil akurasi, presisi, recall dan F1-score sebagai berikut:



Gambar 4. 9 Hasil *Confusion Matrix Naïve Bayes* 70:30



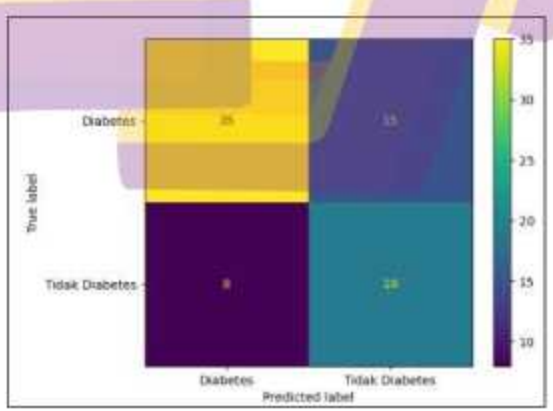
Gambar 4. 10 *Confusion Matrix Naïve Bayes* 80:20

Pada gambar 4.11 merupakan hasil dari *confusion matrix* dari *naïve bayes* dengan perbandingan data training dan data testing sebesar 80:20. Sehingga didapatkan hasil akurasi, presisi, *recall* dan F1-*score* sebagai berikut:

Classification Report

	precision	recall	f1-score	support
0.0	0.83	0.80	0.81	99
1.0	0.66	0.71	0.68	55
accuracy			0.77	154
macro avg	0.75	0.75	0.75	154
weighted avg	0.77	0.77	0.77	154

Gambar 4. 11 Hasil *Confusion Matrix Naïve Bayes* 80:20



Gambar 4. 12 *Confusion Matrix Naïve Bayes* 90:10

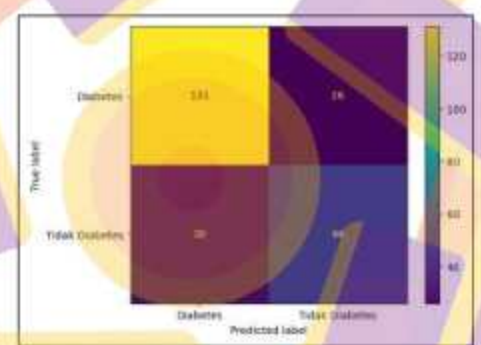
Pada gambar 4.13 merupakan hasil dari *confusion matrix* dari *naïve bayes* dengan perbandingan data training dan data testing sebesar 90:10. Sehingga didapatkan hasil akurasi, presisi, *recall* dan *F1-score* sebagai berikut:

Classification Report

	precision	recall	f1-score	support
0.0	0.81	0.70	0.75	50
1.0	0.56	0.70	0.62	27
accuracy			0.70	77
macro avg	0.69	0.70	0.69	77
weighted avg	0.72	0.70	0.71	77

Gambar 4. 13 Hasil *Confusion Matrix* 90:10

b. *Random Forest*



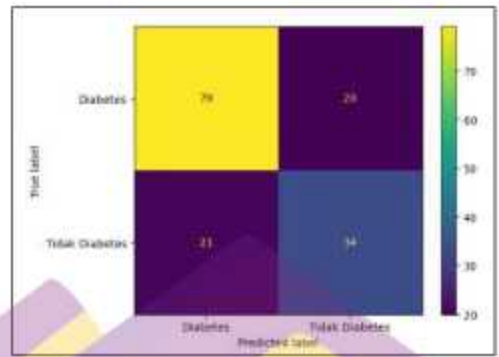
Gambar 4. 14 *Confusion Matrix Random Forest* 70:30

Pada gambar 4.15 merupakan hasil dari *confusion matrix* dari *random forest* dengan perbandingan data training dan data testing sebesar 70:30. Sehingga didapatkan hasil akurasi, presisi, *recall* dan *F1-score* sebagai berikut:

Classification Report

	precision	recall	f1-score	support
0.0	0.81	0.83	0.82	157
1.0	0.63	0.59	0.61	74
accuracy			0.76	231
macro avg	0.72	0.71	0.72	231
weighted avg	0.75	0.76	0.76	231

Gambar 4. 15 Hasil *Confusion Matrix Random Forest* 70:30



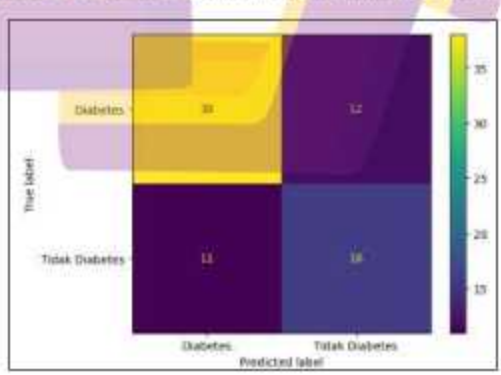
Gambar 4. 16 *Confusion Matrix Random Forest* 80:20

Pada gambar 4.17 merupakan hasil dari *confusion matrix* dari *random forest* dengan perbandingan data training dan data testing sebesar 80:20. Sehingga didapatkan hasil akurasi, presisi, *recall* dan F1-*score* sebagai berikut:

Classification Report

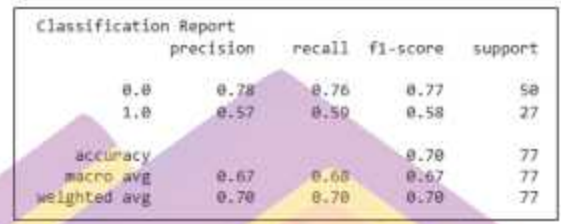
	precision	recall	f1-score	support
0.0	0.79	0.80	0.79	99
1.0	0.63	0.62	0.62	55
accuracy			0.73	154
macro avg	0.71	0.71	0.71	154
weighted avg	0.73	0.73	0.73	154

Gambar 4. 17. Hasil *Confusion Matrix Random Forest* 80:20



Gambar 4. 18 *Confusion Matrix Random Forest* 90:10

Pada gambar 4.19 merupakan hasil dari *confusion matrix* dari *random forest* dengan perbandingan data training dan data testing sebesar 90:10. Sehingga didapatkan hasil akurasi, presisi, *recall* dan F1-*score* sebagai berikut:



Classification Report

	precision	recall	f1-score	support
0.0	0.78	0.76	0.77	50
1.0	0.57	0.59	0.58	27
accuracy			0.70	77
macro avg	0.67	0.68	0.67	77
weighted avg	0.70	0.70	0.70	77

Gambar 4. 19 Hasil *Confusion Matrix Random Forest* 90:10

4.1.5 Analisis Hasil

Langkah selanjutnya dengan menentukan metode yang terbaik antara *naïve bayes* dan *random forest* dalam mengklasifikasi penyakit diabetes dengan melihat rata-rata akurasi dari masing metode sebagai berikut :

Tabel 4. 12 Analisis Hasil Metode

No	Split Data	Akurasi		Presisi		F1 Score	
		Naïve Bayes	Random Forest	Naïve Bayes	Random Forest	Naïve Bayes	Random Forest
1	70 % : 30 %	74%	73%	62%	64%	64%	57%
2	80% : 20%	77%	73%	66%	63%	68%	62%
3	90% : 10%	70%	70%	56%	57%	62%	58%
Rata-Rata		73.66%	72%	61.33%	61.33%	64,66%	59%

Pada tabel 4.11 merupakan analisis hasil metode yang membandingkan kedua metode antara *Naïve Bayes* dan *Random Forest*, pada hasil tersebut didapatkan nilai rata-rata akurasi, presisi, f1 *score Naïve Bayes* lebih tinggi dibandingkan dengan *Random Forest. Naïve Bayes* memiliki rata-rata nilai akurasi sebesar 75%

lebih tinggi dibandingkan dengan nilai rata-rata akurasi *Random Forest* yaitu 74.66%, *Naïve Bayes* juga memiliki nilai presisi sebesar 72,16% lebih tinggi dibandingkan nilai presisi *Random Forest* yaitu 71% dan namun *Naïve Bayes* memiliki nilai f1 *score* 70,83% lebih rendah dibandingkan nilai f1 *score Random Forest* yaitu 71,16% sehingga dapat disimpulkan bahwa metode *naïve bayes* merupakan metode terbaik dalam melakukan analisa klasifikasi penyakit diabetes. Hal tersebut karena jumlah data yang digunakan tidak terlalu banyak sehingga metode *naive bayes* memiliki kinerja yang sangat baik dalam hal akurasi dibanding *random forest*

4.1.6 Analisis Perbandingan

Pada tahap ini penulis akan menganalisis perbandingan penelitian penulis dengan penelitian terkait yang sesuai dengan penelitian perbandingan *naïve bayes* dan *random forest* dalam klasifikasi penyakit diabetes. Adapun penelitian yang akan dibandingkan adalah 5 paper penelitian sebelumnya. Adapun perbedaan antara penelitan antara penelititan penulis dan penelitian sebelumnya adalah sebagai berikut :

Tabel 4. 13 Analisis Hasil Perbandingan

No	Penelitian	Dataset	Normalisasi	Split Data	Akurasi	
					Random Forest	Naïve Bayes
1	(Amri et al., 2025)	Rekam medis pasien diabetes di	metode Min-Max Scaling atau StandardScaler	80:20	89,97%	85,97%,

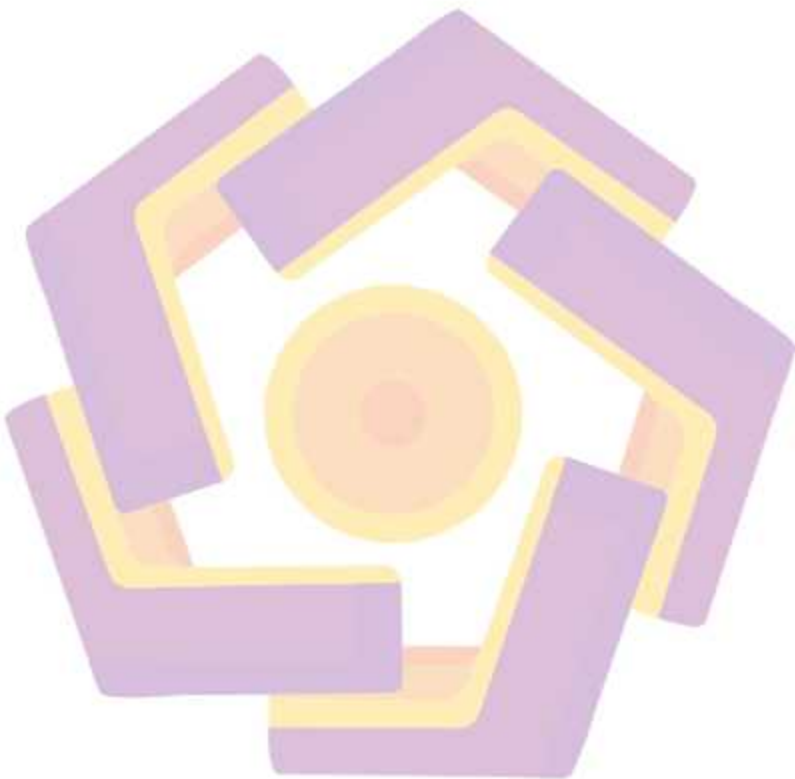
No	Penelitian	Dataset	Normalisasi	Split Data	Akurasi	
					Random Forest	Naïve Bayes
		Klinik Citra Sejati				
2	(Kholish et al., 2024)	Kaagle mengenai data penyakit diabetes	Tidak melakukan proses normalisasi	80:20	72%	77%
3	(Wijaya et al., 2025)	Kaagle mengenai penyakit diabetes	Metode Min-Max Scaling	80:20	85%	92%
4	(Lestari & Homaidi, 2024)	Dataset Breast Cancer Wisconsin	Metode Min-Max Scaling	90:10 80:20 70:30	94,91%	93.51%
5	(Alfian Fadillah & Dewi Sri Mulyani, 2024)	Komunitas global pada studi kasus Electronic Health Record (EHRs)	Tidak menggunakan proses normalisasi	90 : 10, 80 : 20, 70 : 30	97.33%	95.57%
	Penelitian penulis	**Kaagle mengenai penyakit diabetes**	**Metode Min-Max Scaling**	**90:10** **80:20** **70:30** **60:40** **50:50**	**72%**	**73.66%**

Berdasarkan analisis perbandingan antara penelitian penulis dan 5 penelitian terkait, kebanyakan metode *random forest* yang memiliki akurasi yang lebih baik dibandingkan *naïve bayes* hal tersebut dikarenakan adanya perbedaan dari masing-masing metode dalam menangani data yang kompleks, *random forest* tahan terhadap *outlier* dan *noise* dari dataset yang digunakan sedangkan *naïve bayes* sangat sensitif terhadap adanya *noise* atau data *outlier* sehingga distribusi data tidak normal yang mempengaruhi hasil akurasi, berbeda dengan penelitian penulis yang menghasilkan metode *naïve bayes* lebih baik dari metode *random forest* dari segi akurasi data, hal tersebut dikarenakan ukuran dataset ini yang digunakan pada penelitian tidak terlalu besar sehingga tidak membutuhkan banyak waktu dalam proses pelatihan sedangkan *random forest* membutuhkan data yang cukup besar agar dapat berjalan optimal.

Berdasarkan analisis keterkaitan fitur, terdapat beberapa atribut yang saling berkorelasi, khususnya glukosa dengan insulin serta BMI dengan ketebalan lipatan kulit. Selain itu, fitur glukosa merupakan atribut yang paling dominan dalam menentukan hasil prediksi diabetes. Dominasi fitur glukosa menyebabkan algoritma Naïve Bayes mampu memberikan performa yang kompetitif meskipun memiliki asumsi independensi antar fitur, sementara Random Forest belum sepenuhnya optimal akibat keterbatasan ukuran data dan kualitas fitur.

Rendahnya kinerja algoritma dalam penelitian ini tidak semata-mata disebabkan oleh kelemahan metode klasifikasi yang digunakan, melainkan lebih dipengaruhi oleh keterbatasan kualitas dan karakteristik dataset, seperti adanya nilai medis yang tidak logis, outlier ekstrem, distribusi kelas yang tidak seimbang, serta keterbatasan

fitur yang digunakan. Selain itu, kompleksitas penyakit diabetes yang bersifat multifaktorial menyebabkan pola antar kelas menjadi sulit dipisahkan secara jelas, sehingga berdampak langsung pada performa model.



BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan maka kesimpulan pada penelitian ini yaitu :

1. Berdasarkan hasil evaluasi kinerja model klasifikasi menggunakan algoritma *Naïve Bayes* dan *Random Forest*, dapat disimpulkan bahwa kedua algoritma mampu melakukan klasifikasi penyakit diabetes dengan tingkat performa yang relatif sebanding, namun menunjukkan keunggulan pada metrik evaluasi yang berbeda.Dari sisi akurasi, algoritma *Naïve Bayes* memperoleh nilai yang sedikit lebih tinggi dibandingkan *Random Forest*. Hal ini menunjukkan bahwa *Naïve Bayes* lebih konsisten dalam melakukan prediksi secara keseluruhan terhadap data uji, khususnya pada dataset dengan distribusi kelas yang tidak seimbang dan jumlah data yang terbatas.Pada metrik presisi, *Naïve Bayes* juga menunjukkan performa yang lebih baik. Nilai presisi yang lebih tinggi mengindikasikan bahwa *Naïve Bayes* lebih mampu meminimalkan kesalahan prediksi positif (*false positive*), sehingga lebih andal dalam memastikan bahwa pasien yang diprediksi menderita diabetes memang benar termasuk dalam kelas tersebut. Sementara itu, pada metrik F1-score, algoritma *Random Forest* menunjukkan hasil yang sedikit lebih unggul dibandingkan *Naïve Bayes*. Hal ini menandakan bahwa *Random Forest* memiliki keseimbangan yang lebih

baik antara presisi dan recall, khususnya dalam menangani variasi data dan distribusi kelas minoritas.

2. Berdasarkan hasil penelitian, *Naïve Bayes* dapat dinyatakan sebagai algoritma dengan kinerja paling stabil dalam penelitian ini karena unggul pada metrik akurasi dan presisi. Namun, *Random Forest* lebih baik dalam menjaga keseimbangan performa antara presisi dan recall sebagaimana tercermin pada nilai F1-score. Perbedaan kinerja kedua algoritma dipengaruhi oleh karakteristik dataset, keberadaan *imbalance* data, serta nilai ekstrem pada beberapa atribut medis. Dengan demikian, pemilihan algoritma terbaik bergantung pada tujuan penggunaan model. Jika fokus utama adalah ketepatan prediksi secara umum, maka *Naïve Bayes* lebih direkomendasikan. Namun, jika fokus diarahkan pada keseimbangan deteksi kasus diabetes, maka *Random Forest* menjadi alternatif yang lebih sesuai.

5.2. Saran

Saran atau usulan yang dapat diusulkan pada penelitian ini yaitu pada penelitian selanjutnya dapat membandingkan metode *random forest* dengan metode yang lain seperti KNN atau metode *Linear Regression* (LR) dan metode *Support Vector Machine* (SVM) untuk membandingkan metode yang lebih baik dari *Random Forest*. Usulan yang lainnya dapat menggunakan studi kasus yang lain dengan jumlah data yang berbeda untuk melihat tingkat akurasinya

DAFTAR PUSTAKA

Abdi, A. (2018). Three types of Machine Learning Algorithms. *Presentation*, 1–27. https://doi.org/10.13140/RG.2.2.26209.10088

Adnan, M., Mulyati, T., Isworo, J. T., Studi, P., Fakultas, G., Keperawatan, I., & Kesehatan, D. (2013). Hubungan Indeks Massa Tubuh (IMT) Dengan Kadar Gula Darah Penderita Diabetes Mellitus (DM) Tipe 2 Rawat Jalan Di RS Tugurejo Semarang. *JURNAL GIZI UNIVERSITAS MUHAMMADIYAH SEMARANG*, *2*(1), 18–24. http://jurnal.unimus.ac.id

Afif, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus di Rumah Sakit Aisyiah. *Jurnal Ilmu Komputer Dan Matematika* , *1*(2).

Alfian Fadillah, M., & Dewi Sri Mulyani, E. (2024). Komparasi Algoritma C4.5, Naive Bayes, K-Nearest Neighbor, Random Forest Untuk Prediksi Faktor Penyebab Penyakit Diabetes A. *Indonesian Journal of Digital Business*, *4*(1), 37–46. https://ejournal.upi.edu/index.php/IJDB

Alita, D., & Rahman, A. (2020). Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier. In *Jurnal Komputasi* (Vol. 8, Issue 2).

Amari, R. O. (2023). *ANALISIS KINERJA BERBAGAI METODE KLASIFIKASI UNTUK DIAGNOSIS PENYAKIT GINJAL KRONIS*. 31–41.

Amri, M. R. A., Permana, E., Pachadria, P. A., & Fitri, S. (2025). Perbandingan Metode Naïve Bayes dan Random Forest dalam Memprediksi Penyakit

Diabetes Melitus pada Klinik Citra Sejati. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *7*(4), 847–858. https://doi.org/10.35746/jtim.v7i4.747

Anisya, S., Prayudha, J., & Murniyanti, S. (2020). Implementasi Metode Random Forest Pada Sistem Persediaan Bahan Kimia Di Laboraturium Forensik Cabang Medan. In *Jurnal CyberTech: Vol. x. No.x.* https://ojs.trigunadharma.ac.id/

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi*, *10*(1), 163. https://doi.org/10.32520/stmsi.v10i1.1129

Argina, A. M. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. *Indonesian Journal of Data and Science*, *1*(2), 29–33. https://doi.org/10.33096/ijodas.v1i2.11

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., Dwi, A., Putra, P., & Artikel, G. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(2), 2828–9099. https://doi.org/10.55123/jomlai.v1i2.598

Diana, R., Warni, H., & Sutabri, T. (2023). PENGGUNAAN TEKNOLOGI MACHINE LEARNING UNTUK PELAYANAN MONITORING KEGIATAN BELAJAR MENGAJAR PADA SMK BINA SRIWIJAYA PALEMBANG. *JUTEKIN (Jurnal Teknik Informatika)*, *11*(1). https://doi.org/10.51530/jutekin.v11i1.709

Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. *Indonesian Journal of Statistics and Its Applications*, *4*(1), 80–88. https://doi.org/10.29244/ijsa.v4i1.330

Fangatulo Dodo Telaumbanua, D., Hulu, P., Zulfiter Nadeak, T., Romeo Lumbantong, R., & Dharma, A. (2019). Penggunaan Machine Learning. *Jurnal Penelitian Teknik Informatika*, *2*(2), 391–399.

Fathurahman, H., Ariwikri, A., Pratama, G. A., FIKRI, M. A. F. S., & ALRIZKI, M. F. (2023). Perbandingan Akurasi Metode Naive Bayes Classifier Dan Random Forest Menggunakan Reduksi Dimensi Linear Dicriminant Analysis (Lda) Untuk Diagnosis Penyakit Diabetes. *Jurnal Rekayasa Elektro Sriwijaya*, *4*(1), 24–31. https://doi.org/10.36706/jres.v4i1.58

Hidayat, H., Sunyoto, A., & Al Fatta, H. (2023). Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *7*(1), 31–40. https://doi.org/10.47970/siskom-kb.v7i1.464

Informatika, J. T., Sains, F., & Teknologi, D. A. N. (2008). *SISTEM PAKAR DETEKSI PENYAKIT DIABETES MELLITUS DENGAN MENGGUNAKAN PENDEKATAN NAÏVE BAYESIAN BERBASIS WEB*.

Khasanah, L. U., Nasution, Y. N., Deny, F., & Amijaya, T. (2022). Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Ilmiah Matematika*, *1*(1), 41–50. http://jurnal.fmipa.unmul.ac.id/index.php/basis

Kholish, M., Herdianto, A., Setiawan, R. F., Samsinar, R., & Elektro, T. (2024). Perbandingan Algoritma Random Forest dan Naive Bayes dalam Memprediksi Penyakit Diabetes. *SEMINAR NASIONAL & CALL FOR PAPER*, 322–328.

Lestari, I. I., & Homaidi, A. (2024). Komparasi Algoritma Naive Bayes Dan Random Forest PadaKlasifikasi Kanker Payudara. *Gudang Jurnal Multidisiplin Ilmu*, *2*(12), 778–785. https://doi.org/10.59435/gjmi.v2i12.1206

Magaña, P., Del-Rosal-Salido, J., Cobos, M., Lira-Loarca, A., & Ortega-Sánchez, M. (2020). Approaching software engineering for marine sciences: A single development process for multiple end-user applications. *Journal of Marine Science and Engineering*, *8*(5). https://doi.org/10.3390/JMSE8050350

Mashudi, N. A., Ahmad, N., & Mohd Noor, N. (2022). LiWGAN: A Light Method to Improve the Performance of Generative Adversarial Network. *IEEE Access*, *10*(September), 93155–93167. https://doi.org/10.1109/ACCESS.2022.3203065

Maulidah, N., Supriyadi, R., Utami, D. Y., Hasan, F. N., Fauzi, A., & Christian, A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support

Vector Machine dan Naive Bayes. *Indonesian Journal on Software Engineering (IJSE)*, *7*(1), 63–68. http://ejournal.bsi.ac.id/ejurnal/index.php/ijse63

Nainggolan, S. P., & Sinaga, A. (2023). Comparative Analysis of Accuracy of Random Forest and Gradient Boosting Classifier Algorithm for Diabetes Classification. *Sebatik*, *27*(1), 97–102. https://doi.org/10.46984/sebatik.v27i1.2157

Nam Han Cho, Joses Kirigia, Jean Claude Mbanya, Katherine Ogurstova, & Leonor Guariguata. (2017). IDF DIABETES ATLAS. *International Diabetes Federation*, *7*(1), 1–150.

Nicholas, G. (2024). Machine Learning Algorithms for Cloud Computing Security. *Research Proposal*, 1–15. https://doi.org/10.13140/RG.2.2.18676.12168

Nurussakinah, N., & Faisal, M. (2023). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree. *Jurnal Informatika*, *10*(2), 143–149. https://doi.org/10.31294/inf.v10i2.15989

Patimah, N. F., Abdurrohman, M., Rinaldi, A. R., & Rinaldi Dikananda, A. (2021). Implementasi Algoritma Naïve Bayes dalam Klasifikati Penyakit Diabetes. *Jurnal Data Science & Informatika*, *1*(1), 6–10.

Prandika Siregar, A., Priyadi Purba, D., Putri Pasaribu, J., Reza Bakara, K., & Willem Iskandar Pasar Medan Estate, J. V. (2023). Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)*, *2*(4), 155–164.

Punthakee, Z., Goldenberg, R., & Katz, P. (2018). Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome. *Canadian Journal of Diabetes*, *42*, S10–S15. https://doi.org/10.1016/j.jcjd.2017.10.003

Putry, N. M. (2022). Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus. *EVOLUSI : Jurnal Sains Dan Manajemen*, *10*(1). https://doi.org/10.31294/evolusi.v10i1.12514

Ratna Patil, S. T. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. *International Journal of Electrical and Computer Engineering*, *8*(5), 3966–3975. https://doi.org/10.11591/ijece.v8i5.pp3966-3975

Roihan, A., Abas Sunarya, P., & Rafika, A. S. (2019). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *5*(1), 75–82.

Rosita, R., Ananda Agustina Pertiwi, D., & Gina Khoirunnisa, O. (2022). Prediction of Hospital Intesive Patients Using Neural Network Algorithm. *Journal of Soft Computing Exploration*, *3*(1), 8–11. https://doi.org/10.52465/joscex.v3i1.61

Sholekhah, F., Putri, A. D., & Efrizoni, L. (2024). *Comparison of Naive Bayes and K-Nearest Neighbors Algorithms for Metabolic Syndrome Classification Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom*. *4*(April), 507–514.

Sunjana. (2010). APLIKASI MINING DATA MAHASISWA DENGAN METODE KLASIFIKASI DECISION TREE. In *Seminar Nasional Aplikasi Teknologi Informasi*.

Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 67–75. https://doi.org/10.51903/e-bisnis.v13i2.247

Vadim, K. (2018). Overview of different approaches to solving problems of data mining. *Procedia Computer Science*, *123*, 234–239. https://doi.org/10.1016/j.procs.2018.01.036

Widya Ningsih, E. (2020). Penerapan Algoritma Naïve Bayes Dalam Penentuan Kelayakan Penerima Kartu Jakarta Pintar Plus. *Jurnal Teknik Komputer AMIK BSI* , *6*(1), 15–20. https://doi.org/10.31294/jtk.v4i2

Wijaya, A. P., Penulis, *, & Diajukan, K. (2025). Perbandingan Algoritma Klasifikasi Random Foresst dengan Naïve Bayes Classifier pada Studi Penyakit Berdasarkan Pola Nutrisi. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *9*(1). https://doi.org/10.33395/remik.v9i1.14652

Yudha Prawira, D., Pratama, Y., & Yanti, E. (2024). Analisis Dan Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerimaan Beasiswa PIP (Studi Kasus : SMPN 7 Kota Jambi). *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)*, *4*(2). https://doi.org/10.33998/jakakom.v4i2

Zainal Macfud, A., Pandu Kusuma, A., Dwi Puspitasari, W., Balitar Blitar Jl Majapahit No, I., Sananwetan, K., Blitar, K., & Timur, J. (2023). ANALISIS ALGORITMA NAIVE BAYES CLASSIFIER (NBC) PADA KLASIFIKASI TINGKAT MINAT BARANG DI TOKO VIOLET CELL. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 1).

LAMPIRAN