ANALISIS KLASIFIKASI PENYAKIT HEPATITIS C MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

TESIS

Oleh:
IDA OKTAVIA SALSAVANA
NIM. 23065220012

PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS NEGRI MAULANA MALIK IBRAHIM
MALANG
2026

HALAMAN PENGAJUAN

ANALISIS KLASIFIKASI PENYAKIT HEPATITIS C MENGGUNAKAN ALGORITMA *NAIVE BAYES* DAN *RANDOM FOREST*

TESIS

Diajukan Kepada:

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

Oleh:
IDA OKTAVIA SALSAVANA
NIM. 23065220012

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

MOTTO

"Sesungguhnya beserta kesulitan ada kemudahan." Q.S. Al-Insyirah: 6 “

Tidak ada perjuangan yang sia-sia selama doa dan ikhtiar terus mengalir. Setiap hambatan dalam penyusunan tugas akhir ini pada akhirnya bermuara pada keindahan sebuah pencapaian.

HALAMAN PERSEMBAHAN

Bismillahirrahmanirrahim

Dengan mengucapkan puji syukur ke hadirat Allah Swt. atas segala rahmat, karunia, dan ridha-Nya, sehingga penulis dapat menyelesaikan tesis ini. Karya sederhana ini penulis persembahkan dengan penuh rasa syukur, hormat, dan kasih sayang kepada:

1. kedua orang tua saya tercinta, Bapak Joko Kristanto Dan Ibu Kusniyawati. Terima kasih atas segala doa yang tidak pernah putus, kasih sayang yang tulus, pengorbanan, dukungan, serta semangat yang selalu diberikan kepada penulis. Terima kasih telah menjadi alasan terbesar bagi penulis untuk terus berjuang dan menyelesaikan pendidikan hingga tahap ini. Semoga Allah Swt. senantiasa memberikan kesehatan, kebahagiaan, dan keberkahan kepada Bapak dan Ibu.
2. Adik-adik tersayang, Ainurahma Azahra, Habibi Ahmad Rahmatullah, Rangga Wardana Setiawan, dan Syahnaz Riski Setiawan.Terima kasih atas doa, dukungan, canda, dan kebersamaan yang selalu menjadi penyemangat bagi penulis dalam setiap proses perjalanan menyelesaikan tesis ini. Semoga kalian selalu diberikan kemudahan dalam meraih cita-cita dan menjadi kebanggaan keluarga.
3. Kakek dan nenek tercinta, Bapak Darga dan Ibu Rasiyah.Terima kasih atas kasih sayang, doa, perhatian, dan nasihat yang selalu mengiringi setiap langkah penulis. Semoga Allah Swt. senantiasa melimpahkan kesehatan, umur yang berkah, serta kebahagiaan kepada Kakek dan Nenek.
4. Seluruh keluarga besar.Terima kasih atas doa, dukungan, perhatian, dan motivasi yang telah diberikan kepada penulis selama menempuh pendidikan. Kehadiran dan kebersamaan keluarga menjadi kekuatan yang sangat berarti dalam menyelesaikan tesis ini.
5. Teman-teman seperjuangan.Terima kasih atas kebersamaan, bantuan, dukungan, motivasi, serta pengalaman yang telah dilalui bersama selama masa perkuliahan hingga penyusunan tesis ini.
6. Untuk diri saya sendiri.Terima kasih telah mampu bertahan, bangkit dari setiap kesulitan, dan tidak menyerah meskipun perjalanan ini penuh dengan tantangan, rasa lelah, dan keraguan. Terima kasih karena telah memilih untuk terus melangkah hingga mampu menyelesaikan tesis ini. Semoga pencapaian ini menjadi awal dari perjalanan yang lebih baik, membuka pintu kesempatan yang lebih luas.

KATA PENGANTAR

Bismillahirrahmanirrahim

Puji syukur kehadirat Allah Subḥānahu wa Ta‘ālā atas segala rahmat, taufik, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan studi pada Program Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang serta menyelesaikan tesis ini dengan baik. Shalawat serta salam senantiasa tercurah kepada Nabi Muhammad SAW yang telah membawa umat manusia menuju jalan yang penuh ilmu pengetahuan dan keberkahan.

Penulis menyadari bahwa penyusunan tesis ini tidak terlepas dari bantuan, bimbingan, motivasi, doa, dan dukungan dari berbagai pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih kepada:

1. Allah Subḥānahu wa Ta‘ālā atas segala rahmat, karunia, kesehatan, kekuatan, dan kemudahan yang diberikan kepada penulis.
2. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Prof. Dr. H. Agus Mulyono, S.Pd., M.Kes., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Ir. Muhammad Faisal, S.Kom., M.T., selaku Ketua Program Studi Magister Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
5. Dr. Mokhamad Amin Hariyadi, M.T., selaku Dosen Pembimbing I yang telah memberikan bimbingan, arahan, kritik, dan saran selama penyusunan tesis.
6. Dr. Fresy Nugroho,S.T.,M.T.,IPM.,ASEAN Eng selaku Dosen Pembimbing II yang telah memberikan bimbingan, motivasi, dan masukan yang membangun.
7. Dr. Cahyo Crisdiyan, S.T.,M.Cs selaku Dosen Penguji I yang telah memberikan kritik dan saran untuk penyempurnaan tesis.

8. Dr. Agung Teguh Wibowo Almais,S.Kom.,M,T, selaku Dosen Penguji II yang telah memberikan masukan dan koreksi yang sangat bermanfaat.
9. Seluruh Dosen Program Studi Magister Informatika dan Civitas Akademika ,yang telah memberikan ilmu dan pengalaman selama masa perkuliahan.

Akhir kata**,** penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan doa, dukungan, bimbingan, dan bantuan selama proses penyusunan tesis ini. Penulis berharap semoga tesis ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan serta menjadi referensi bagi penelitian selanjutnya dibidang Informatika. Penulis juga menyadari bahwa tesis ini masih memiliki keterbatasan, Semoga Allah Subḥānahu wa Ta‘ālā senantiasa melimpahkan rahmat dan keberkahan kepada kita semua. Āmīn yā Rabbal 'Ālamīn.

Malang,06 Juni 2026

Penulis,

Ida oktavia Salsavana

DAFTAR ISI

HALAMAN PENGAJUAN……..…..ii
HALAMAN PERSETUJUAN…….iii
MOTTO……iv
HALAMAN PERSEMBAHAN……v
KATA PENGANTAR…….....vi
DAFTAR GAMBAR 13
DAFTAR TABEL 14
ABSTRAK 15
ABSTRACT 16
BAB I 1
PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Pernyataan Masalah..... 6
1.3 Tujuan Penelitian..... 6
1.4 Batasan Masalah..... 6
1.5 Manfaat Penelitian..... 7
STUDI PUSTAKA 8
2.1 Klasifikasi Penyakit Hepatitis C 8
2.2 Landasan Teori..... 17
2.2.1 Pengertian Hepatitis C 17
2.2.2Klasifikasi dalam Diagnosis Medis..... 19
2.2.3 Pembahasan Al-Qur'an dan Hadis Tentang Penyakit 19
2.2.4 Algoritma Naïve Bayes 23
2.2.5Algoritma *Random Forest*..... 24
2.2.6 Naive Bayes dan Random Forest 27
2.3 Kerangka Teori..... 28
BAB III 30
METODOLOGI PENELITIAN..... 30
3.1 Desain Penelitian..... 30
3.1.1 Pengumpulan Data 31
3.1.2 Validasi Data 33

3.1.3 Desain Sistem 33
3.1.4 Eksperimen 33
3.1.5 Hasil dan Evaluasi 34
3.16 Analisis dan Kesimpulan 34
3.2 Kerangka Konseptual 34
3.3 Skenario Uji Coba 37
3.4 Skenario Pengujian Hasil 39
BAB IV 40
IMPLEMENTASI METODE *NAÏVE BAYES* 40
4.1 Desain Sistem Naïve Bayes 40
4.2 Uji Coba Niave Bayes 47
4.1.1 Split Dataset 47
4.1.2 Training Model 48
4.1.3 Model dan Akurasi 48
BAB V 51
5.2 Uji coba Random Forest 59
5.2.1 Split Dataset 60
5.2.2 Training Model 60
5.2.3 Test Model dan Akurasi 61
BAB VI 63
HASIL DAN PEMBAHASAN 63
6.1 Pembahasan 63
6.1.1 Pustaka Python yang Digunakan 63
6.1.2 Cleaning Data 64
6.1.3 Transpormasi Data 65
6.1.4 Analisis Data 65
6.1.6 Heatmap Korelasi 69
6.5 Perbandingan Akurasi Model 80
6.6 ROC Curve Comparison 82
BAB VII KESIMPULAN 85
7.1 Kesimpulan 85
7.2 Saran 86
DAFTAR PUSTAKA 87

DAFTAR GAMBAR

Gambar 2.1 Majority-Voting..... 25
Gambar 2.2: Kerangka Teori..... 29
Gambar 3. 1: Desain Penelitian..... 31
Gambar 4.1 Flowchart Naïve Bayes..... 41
Gambar 4.2 Confusion Matrix NB..... 50
Gambar 5.1 flowchat Random Forest..... 52
Gambar 5.2 Confusion Matrix RF..... 6
Gambar 6.1 Histogram Hepatitis C..... 67
Gambar 6.2: Heatmap..... 70
Gambar 6.3: Age Distribution..... 72
Gambar 6.4 Albumin Distribution..... 73
Gambar 6.5 ALT Distribution..... 75
Gambar 6.7 Feature Importance dari model..... 76
Gambar 6.9 ROC Curve Comparison..... 82

DAFTAR TABEL

Tabel 3.1 Data Atteribut 32
Tabel 4.1: Tabel Probabilitas Tes Sampel 46
Tabel 4.2 Hasil Akurasi NB 48
Tabel 6.1 Sebelum Cleaning 64
Tabel 6.2 Cek Tipe Data 65
Tabel 6. 3 Lood Dataset 66
Tabel 6. 4: Nilai Ferforma *Random Forest* 77
Tabel 6.5 Nilai Ferforma *Naive Bayes* 79
Tabel 6.6 Perbandingan Akurasi Model 80
Tabel 6.7 Tabel hasil klasifikasi 83

ABSTRAK

Salsavana Oktavia Ida 2026. **Analisis Klasifikasi Penyakit Hepatitis C Menggunakan Algoritma *Naive Bayes* Dan *Random Forest.*** Tesis. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (I) Dr.Ir.Mokhamad Amin Hriyadi,M.T. Pembimbing (II) Dr. Fresy Nugroho,S.T,M.T.,I.PM.,ASEAN Eng.

Hepatitis C merupakan penyakit infeksi hati yang memerlukan deteksi dini untuk mencegah komplikasi yang lebih serius. Penelitian ini bertujuan untuk membangun model klasifikasi diagnosis Hepatitis C menggunakan algoritma Naïve Bayes dan Random Forest serta membandingkan kinerja kedua algoritma tersebut. Dataset klinis yang digunakan melalui tahapan preprocessing yang meliputi data cleaning, transformasi data, dan penyeimbangan kelas menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan data. Hasil analisis menunjukkan bahwa variabel Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Gamma-Glutamyl Transferase (GGT), Age, dan Albumin (ALB) merupakan faktor yang paling berpengaruh dalam proses klasifikasi penyakit Hepatitis C. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Berdasarkan hasil pengujian dengan rasio pembagian data 70% data latih dan 30% data uji, algoritma Random Forest memperoleh performa terbaik dengan nilai akurasi sebesar 95,14%, precision 0,74, recall 0,68, dan F1-score 0,70. Sementara itu, algoritma Naïve Bayes memperoleh akurasi sebesar 91,35%, precision 0,63, recall 0,70, dan F1-score 0,66. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu membantu meningkatkan keseimbangan distribusi kelas sehingga model dapat melakukan klasifikasi dengan lebih baik. Berdasarkan perbandingan performa yang diperoleh, algoritma Random Forest terbukti lebih unggul dibandingkan Naïve Bayes dalam klasifikasi penyakit Hepatitis C dan berpotensi digunakan sebagai metode pendukung diagnosis yang lebih akurat.

Kata Kunci: *Hepatitis C, Naïve Bayes, Random Forest, SMOTE,Classification.*

ABSTRACT

Salsavana Oktavia Ida 2026. **Analisis Klasifikasi Penyakit Hepatitis C Menggunakan Algoritma *Naive Bayes* Dan *Random Forest.*** Tesis. Program Study Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing (I) Dr.Ir.Mokhamad Amin Hriyadi,M.T. Pembimbing (II) Dr. Fresy Nugroho,S.T,M.T.,I.PM.,ASEAN Eng.

Hepatitis C is a liver infection that requires early detection to prevent more serious complications. This study aims to build a Hepatitis C diagnosis classification model using the Naïve Bayes and Random Forest algorithms and compare the performance of the two algorithms. The clinical dataset used went through preprocessing stages including data cleaning, data transformation, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE) method to address data imbalance. The analysis results showed that the variables Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Gamma-Glutamyl Transferase (GGT), Age, and Albumin (ALB) were the most influential factors in the Hepatitis C disease classification process. Model evaluation was carried out using accuracy, precision, recall, and F1-score metrics. Based on the test results with a data sharing ratio of 70% training data and 30% test data, the Random Forest algorithm obtained the best performance with an accuracy value of 95.14%, precision of 0.74, recall of 0.68, and F1-score of 0.70. Meanwhile, the Naïve Bayes algorithm achieved an accuracy of 91.35%, a precision of 0.63, a recall of 0.70, and an F1-score of 0.66. The results showed that the application of SMOTE was able to help improve the balance of class distribution so that the model could perform better classification. Based on the performance comparison obtained, the Random Forest algorithm was proven to be superior to Naïve Bayes in classifying Hepatitis C and has the potential to be used as a supporting method for more accurate diagnosis.

Keywords: Hepatitis C, *Naive Bayes*, *Random Forest*, *SMOTE*Classification.

ملخص

سالسافانا أوكتافيا إيدا، 2026. تحليل تصنيف مرض التهاب الكبد الوبائي سي باستخدام خوارزميتي بايز الساذج العشوائية. رسالة ماجستير في المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم والغابة الإسلامية الحكومية، مالانج. المشرف الأول: د. محمد أمين هريادي، ماجستير في الهندسة. المشرف الثاني: د. فريسي نوغروهو، بكالوريوس في الهندسة، ماجستير في الهندسة، مدير برنامج معتمد، مهندس معتمد من رابطة دول جنوب شرق آسيا (آسيان).

يُعدّ التهاب الكبد الوبائي سي عدوى تصيب الكبد، ويتطلب الكشف المبكر عنها تجنب مضاعفات أكثر خطورة. تهدف هذه الدراسة إلى بناء نموذج لتصنيف تشخيص التهاب الكبد الوبائي سي باستخدام خوارزميتي بايز الساذج والغابة العشوائية، ومقارنة أداء الخوارزميتين. خضعت مجموعة البيانات السريرية المستخدمة لمراحل معالجة مسبقة شملت تنظيف البيانات، وتحويلها، وموازنة الفئات باستخدام تقنية التضخيم الاصطناعي للأقلية (SMOTE) ، وناقلة (AST) لمعالجة عدم توازن البيانات. أظهرت نتائج التحليل أن متغيرات ناقلة أمين الأسبارتات كانت من أكثر العوامل تأثيرًا في عملية (ALB) ، والعمر، والألبومين(GGT) ، وناقلة غاما غلوتاميل(ALT) أمين الألانين تصنيف مرض التهاب الكبد الوبائي سي. وقد تم تقييم النموذج باستخدام مقاييس الدقة، والضبط، والاستدعاء، وبناءً على نتائج الاختبار بنسبة مشاركة بيانات 70% لبيانات التدريب و30% لبيانات الاختبار، حققت F1. ومقياس بلغ F1 خوارزمية الغابة العشوائية أفضل أداء بدقة بلغت 95.14%، وضبط 0.74، واستدعاء 0.68، ومقياس 0.70. في المقابل، حققت خوارزمية بايز البسيطة دقة بلغت 91.35%، وضبط 0.63، واستدعاء 0.70، ومقياس ساهم في تحسين توازن توزيع الفئات، مما مكّن النموذج SMOTE بلغ 0.66. أظهرت النتائج أن تطبيق تقنية F1 من تصنيف البيانات بشكل أفضل. وبناءً على مقارنة الأداء، تبيّن أن خوارزمية الغابة العشوائية تتفوق على خوارزمية بايز البسيطة في تصنيف التهاب الكبد الوبائي سي، ولها إمكانية استخدامها كطريقة مساعدة لتشخيص أكثر دقة.

الكلمات المفتاحية: التهاب الكبد الوبائي سي، بايز البسيطة، الغابة العشوائية، تصنيف

BAB I

PENDAHULUAN

1.1 Latar Belakang

Hepatitis C merupakan salah satu penyakit infeksi hati yang disebabkan oleh virus Hepatitis C (HCV) dan menjadi masalah kesehatan masyarakat di berbagai negara. Penyakit ini dapat berkembang menjadi kondisi kronis yang berpotensi menyebabkan sirosis hati, gagal hati, hingga kanker hati apabila tidak terdeteksi dan ditangani sejak dini. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa jutaan orang di seluruh dunia hidup dengan infeksi Hepatitis C kronis, dan sebagian besar kasus tidak terdiagnosis karena gejalanya sering kali tidak muncul pada tahap awal penyakit (WHO, 2024).

Hepatitis C ditularkan melalui kontak dengan darah yang terkontaminasi virus, seperti penggunaan jarum suntik yang tidak steril, transfusi darah yang tidak aman, serta penggunaan alat medis yang tidak memenuhi standar kebersihan. Selain itu, faktor risiko lain yang dapat meningkatkan kemungkinan seseorang terinfeksi Hepatitis C meliputi riwayat penggunaan narkoba suntik, prosedur medis invasif, tato atau tindik dengan peralatan yang tidak steril, serta riwayat kontak dengan penderita Hepatitis C. Penyakit ini dapat menyerang berbagai kelompok usia, meskipun prevalensinya lebih tinggi pada orang dewasa (WHO, 2024).

Berdasarkan tingkat keparahannya, Hepatitis C dapat dibedakan menjadi infeksi akut dan infeksi kronis. Hepatitis C akut terjadi dalam enam bulan pertama setelah seseorang terpapar virus. Sebagian penderita dapat sembuh secara spontan, namun sebagian besar berkembang menjadi Hepatitis C kronis.

Pada kondisi kronis, virus terus menginfeksi hati dalam jangka waktu yang lama sehingga meningkatkan risiko terjadinya kerusakan hati permanen. Oleh karena itu, deteksi dan diagnosis dini menjadi langkah penting dalam mencegah komplikasi yang lebih serius (Alghamdi et al., 2023).

Terdapat ayat Al-Qur'an yang mengajarkan manusia untuk menjaga kesehatan dan menghindari perbuatan yang dapat membahayakan diri sendiri. Allah SWT berfirman dalam Surah Al-Baqarah ayat 195 yang artinya:

وَاَنْفِقُوْا فِيْ سَبِيْلِ اللّٰهِ وَلَا تُلْقُوْا بِاَيْدِيْكُمْ اِلَى التَّهْلُكَةِ ۛ وَاَحْسِنُوْا ۛ اِنَّ اللّٰهَ يُحِبُّ الْمُحْسِنِيْنَ

"Dan janganlah kamu menjatuhkan dirimu sendiri ke dalam kebinasaan." (Q.S. Al-Baqarah: 195).

Menurut Tafsir Ibnu Katsir, ayat ini mengandung perintah agar manusia menjaga dirinya dari berbagai hal yang dapat menyebabkan kerusakan dan kebinasaan, termasuk penyakit yang timbul akibat kurangnya perhatian terhadap kesehatan. Dalam konteks Hepatitis C, ayat ini mengajarkan pentingnya pencegahan melalui perilaku hidup sehat, menjaga kebersihan, serta menghindari berbagai faktor yang dapat menyebabkan

penularan penyakit.

Allah SWT juga berfirman dalam Surah Asy-Syu'ara ayat 80:

Dalam era digital dan perkembangan teknologi informasi yang pesat, data mining telah menjadi salah satu pendekatan yang banyak dimanfaatkan dalam bidang kesehatan. Data mining memungkinkan peneliti dan tenaga medis untuk mengolah data kesehatan dalam jumlah besar guna menemukan pola, hubungan, dan informasi penting yang sulit diidentifikasi secara manual. Dalam konteks deteksi penyakit Hepatitis C, penerapan teknik klasifikasi memiliki potensi besar dalam membantu pengembangan model prediktif yang mampu mengidentifikasi status penyakit berdasarkan data klinis dan hasil pemeriksaan laboratorium pasien (Chao & Li, 2022).

Dalam konteks ini, pemilihan metode yang tepat untuk klasifikasi penyakit Hepatitis C menjadi sangat penting. Dua algoritma yang sering digunakan dalam klasifikasi data medis adalah Naive Bayes dan Random Forest. Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi independensi antar atribut. Keunggulan algoritma ini terletak pada proses komputasi yang cepat, sederhana, serta mampu memberikan hasil klasifikasi yang baik meskipun digunakan pada dataset yang besar dan kompleks. Naive Bayes juga efektif dalam menangani data medis yang memiliki banyak atribut dan sering digunakan dalam berbagai penelitian kesehatan (Zhang, 2004).

Random Forest merupakan salah satu metode ensemble learning yang menggabungkan sejumlah pohon keputusan (decision tree) untuk

menghasilkan model klasifikasi yang lebih stabil dan akurat. Keunggulan utama algoritma ini adalah kemampuannya dalam mengolah data yang kompleks, mengurangi risiko overfitting, serta mengidentifikasi variabel yang memiliki pengaruh besar terhadap hasil klasifikasi.

Kemampuan tersebut menjadikan Random Forest sebagai salah satu algoritma yang banyak digunakan dalam klasifikasi penyakit dan analisis data kesehatan (Breiman, 2001). Pendekatan Naive Bayes dan Random Forest untuk klasifikasi penyakit Hepatitis C melibatkan penggunaan berbagai parameter klinis dan hasil pemeriksaan laboratorium, seperti kadar Albumin,Bilirubin, Creatinine, Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), serta indikator kesehatan lainnya yang berkaitan dengan fungsi hati. Dengan memanfaatkan dataset tersebut, model klasifikasi dapat dikembangkan untuk membantu mengidentifikasi pasien yang terindikasi menderita Hepatitis C secara lebih cepat dan akurat (Kaur & Kumari, 2022).

Naive Bayes telah terbukti efektif dalam berbagai penelitian klasifikasi penyakit karena kemampuannya menghasilkan model yang sederhana namun memiliki tingkat akurasi yang cukup baik. Algoritma ini mampu bekerja secara efisien meskipun jumlah data yang digunakan cukup besar. Sementara itu, Random Forest sering digunakan dalam klasifikasi penyakit karena kemampuannya menangani data yang kompleks dan menghasilkan performa klasifikasi yang tinggi. Beberapa penelitian menunjukkan bahwa Random Forest mampu memberikan hasil yang lebih

baik dalam mengidentifikasi pola hubungan antar variabel dibandingkan metode klasifikasi lainnya (Breiman, 2001; Alghamdi et al., 2023).

Tujuan penelitian ini berfokus pada perbandingan algoritma Naive Bayes dan Random Forest dalam klasifikasi penyakit Hepatitis C dengan pendekatan data mining untuk menganalisis tingkat akurasi kedua algoritma. Sementara banyak penelitian sebelumnya telah menerapkan berbagai algoritma machine learning dalam klasifikasi penyakit Hepatitis C, penelitian ini tidak hanya berfokus pada tingkat akurasi klasifikasi, tetapi juga bertujuan untuk mengevaluasi efektivitas kedua algoritma berdasarkan nilai precision, recall, dan F1-score dalam menghasilkan model klasifikasi yang optimal. Dengan demikian, penelitian ini bertujuan untuk mengkaji lebih mendalam kemampuan algoritma Naive Bayes dan Random Forest dalam meningkatkan kualitas klasifikasi penyakit Hepatitis C sehingga dapat mendukung proses deteksi dini dan pengambilan keputusan medis yang lebih akurat (Kaur & Kumari, 2022).

Pendekatan ini relevan dengan berbagai penelitian sebelumnya yang menunjukkan bahwa Random Forest memiliki kemampuan yang baik dalam menangani data medis yang kompleks dan melibatkan banyak variabel, sedangkan Naive Bayes mampu memberikan performa klasifikasi yang efisien dengan waktu komputasi yang relatif cepat pada dataset kesehatan (Breiman, 2001; Zhang, 2004). Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi dalam menentukan algoritma yang lebih efektif untuk klasifikasi penyakit Hepatitis C. Hasil penelitian ini

diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan di bidang kesehatan yang memanfaatkan teknik data mining dan machine learning untuk meningkatkan akurasi diagnosis serta membantu upaya pencegahan dan penanganan penyakit Hepatitis C secara lebih dini (Alghamdi et al., 2023).

1.2 Pernyataan Masalah

1. Bagaimana mencari akurasi dalam klasifikasi diagnosis penyakit *Hepatitis C* menggunakan algoritma *Naive Bayes* dan *Random Forest*?
2. Bagaimana Hasil klasifikasi diagnosis penyakit *Hepatitis C* ?

1.3 Tujuan Penelitian

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

1.4 Batasan Masalah

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

1. Keterbatasan *Dataset* yang Digunakan Penelitian yg di lakukan menggunakan *dataset* yang tersedia secara publik, seperti *dataset kaggle*, sehingga hasil klasifikasi dan *model prediktif* yang

dikembangkan mungkin tidak mencerminkan populasi yang lebih luas atau faktor risiko yang unik di luar data tersebut.

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

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

4. Tidak menangani aspek klinis secara mendalam

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

1.5 Manfaat Penelitian

1. Bagi Peneliti

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

2. Bagi akademis atau Pembaca

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

BAB II

STUDI PUSTAKA

2.1 Klasifikasi Penyakit Hepatitis C

Penelitian yang dilakukan oleh *Zulfiqar et al. (2024)* bertujuan untuk menganalisis dan membandingkan performa beberapa algoritma machine learning dalam klasifikasi penyakit Hepatitis C. Penelitian ini menggunakan dataset Hepatitis C yang terdiri dari data demografis pasien, hasil pemeriksaan laboratorium, serta indikator kesehatan lainnya yang berhubungan dengan fungsi hati. Algoritma yang digunakan dalam penelitian ini meliputi Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), dan Multilayer Perceptron (MLP). Evaluasi model dilakukan menggunakan beberapa metrik pengukuran, yaitu accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma Multilayer Perceptron memperoleh performa terbaik dengan tingkat akurasi sebesar 92,68%, sedangkan SVM memperoleh akurasi sebesar 91,86%. Sementara itu, algoritma Naive Bayes dan Random Forest masing-masing menghasilkan akurasi sebesar 89,43%. Penelitian ini menunjukkan bahwa metode machine learning mampu membantu proses klasifikasi penyakit Hepatitis C secara efektif serta dapat digunakan sebagai pendukung pengambilan keputusan dalam bidang kesehatan. *Zulfiqar et al. (2024).*

Penelitian yang dilakukan oleh *Safdari et al. (2022)* bertujuan untuk mengembangkan model klasifikasi pasien yang diduga terinfeksi virus Hepatitis C dengan memanfaatkan berbagai teknik data mining. Dataset yang digunakan berasal dari UCI Machine Learning Repository yang berisi data hasil pemeriksaan laboratorium pasien.

Dalam penelitian ini digunakan beberapa algoritma klasifikasi, yaitu Random Forest, Gaussian Naive Bayes, Support Vector Machine, Decision Tree, Logistic Regression, dan K-Nearest Neighbor. Sebelum proses klasifikasi dilakukan, data terlebih dahulu melalui tahap preprocessing dan penyeimbangan kelas menggunakan metode Synthetic Minority Oversampling Technique (SMOTE). Hasil pengujian menunjukkan bahwa Random Forest memberikan performa terbaik dibandingkan algoritma lainnya dengan tingkat akurasi sebesar 97,29% dan nilai Area Under Curve (AUC) sebesar 0,998. Penelitian ini membuktikan bahwa Random Forest memiliki kemampuan yang sangat baik dalam mengidentifikasi pola pada data medis sehingga dapat digunakan sebagai metode yang efektif dalam membantu proses diagnosis penyakit Hepatitis C. *Safdari et al. (2022).*

Penelitian yang dilakukan oleh *Lilhore et al. (2023)* bertujuan untuk meningkatkan akurasi klasifikasi penyakit Hepatitis C dengan mengembangkan model hibrida yang menggabungkan Improved Random Forest dan Support Vector Machine.

Penelitian ini menggunakan dataset Hepatitis C yang diperoleh dari UCI Machine Learning Repository dengan berbagai atribut klinis seperti Albumin, Bilirubin, Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), dan Creatinine. Untuk meningkatkan kualitas data, peneliti menerapkan teknik feature selection dan metode Synthetic Minority Oversampling Technique (SMOTE) guna mengatasi ketidak seimbangan data. Hasil penelitian menunjukkan bahwa penggunaan Improved Random Forest mampu meningkatkan kemampuan model dalam memilih fitur-fitur penting yang berpengaruh terhadap klasifikasi penyakit. Selain itu, kombinasi Improved Random Forest dan SVM menghasilkan performa yang lebih baik dibandingkan penggunaan algoritma tunggal. Penelitian ini menunjukkan bahwa pengembangan algoritma Random Forest dapat memberikan kontribusi yang signifikan dalam meningkatkan akurasi klasifikasi penyakit Hepatitis *C. Lilhore et al. (2023).*

Penelitian yang dilakukan oleh *Bagur dan Pratama (2025)* bertujuan untuk membandingkan performa algoritma Random Forest, Explainable Boosting Machine (EBM), dan Ensemble Stacking dalam klasifikasi penyakit Hepatitis C. Dataset yang digunakan terdiri dari 615 data pasien yang dikategorikan ke dalam beberapa tingkat kondisi kesehatan. Penelitian ini menerapkan teknik oversampling untuk mengatasi permasalahan ketidakseimbangan kelas yang sering ditemukan pada data medis. Evaluasi

model dilakukan menggunakan accuracy, precision, recall, F1-score, dan Area Under Curve (AUC). Hasil penelitian menunjukkan bahwa penerapan oversampling mampu meningkatkan performa seluruh model yang digunakan. Random Forest menghasilkan nilai F1-score sebesar 0,8086 dan nilai AUC sebesar 0,9710 sehingga mampu mengidentifikasi kelas minoritas dengan baik. Penelitian ini membuktikan bahwa Random Forest merupakan algoritma yang efektif untuk menangani permasalahan klasifikasi data medis yang memiliki distribusi data tidak seimbang. *Bagur dan Pratama (2025).*

Penelitian yang dilakukan oleh *Senbagamalar dan Logeswari (2024)* bertujuan untuk mengembangkan model klasifikasi multi-kelas pada data kanker menggunakan metode Divergent Random Forest yang dikombinasikan dengan Genetic Clustering Algorithm sebagai teknik seleksi fitur. Dataset yang digunakan terdiri dari data ekspresi gen berbagai jenis kanker, yaitu kanker payudara, kanker paru-paru, kanker ginjal, kanker usus besar, dan kanker prostat. Tingginya dimensi data ekspresi gen menyebabkan proses klasifikasi menjadi lebih kompleks sehingga diperlukan metode seleksi fitur yang efektif. Hasil penelitian menunjukkan bahwa Genetic Clustering Algorithm mampu mengurangi jumlah fitur yang tidak relevan dan meningkatkan kualitas data yang digunakan dalam proses klasifikasi. Selain itu, Divergent Random Forest menghasilkan tingkat akurasi yang tinggi dalam membedakan berbagai jenis kanker berdasarkan karakteristik genetiknya. Penelitian ini membuktikan bahwa Random Forest

sangat efektif digunakan pada data medis yang kompleks dan berdimensi tinggi karena mampu mempertahankan akurasi klasifikasi yang baik. *Senbagamalar dan Logeswari (2024).*

Penelitian yang dilakukan oleh *Fan et al. (2023)* bertujuan untuk mengembangkan sistem prediksi penyakit Hepatitis C yang bersifat interpretable atau mudah dijelaskan menggunakan pendekatan machine learning. Penelitian ini menggunakan dataset Hepatitis C dari UCI Machine Learning Repository yang berisi data pemeriksaan serologis pasien, seperti AST, ALT, ALP, bilirubin, dan indikator kesehatan hati lainnya. Algoritma yang digunakan dalam penelitian ini meliputi Random Forest (RF), Support Vector Machine (SVM), dan AdaBoost. Selain itu, penelitian ini juga menerapkan metode Explainable Artificial Intelligence (XAI) menggunakan SHAP dan LIME untuk menjelaskan proses pengambilan keputusan model. Evaluasi model dilakukan menggunakan accuracy, precision, recall, F1-score, dan AUC. Hasil penelitian menunjukkan bahwa algoritma Random Forest yang dioptimasi dengan Bayesian Optimization memperoleh performa terbaik dengan accuracy sebesar 99,44% dan AUC sebesar 0,9986 pada dataset UCI. Penelitian ini menunjukkan bahwa kombinasi machine learning dan explainable AI mampu meningkatkan akurasi prediksi sekaligus memberikan interpretasi yang transparan terhadap proses klasifikasi penyakit Hepatitis C. *Fan et al. (2023).*

Penelitian yang dilakukan oleh *I Gede Hendrayana et al. (2025)* bertujuan untuk menganalisis implementasi algoritma Random Forest

dengan teknik resampling dan tanpa resampling dalam klasifikasi penyakit Hepatitis C. Dataset yang digunakan berasal dari Kaggle dengan jumlah 615 data pasien yang terdiri dari atribut demografis, hasil laboratorium, dan indikator kesehatan hati.

Penelitian ini membandingkan tiga pendekatan, yaitu Random Forest tanpa resampling, Random Forest dengan SMOTE+ENN, dan Random Forest dengan SMOTE+OSS untuk mengatasi masalah ketidakseimbangan kelas data. Evaluasi model dilakukan menggunakan accuracy, precision, recall, F1-score, dan AUC-ROC. Hasil penelitian menunjukkan bahwa metode Random Forest dengan teknik SMOTE+ENN memberikan performa terbaik dengan accuracy sebesar 99,19% dan ROC AUC sebesar 0,9999 serta mampu mendeteksi kelas minoritas secara lebih optimal dibandingkan metode lainnya. Penelitian ini membuktikan bahwa penggunaan teknik resampling dapat meningkatkan performa klasifikasi penyakit Hepatitis C dan membantu pengembangan sistem diagnosis berbasis kecerdasan buatan yang lebih akurat. *I Gede Hendrayana et al. (2025).*

Penelitian yang dilakukan oleh *Shalini Kumari et al. (2025)* bertujuan untuk memprediksi penyakit Hepatitis C menggunakan berbagai model statistical learning dan machine learning. Dataset yang digunakan berasal dari UCI Machine Learning Repository dengan jumlah 615 data pasien yang terdiri dari data demografis serta hasil pemeriksaan laboratorium. Algoritma yang digunakan dalam penelitian ini meliputi

Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), dan Artificial Neural Network (ANN). Penelitian ini juga menerapkan teknik Synthetic Minority Oversampling Technique (SMOTE) untuk menangani ketidakseimbangan data. Evaluasi model dilakukan menggunakan accuracy, Brier Score (BS), Matthews Correlation Coefficient (MCC), serta berbagai metrik klasifikasi lainnya. Hasil penelitian menunjukkan bahwa algoritma Random Forest memperoleh performa terbaik dengan accuracy sebesar 97,41%, Brier Score sebesar 0,026, dan MCC sebesar 0,947 setelah penerapan feature selection dan SMOTE. Penelitian ini menunjukkan bahwa penerapan machine learning dengan teknik penanganan imbalance data mampu meningkatkan akurasi prediksi penyakit Hepatitis C secara efektif dan dapat membantu proses deteksi dini penyakit. *Shalini Kumari et al. (2025).*

Penelitian yang dilakukan oleh *Alfina Damayanti dan Gusmelia Testiana (2023)* bertujuan untuk menerapkan teknik data mining dalam prediksi penyakit Hepatitis C menggunakan algoritma Naïve Bayes. Dataset yang digunakan berasal dari Kaggle dengan jumlah 586 data pasien yang terdiri dari 13 atribut, meliputi data donor darah, hepatitis, fibrosis, dan cirrhosis. Penelitian ini memanfaatkan data pasien untuk menggali informasi baru yang dapat digunakan dalam proses klasifikasi penyakit Hepatitis C. Evaluasi model dilakukan untuk mengetahui tingkat akurasi algoritma Naïve Bayes dalam memprediksi kemungkinan pasien menderita Hepatitis C. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes

mampu diterapkan dalam klasifikasi penyakit Hepatitis C dengan menghasilkan nilai akurasi sebesar 93,61% yang termasuk dalam kategori Good Classification. Penelitian ini menunjukkan bahwa metode data mining menggunakan algoritma Naïve Bayes dapat membantu proses prediksi penyakit Hepatitis C secara cepat dan efektif. *Alfina Damayanti dan Gusmelia Testiana (2023).*

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

Tabel 2.1 Penelitian terkait 1

No	Peneliti	Topik	Metode	Hasil
1	Zulfiqar et al. (2024)	Klasifikasi penyakit Hepatitis C	SVM, Naive Bayes, Random Forest, dan MLP	MLP memperoleh akurasi terbaik sebesar 92,68%, SVM 91,86%, sedangkan Naive Bayes dan Random Forest masing-masing 89,43%.
2	Safdari et al. (2022)	Klasifikasi pasien terduga Hepatitis C	Random Forest, Gaussian Naive Bayes, SVM, Decision Tree, Logistic Regression, dan KNN dengan SMOTE	Random Forest memberikan performa terbaik dengan akurasi 97,29% dan nilai AUC sebesar 0,998.
3	Lilhore et al. (2023)	Peningkatan akurasi klasifikasi Hepatitis C	Improved Random Forest, SVM, dan SMOTE	Kombinasi Improved Random Forest dan SVM menghasilkan performa klasifikasi lebih baik dibanding algoritma tunggal.
4	Bagur dan Pratama (2025)	Klasifikasi penyakit Hepatitis C pada data tidak seimbang	Random Forest, Explainable Boosting Machine (EBM), dan Ensemble Stacking	Random Forest menghasilkan F1-score sebesar 0,8086 dan AUC sebesar 0,9710 sehingga efektif mendeteksi kelas minoritas.

5	Senbagamalar dan Logeswari (2024)	Klasifikasi multi-kelas data kanker	Divergent Random Forest dan Genetic Clustering Algorithm	Genetic Clustering Algorithm meningkatkan kualitas seleksi fitur dan Divergent Random Forest menghasilkan akurasi klasifikasi tinggi.
6	Fan et al. (2023)	Prediksi penyakit Hepatitis C berbasis Explainable AI	Random Forest, SVM, AdaBoost, SHAP, dan LIME	Random Forest dengan Bayesian Optimization memperoleh accuracy sebesar 99,44% dan AUC sebesar 0,9986.
7	I Gede Hendrayana et al. (2025)	Klasifikasi Hepatitis C dengan teknik resampling	Random Forest, SMOTE+ENN, dan SMOTE+OSS	Random Forest dengan SMOTE+ENN memperoleh accuracy sebesar 99,19% dan ROC AUC sebesar 0,9999.
8	Shalini Kumari et al. (2025)	Prediksi Hepatitis C menggunakan statistical learning	Naive Bayes, Random Forest, SVM, LR, DT, ANN, dan SMOTE	Random Forest memperoleh accuracy sebesar 97,41%, Brier Score 0,026, dan MCC sebesar 0,947.
9	Alfina Damayanti dan Gusmelia Testiana (2023)	Prediksi penyakit Hepatitis C menggunakan data mining	Algoritma Naïve Bayes	Algoritma Naïve Bayes menghasilkan akurasi sebesar 93,61% dalam kategori Good Classification.

2.2 Landasan Teori

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

2.2.1 Pengertian Hepatitis C

Hepatitis C merupakan salah satu penyakit infeksi pada hati yang disebabkan oleh virus Hepatitis C atau *Hepatitis C Virus* (HCV). Penyakit ini termasuk dalam kelompok penyakit hepatitis virus yang menyerang organ hati dan dapat menyebabkan peradangan akut maupun kronis. HCV merupakan virus RNA yang termasuk dalam famili *Flaviviridae* dan genus *Hepacivirus* (WHO, 2025). Virus ini menyerang sel hati (*hepatosit*) sehingga mengakibatkan kerusakan fungsi hati secara bertahap. Infeksi Hepatitis C sering disebut sebagai *silent disease* karena sebagian besar penderita tidak menunjukkan gejala pada tahap awal infeksi. Penyakit Hepatitis C disebabkan oleh infeksi virus Hepatitis C (HCV). Virus ini dapat masuk ke dalam tubuh manusia melalui aliran darah dan kemudian menyerang organ hati. Penularan virus Hepatitis C dapat terjadi melalui beberapa faktor, antara lain penggunaan jarum suntik secara bergantian, transfusi darah yang terkontaminasi virus HCV, penggunaan alat medis yang tidak steril, transplantasi organ dari donor yang terinfeksi, serta penggunaan alat tato atau tindik yang tidak higienis (WHO, 2025).

Pada tahap awal, penderita Hepatitis C sering kali tidak menunjukkan gejala yang spesifik. Namun, seiring perkembangan penyakit, beberapa gejala yang dapat muncul antara lain mudah lelah, mual, muntah, penurunan nafsu makan, nyeri pada bagian perut kanan atas, demam ringan, serta kulit dan mata yang menguning (*jaundice*) (WHO, 2025).Apabila infeksi berlangsung dalam jangka waktu lama tanpa penanganan yang tepat, maka penyakit dapat berkembang menjadi fibrosis, sirosis hati, gagal hati, bahkan kanker hati (*hepatocellular carcinoma*) (Safdari et al., 2022). Oleh karena itu, pemeriksaan kesehatan secara rutin sangat penting dilakukan untuk mendeteksi penyakit sejak dini.

Diagnosis penyakit Hepatitis C dilakukan melalui pemeriksaan laboratorium untuk mendeteksi keberadaan antibodi maupun RNA virus Hepatitis C dalam darah pasien. Pemeriksaan yang umum dilakukan meliputi tes Anti-HCV, tes HCV RNA, pemeriksaan fungsi hati seperti ALT, AST, bilirubin, albumin, serta pemeriksaan biopsi hati atau *fibroscan* untuk mengetahui tingkat kerusakan hati (WHO, 2025).

Pengobatan Hepatitis C bertujuan untuk menghilangkan virus dari tubuh serta mencegah terjadinya kerusakan hati yang lebih parah. Saat ini pengobatan modern menggunakan *Direct Acting Antiviral* (DAA) yang memiliki tingkat keberhasilan sangat tinggi dalam menyembuhkan infeksi Hepatitis C (WHO, 2025).

2.2.2Klasifikasi dalam Diagnosis Medis

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

2.2.3 Pembahasan Al-Qur'an dan Hadis Tentang Penyakit

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

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

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

مَا أَنْزَلَ اللهُ دَاءً إِلاَّ أَنْزَلَ لَهُ شِفَاءً

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

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

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

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

Penerapan teknologi untuk mempermudah diagnosis dan klasifikasi penyakit adalah bentuk nyata menolong sesama**.** Dengan membantu pasien atau tenaga medis mengenali risiko hepatitis c secara lebih cepat dan akurat, maka peneliti ikut meringankan beban penderita.

Konsep tersebut sejalan dengan prinsip kesehatan modern yang menekankan pentingnya keseimbangan nutrisi serta penerapan sikap moderasi dalam berbagai aspek kehidupan. Dengan demikian, penelitian ini tidak hanya fokus pada aspek medis semata, tetapi juga hubungan

dengan nilai-nilai keseimbangan yang menjadi dasar dalam menjaga kesehatan secara holistik, tetapi juga menghubungkannya dengan aspek spiritual yang dapat membantu dalam pengelolaan hepatitis c secara holistik *holistik* Care & Suppl. (2020).

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

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

1. Tawakal dan ikhtiar yang seimbang

وَإِذَا مَرِضْتُ فَهُوَ يَشْفِينِ

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

Ayat ini menanamkan keyakinan bahwa kesembuhan mutlak di tangan Allah. Namun, keyakinan ini tidak bertentangan dengan upaya penelitian.penelitian Anda adalah sebab (wasilah) yang diciptakan dan dimudahkan oleh Allah untuk mencapai kesembuhan. Seorang peneliti Muslim meyakini bahwa algoritma *Random Forest* dan *Naive Bayes* yang berhasil dikembangkan adalah atas izin dan pertolongan-Nya. Ini mencegah kesombongan ilmiah dan senantiasa mengingatkan pada Sang Maha Penyembuh.

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

مَا أَنْزَلَ ٱللَّهُ دَاءً إِلَّ أَنْزَلَ لَهُ شِفَاءً

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

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

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

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

وَتَعَاوَنُوا عَلَى الْبِ ِر وَالتَّقْوٰى ۖ وَلَ تَعَاوَنُوا عَلَى الْثِْمِ وَالْعُدْوَانِ

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

Menyelamatkan orang dari penyakit kronis adalah kebaikan (al-birr) yang sangat nyata. Kolaborasi antara dunia teknologi informasi (Anda sebagai peneliti Informatika) dengan dunia medis untuk menangani hepatitis adalah bentuk ta'awun'alal birri yang modern. Penelitian Anda adalah kontribusi nyata dari disiplin ilmu komputer untuk kemanusiaan.

2. Menjadi sebab terlepasnya kesusahan orang lain

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

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

Seorang pasien yang tidak terdiagnosis hepatitis atau terlambat

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

2.2.4 Algoritma Naïve Bayes

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

Rumus *Naive Bayes* didasarkan pada Teorema Bayes:

$$P(C|X) = \frac{P(X|C)\cdot P(C)}{P(X)} \qquad (1)$$

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

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

1. *Preprocessing Data*:

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

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

- Probabilitas setiap kelas dihitung dari distribusi data pelatihan.

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

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

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

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

4. Kalkulasi probabilitas Posterior *P(C|X)):*

- *Gabungkan Prior dan likelihood:*

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

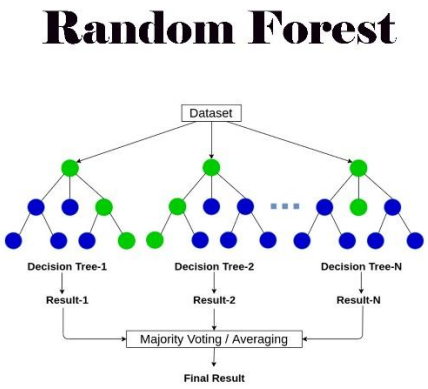
5. Klasifikasi

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

2.2.5Algoritma *Random Forest*

Dikutip dari Carstensen *et al.* (2020) *Random Forest* adalah jenis algoritma *Supervised Learning* yang memanfaatkan *Decision Trees* untuk membuat Klasifikasi. *Random Forest* termasuk dalam *Ensemble Learning,* yaitu dimana pendekatan dilakukan dengan beberapa algoritma yang digabungkan untuk meningkatkan hasil Klasifikasi. Dalam hal ini *Random Forest* menggunakan *Decision Tree.* Secara khusus, *Random Forest*

digunakan untuk klasifikasi data kedalam kategori atau label yang telah ditentukan. *Random Forest*), *Decision Tree* yang dibangung akan menggunakan subset acak dari data pelatihan dengan penggantian sehingga menciptakan variasi yang membantu mengurangi *overfitting*. Setelah *Decision Trees* dibangun, Klasifikasi akhir diperoleh dengan *voting* dari semua *Decision Trees* Yesa *et al.* (2023).



2.1 Gambar Majority-Voting 1

Berikut adalah langkah dan elemen penting dari algoritma Random Forest

1. Pemilihan Fitur

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

$$F = RandomSubset\ (M) \tag{6}$$

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

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

Entropi $Entropy\ (D)_{i=\overline{T}} - \sum^{n} pi \log (Pi)$ (7)

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

Dimana:

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

Split terbaik adalah yang meminimalkan entropi atau Gini Index.

3. Klasifikasi *Voting* Mayoritas

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

$$y = mode\ (y1, y2\ \ Yn) \quad (8)$$

Dimana:

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

Rumus Skor Akhir (Probabilitas untuk Setiap Kelas)

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

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

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

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

2.2.6 Naive Bayes dan Random Forest

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

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

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

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

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

Rumus ini juga dapat dijelaskan sebagai berikut:

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

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

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

Rumus ini juga dapat dijelaskan sebagai berikut:

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

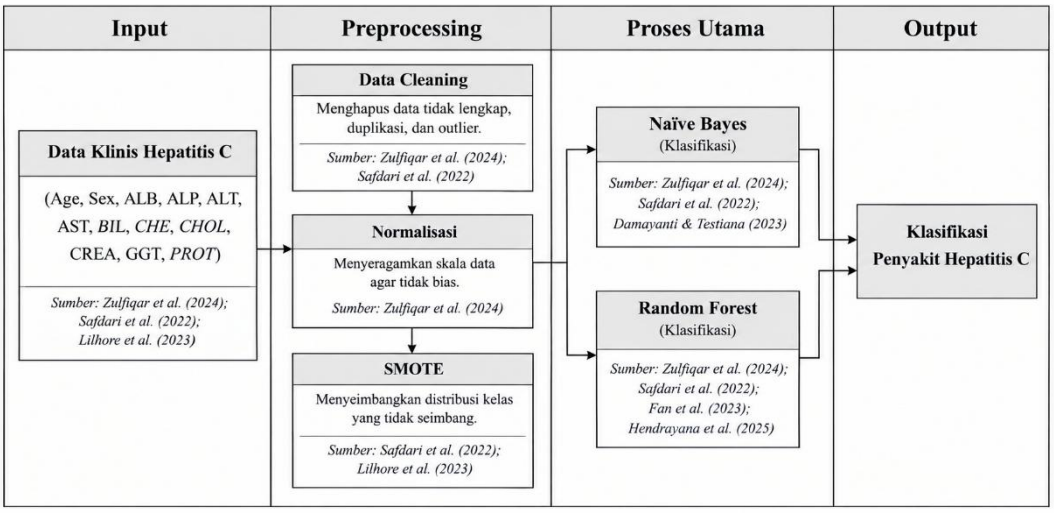
F1-Score merupakan rata-rata harmonis antara *precision* dan *recall*, memberikan keseimbangan antara keduanya.

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

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

2.3 Kerangka Teori

Kerangka teori (Theoretical Framework) adalah suatu struktur (bangunan) teori atau konsep konsep yang saling terhubung untuk menjelaskan eksistensi/keberadaan masalah penelitian yang kita bahas. Theoretical framework akan digambarkan pada gambar 2.2 :



Gambar 2.2: Kerangka Teori

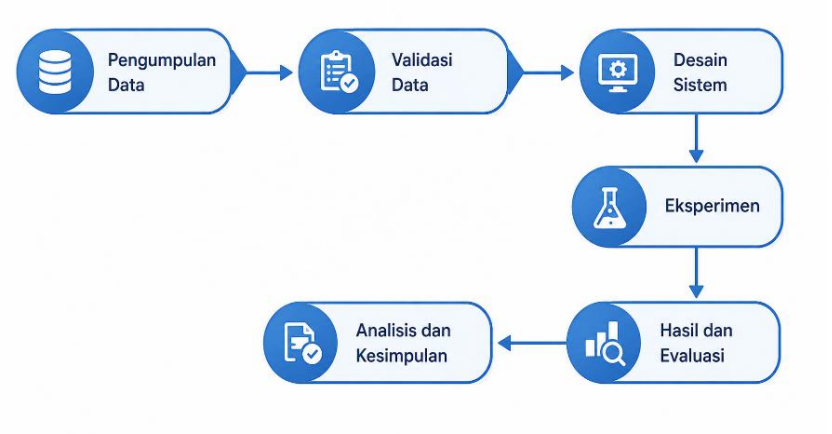
2.2 tersebut menunjukankerangka teori yang menggambarkan alur kerja penelitian dalam membangun model prediksi status Hepatitis C berbasis teknik data mining. Pada komponen Input, penelitian memanfaatkan data klinis dan biokimia pasien yang meliputi usia, jenis kelamin, status donor darah, kadar bilirubin, albumin, enzim hati (AST, ALT, γ-GT, ALP), kolesterol, total protein, kreatinin, serta parameter laboratorium lainnya yang memiliki keterkaitan dengan fungsi hati. Variabel-variabel tersebut dipilih berdasarkan literatur yang menunjukkan bahwa parameter tersebut merupakan indikator yang signifikan dalam mendeteksi gangguan atau infeksi pada organ hati, termasuk Hepatitis C.

BAB III

METODOLOGI PENELITIAN

3.1 Desain Penelitian

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



Gambar 3.1 Desain Penelitian

3.1.1 Pengumpulan Data

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

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

b. Jurnal yang digunakan oleh penulis yaitu penelitian dari 5 tahun terakhir dengan tema Hepatitis C.

A. Atteribut Data

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

Tabel 3.1 Data Atteribut

No.	Atribut	Tipe	Keterangan
1	Att1	integer	(Pasien ID/No.)
2	Age	integer	(Usia)
3	Sex	polynominal	(Jenis Kelamin)
4	ALB	polynominal	(Jumlah albumin)
5	ALP	polynominal	(Jumlah alkaline phosphatase)
6	ALT	polynominal	(Jumlah alanin transaminase)
7	AST	polynominal	(Jumlah aspartat aminotransferase)
8	BIL	polynominal	(Jumlah bilirubin)
9	CHE	polynominal	(Jumlah kolinesterase)
10	CHOL	polynominal	(Jumlah kolesterol)
11	CREA	polynominal	(Jumlah kreatinin)
12	GGT	polynominal	(Jumlah gamma-glutamil transferase)
13	PROT	polynominal	(Jumlah protein)
14	Category (Diagnosa)	polynominal	0= Blood Donor (0= donor darah) 3= cirrhosis (penyakit hati stadium akhir) 2= Fobrosis (Jaringan parut fibrotik) 1=Hepatitis (1=peradangan hati) 0_S= Suspect Blood Donor (0s=Dugaan Donor Darah)

Untuk menemukan atribut yang tepat atau optimal bagi permasalahan tersebut, diperlukan klasifikasi menyeluruh terhadap semua atribut yang ada, serta mengabaikan atribut yang tidak relevan. *Dataset* input yang tercantum pada Tabel

3.1 mencakup berbagai atribut beserta deskripsinya. Pemilihan atribut yang sesuai dengan kualitasnya akan meningkatkan kualitas kumpulan data masukan dan hasil klasifikasi yang diharapkan.

3.1.2 Validasi Data

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

3.1.3 Desain Sistem

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

3.1.4 Eksperimen

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

hepatitis c atau non- hepatitis c.

3.1.5 Hasil dan Evaluasi

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

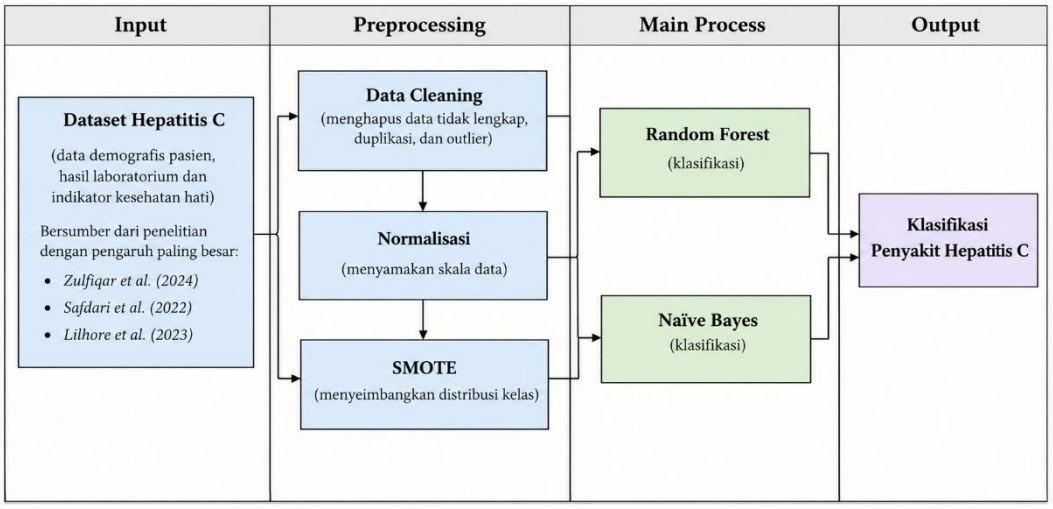
3.16 Analisis dan Kesimpulan

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

3.2 Kerangka Konseptual

Kerangka konseptual menggambarkan tahapan proses serta keterkaitan antar komponen yang terlibat dalam penelitian ini. Melalui kerangka tersebut, alur mulai dari data masukan, proses prapengolahan, proses utama (yang mencakup ekstraksi fitur dan klasifikasi menggunakan algoritma *Naïve Bayes* dan *Random Forest*), hingga menghasilkan keluaran dapat dijelaskan secara sistematis.

Kerangka ini memberikan pemahaman menyeluruh mengenai hubungan antar elemen dalam mencapai tujuan akhir, yaitu klasifikasi penyakit Hepatitis C. Kerangka konsep disajikan pada Gambar 3.2.



Gambar 3.2 Kerangka Konseptual

Kerangka konseptual pada penelitian ini menggambarkan alur sistematis dalam proses klasifikasi penyakit Hepatitis C berdasarkan data pasien yang diperoleh dari dataset Hepatitis C. Data yang digunakan terdiri dari atribut demografis dan hasil pemeriksaan laboratorium, yaitu Age, Sex, Albumin (ALB), Alkaline Phosphatase (ALP), Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Bilirubin (BIL), Cholinesterase (CHE), Cholesterol (CHOL), Creatinine (CREA), Gamma Glutamyl Transferase (GGT), Protein (PROT), serta kategori diagnosis yang menunjukkan kondisi pasien seperti Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis, dan Cirrhosis. Atribut-atribut tersebut dipilih karena memiliki hubungan yang erat dengan fungsi hati dan sering digunakan dalam penelitian klasifikasi Hepatitis C.

Tahap preprocessing dilakukan untuk meningkatkan kualitas data sebelum proses pemodelan. Pada tahap ini diterapkan metode normalisasi untuk menyamakan rentang nilai setiap atribut sehingga perbedaan skala

antar fitur tidak memengaruhi proses pembelajaran model. Selain itu, digunakan teknik Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi permasalahan ketidakseimbangan kelas yang umum ditemukan pada data medis. Penerapan SMOTE bertujuan meningkatkan jumlah data pada kelas minoritas sehingga model dapat mengenali pola setiap kelas dengan lebih baik dan menghasilkan performa klasifikasi yang lebih optimal.

Pada tahap utama (main process), data yang telah melalui proses preprocessing kemudian diklasifikasikan menggunakan dua algoritma machine learning, yaitu Random Forest dan Naïve Bayes. Algoritma Random Forest dipilih karena memiliki kemampuan yang baik dalam menangani data dengan banyak atribut, mampu mengurangi risiko overfitting, serta menghasilkan tingkat akurasi yang tinggi melalui pendekatan ensemble dari beberapa decision tree. Sementara itu, algoritma Naïve Bayes digunakan karena memiliki proses komputasi yang sederhana, efisien, dan mampu melakukan klasifikasi berdasarkan pendekatan probabilistik sehingga cocok digunakan sebagai metode pembanding.

Berdasarkan berbagai penelitian terdahulu, seperti Safdari et al. (2022), Fan et al. (2023), dan I Gede Hendrayana et al. (2025), algoritma Random Forest menunjukkan performa yang sangat baik dalam klasifikasi penyakit Hepatitis C, terutama ketika dikombinasikan dengan teknik penanganan ketidakseimbangan data seperti SMOTE. Oleh karena itu,

kedua algoritma tersebut dipilih untuk dianalisis dan dibandingkan pada penelitian ini.

Hasil dari proses klasifikasi kemudian dievaluasi menggunakan beberapa metrik pengukuran, seperti accuracy, precision, recall, dan F1-score untuk mengetahui tingkat performa masing-masing algoritma. Output akhir dari kerangka konseptual ini adalah keputusan klasifikasi kondisi pasien ke dalam kategori penyakit Hepatitis C berdasarkan karakteristik data yang dimiliki. Hasil tersebut diharapkan dapat membantu proses identifikasi dan pendukung keputusan dalam diagnosis penyakit Hepatitis C secara lebih cepat, objektif, dan akurat.

3.3 Skenario Uji Coba

Tujuan skenario uji coba adalah rancangan atau skema pengujian yang dibuat untuk memastikan bahwa sistem atau model yang dikembangkan diuji secara terstruktur, terukur, dan objektif berdasarkan tujuan penelitian. Dalam konteks penelitian Anda yang menggunakan algoritma *Random Forest* dan *Naïve Bayes* untuk klasifikasi penyakit Hepatitis C, skenario uji coba menjelaskan bagaimana proses pengujian dilakukan, data apa yang digunakan, tahapan evaluasinya, dan kriteria keberhasilan model.

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

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

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

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

Rumus ini juga dapat dijelaskan sebagai berikut:

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

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

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

Rumus ini juga dapat dijelaskan sebagai berikut:

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

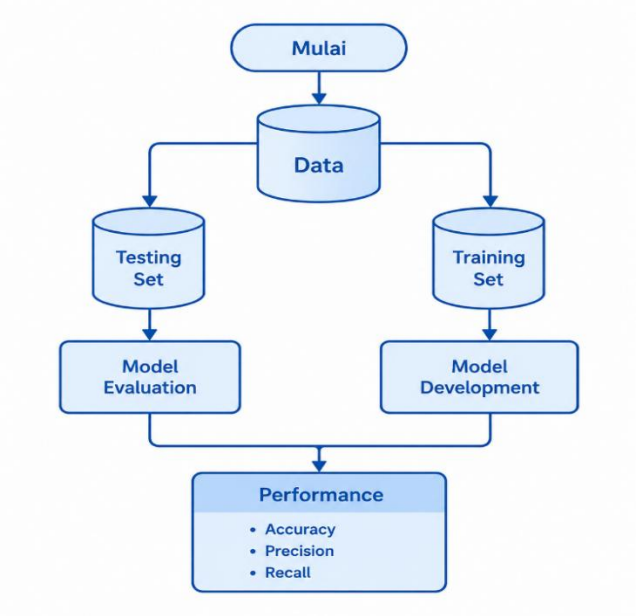
F1-Score merupakan rata-rata harmonis antara *precision* dan *recall*, memberikan keseimbangan antara keduanya.

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

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

3.4 Skenario Pengujian Hasil

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



Gambar 3.3 Pengujian Hasil

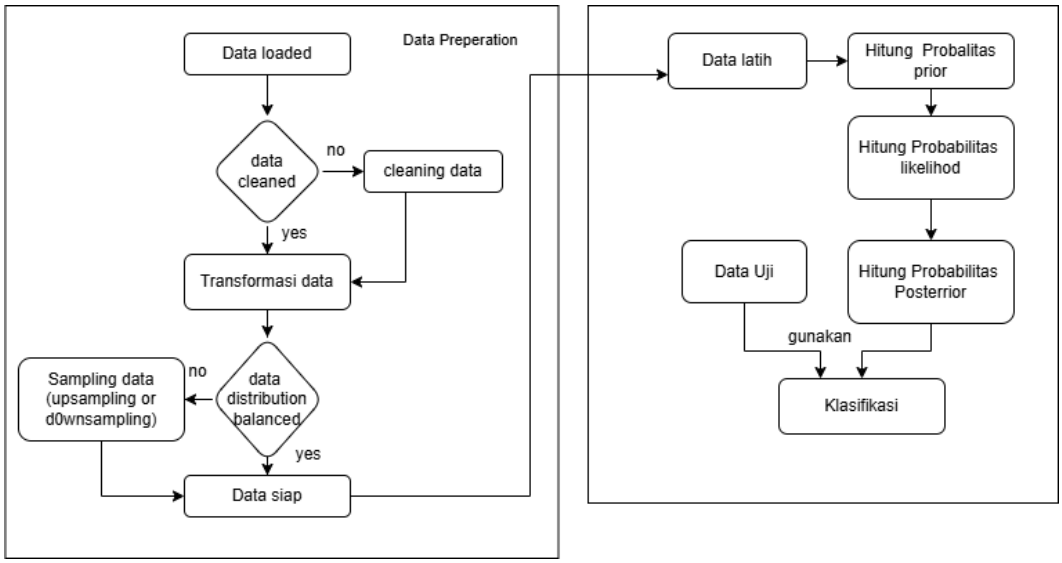
Gambar menunjukkan alur pembentukan dan evaluasi model klasifikasi. Data dibagi menjadi training set untuk membangun model dan testing set untuk menguji model. Kinerja model kemudian dievaluasi menggunakan metrik accuracy, precision, dan recall.

BAB IV

IMPLEMENTASI METODE *NAÏVE BAYES*

4.1 Desain Sistem Naïve Bayes

Desain sistem Naïve Bayes pada penelitian ini menggambarkan alur proses klasifikasi penyakit Hepatitis C mulai dari tahap data preparation hingga menghasilkan keputusan klasifikasi. Proses diawali dengan memuat dataset, melakukan data cleaning apabila masih terdapat data yang tidak lengkap atau tidak konsisten, kemudian mentransformasikan data ke dalam format yang sesuai untuk proses pembelajaran. Selanjutnya dilakukan pemeriksaan distribusi kelas dan penerapan teknik penyeimbangan data apabila diperlukan sehingga diperoleh dataset yang siap digunakan. Dataset kemudian dibagi menjadi data latih dan data uji, di mana data latih digunakan untuk menghitung probabilitas prior, likelihood, dan posterior berdasarkan Teorema Bayes, sedangkan data uji digunakan untuk menghasilkan prediksi kelas berdasarkan nilai probabilitas posterior tertinggi. Alur lengkap desain sistem Naïve Bayes pada penelitian ini ditunjukkan pada Gambar 4.1.



Gambar 4.1 Flowchart Naïve Bayes

Fakhri *et al*. (2023)

1. Tahap Data Preparation (Persiapan Data)

a. Data Loaded (Data Dimuat)

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

b. Data Cleaned (Pembersihan Data)

Flowchart menunjukkan pengambilan keputusan: apakah data sudah bersih atau belum

Jika belum, maka dilakukan proses cleaning data seperti:

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

c. Transformasi Data

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

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

d. Pengecekan Distribusi Data

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

2. Tahap Modeling dengan Algoritma Naïve Bayes

a. Pemisahan Data

Data yang telah siap dibagi menjadi:

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

b. Perhitungan Probabilitas Prior

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

- P(hepatitis C) dan P(tidak hepatitis C).

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

c. Perhitungan Probabilitas Likelihood

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

d. Perhitungan Probabilitas Posterior

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

Posterior = prior × likelihood

e. Proses Klasifikasi

- Data uji dimasukkan ke dalam model,
- Model menggunakan probabilitas posterior untuk menentukan apakah pasien termasuk kategori hepatitis C atau Blood Donor (Sehat),
- Hasil akhirnya berupa prediksi klasifikasi.

4.1 Implementasi *Naive Bayes*

Adalah algoritma yang digunakan untuk klasifikasi dengan pendekatan probabilistik. Algoritma ini efektif karena hanya memerlukan jumlah data yang sedikit untuk menentukan parameter klasifikasi. Berikut tahapan untuk melakukan perhitungan menggunakan algoritma *Naive Bayes* dengan jumlah data sebanyak 615 data dan setelah dilakukan pembersihan data dan oversampling menjadi 1066 data. Tahapan dalam

menggunakan algoritma *Naive Bayes* adalah sebagai berikut:

1. Teorema *Naive Bayes*

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

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

2. Menghitung Probabilitas Prior

   Dihitung berdasarkan jumlah data dalam setiap kelas dibandingkan total sampel.

$$\text{P(hepatitis} = \text{yes)} = \frac{\text{jumlah sakit}}{\text{total sampel}} = \frac{378}{756} = 0.50$$

$$\text{P(hepatitis} = \text{yes)} = \frac{\text{jumlah sakit}}{\text{total sampel}} = \frac{378}{756} = 0.50$$

3. Menghitung Mean (μ)

   Untuk kelas "yes":

   μage, yes $= \frac{44+\cdots+61}{373} = 48.50$

   μALB, yes $= \frac{30.1 + \ldots + 38.2}{373} = 35.80$

   μAST, yes $= \frac{45.2 + \ldots + 80.5}{373} = 65.40$

   μAST, yes $= \frac{50.1 + \ldots + 110.2}{373} = 82.10$

   Untuk kelas "no":

   μage, no $= \frac{32 + \ldots + 45}{373} = 40.20$

   μALB, no $= \frac{40.5 + \ldots + 46.2}{373} = 42.50$

   μAST, no $= \frac{15.2 + \ldots + 25.5}{373} = 20.30$

   μAST, no $= \frac{18.1 + \ldots + 26.4}{373} = 22.40$

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

Likelihood dihitung menggunakan distribusi *Gaussian*:

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

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

$(x\ -\ \mu)2\ 2\sigma 2^{)}$Untuk kelas "yes":

$$P(Age\ =\ 50\ |\ yes)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 120.5}}\ \cdot exp\ (\frac{(50\ -\ 48.50)^2}{(2\ \cdot\ 120.5)})\ =\ 0.036$$

$$P(ALB\ =\ 32\ |\ yes)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 45.2}}\ \cdot exp\ (\frac{(32\ -\ 35.80)^2}{(2\ \cdot\ 45.2)})\ =\ 0.051$$

$$P(ALT\ =\ 70\ |\ yes)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 210.4}}\cdot exp\ (\frac{(70\ -\ 65.40)^2}{(2\ \cdot\ 210.4)})\ =\ 0.026$$

$$P(AST\ =\ 85\ |\ yes)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 315.6}}\ \cdot exp\ (\frac{(85\ -\ 82.10)^2}{(2\ \cdot\ 315.6)})\ =\ 0.022$$

Untuk kelas "no":

$$P(Age\ =\ 50\ |\ no)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 95.4}}\ \cdot exp\ (\frac{(50\ -\ 40.20)^2}{(2\ \cdot\ 95.4)})\ =\ 0.024$$

$$P(ALB\ =\ 32\ |\ no)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 30.5}}\ \cdot exp\ (\frac{(32\ -\ 42.50)^2}{(2\ \cdot\ 30.5)})\ =\ 0.012$$

$$P(ALT\ =\ 70\ |\ no)\ =\ \frac{1}{\sqrt{2\pi\ \cdot 50.2}}\cdot exp\ (\frac{(70\ -\ 20.30)^2}{(2\ \cdot\ 50.2)})\ =\ 0.0001$$

$$P(AST\ =\ 85\ |\ no)\ =\ \frac{1}{\sqrt{2\pi\ \cdot\ 45.8}}\ \cdot exp\ (\frac{(85\ -\ 22.40)^2}{(2\ \cdot\ 45.8)})\ =\ 0.00005$$

5. Menghitung Posterior

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

Untuk kelas Hepatitis C “yes”:

P(yes | X) $\propto$ P(yes) · P(Age=50 | yes) · P(ALB=32 | yes) · P(ALT=70 | yes) · P(AST=85 | yes)

$= 0.50 \cdot 0.036 \cdot 0.051 \cdot 0.026.0.002$

$= 5.25 \text{ x } 10^{-7}$

Untuk kelas Hepatitis C“no”:

P(no | X) $\propto$ P(no) · P(Age=50 | no) · P(ALB=32 | no) · P(ALT=70 | no) · P(AST=85 | no)

$= 0.50 \cdot 0.024 \cdot 0.012 \cdot 0.0001 \cdot 0.00005$

$= 7.20 \text{ x } 10^{-13}$

6. Memilih Kelas dengan Probabilitas

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

Tabel 4.1: Tabel Probabilitas Tes Sampel

	Ya	**Tidak**
Prior-Probabilitas	0.50	0.50
Age (52.0)	0.036	0.024
ALB (41.2)	0.051	0.012
ALT (20.4)	0.026	0.0001
AST (18.8)	0.022	0.00005
Posterior Probabilitas	**5.25 x 10^{-7}**	**7.20 x 10^{-13}**

4.2 Uji Coba Niave Bayes

Uji coba model Naïve Bayes dilakukan untuk mengukur kinerja model dalam mengklasifikasikan data diagnosis penyakit Hepatitis C. Pengujian ini penting untuk menilai sejauh mana model mampu mengenali pola dan memklasifikasi kelas targetsecara akurat. Dalam uji coba ini, digunakan data uji yang sebelumnya telah dipisahkan dari data pelatihan agar evaluasi lebih objektif.

4.1.1 Split Dataset

Dataset pada penelitian ini telah dibersihkan dari missing value dan diseimbangkan menggunakan teknik oversampling sehingga total data menjadi 1.066 baris. Data ini dibagi menjadi dua bagian utama, yaitu data training dan data testing, dengan proporsi masing-masing sebesar 70% untuk training (746 data) dan 30% untuk testing (320 data). Proporsi ini ditentukan menggunakan parameter test_size = 0.3, yang bertujuan untuk memastikan sebagian besar data digunakan dalam proses pelatihan model, sementara sisanya dialokasikan untuk menguji kinerja model yang telah dilatih. Pembagian dilakukan secara acak dengan menetapkan parameter random_state = 42, sehingga hasil pembagian akan konsisten setiap kali kode dijalankan ulang.

4.1.2 Training Model

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

4.1.3 Model dan Akurasi

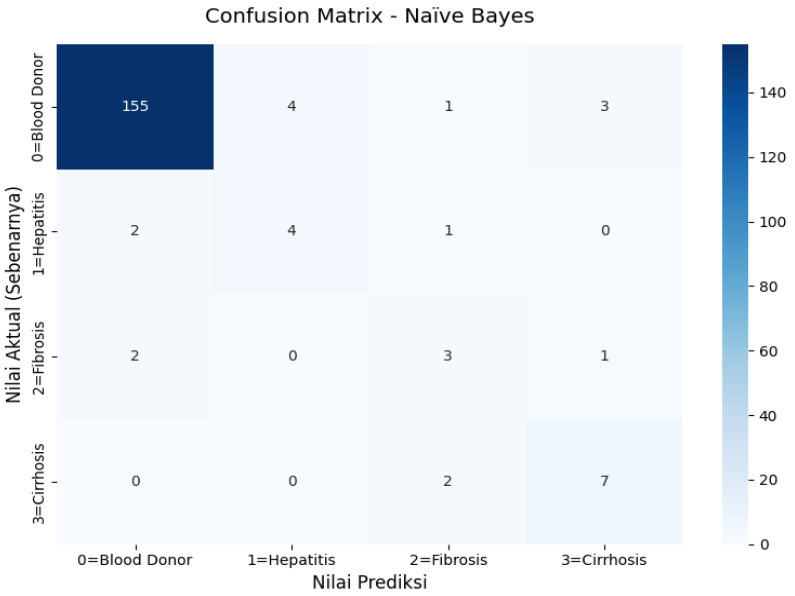
Setelah proses training, selanjutnya model akan melakukan klasifikasi dengan data test. Setelah pelatihan selesai, model akan dievaluasi menggunakan metrik *accuracy, F1-score, recall, dan precision.*

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

Tabel 4.2 Hasil Akurasi NB

	precision	recall	f1-score	support
0=Blood Donor	0.97	0.95	0.96	163
1=Hepatitis	0.50	0.57	0.53	7
Accuracy			0.91	185
macro avg	0.63	0.70	0.66	185
weighted avg	0.92	0.91	0.91	185

Tabel 4.2 tersebut menampilkan laporan evaluasi model Naïve Bayes dalam mengklasifikasikan data Hepatitis C. Model ini mencapai akurasi sebesar 84%, yang menunjukkan bahwa sebagian besar klasifikasi yang dilakukan oleh sistem sudah sesuai dengan label aktualnya. Untuk kelas 0 (bukan Hepatitis C), diperoleh nilai precision sebesar 0,78 dan recall yang sangat tinggi yaitu 0,98. Hal ini menandakan bahwa model sangat peka dan unggul dalam mengenali pasien yang benar-benar negatif. Sementara itu, untuk kelas 1 (Hepatitis C), model mencatatkan nilai precision yang sangat tinggi mencapai 0,97. Artinya, ketika model memprediksi seseorang positif menderita Hepatitis C, tingkat kepastian prediksi tersebut bernilai benar adalah 97%. Namun, nilai recall pada kelas 1 sebesar 0,68 menunjukkan bahwa model masih melewatkan beberapa kasus pasien yang sebenarnya positif (diklasifikasikan keliru sebagai negatif). Secara keseluruhan, nilai macro average F1-score sebesar 0,83 dan weighted average F1-score sebesar 0,84 membuktikan bahwa performa model tergolong seimbang dan baik dalam menangani kedua kelas tersebut, berkat proses penyeimbangan data latih di tahap sebelumnya.



Gambar 4.2 Confusion Matrix NB 1

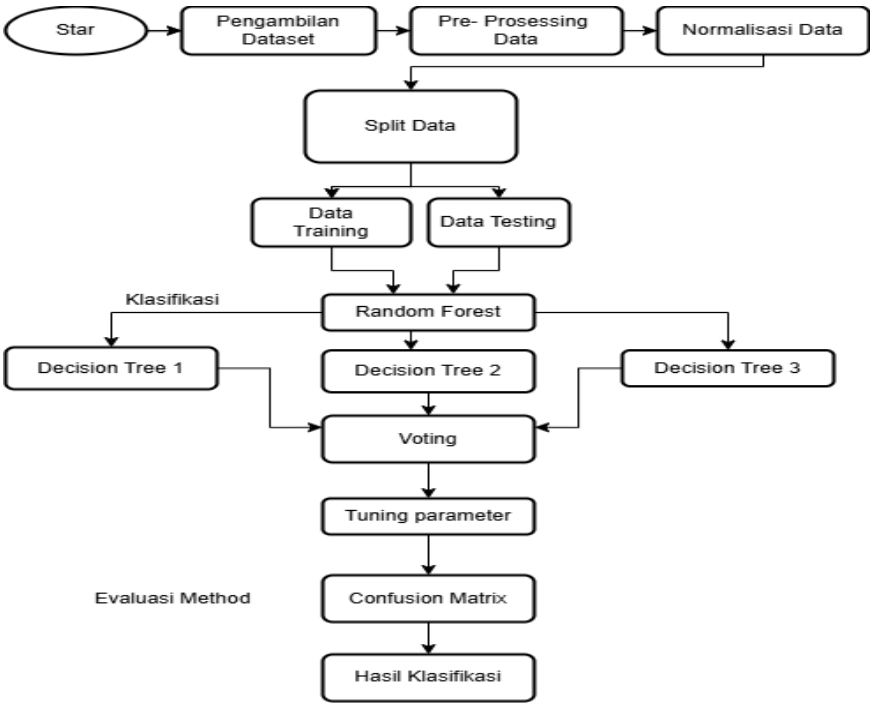
Gambar confusion matrix menunjukkan kinerja model Naïve Bayes dalam mengklasifikasikan data Hepatitis C secara multi-kelas. Dari total 185 data uji, model Naïve Bayes menunjukkan kemampuan klasifikasi yang baik dalam membedakan Blood Donor, namun masih menunjukkan misklasifikasi pada kelas Hepatitis, Fibrosis, dan Cirrhosis. Hal ini dikarenakan karakteristik pola pada kelas penyakit tersebut saling beririsan.

BAB V

IMPLEMENTASI METODE *RANDOM FOREST*

5.1 Desain Sistem *Random Forest*

Pada penelitian ini, algoritma Random Forest digunakan sebagai salah satu metode klasifikasi untuk mendiagnosis penyakit Hepatitis C berdasarkan data klinis pasien. Desain sistem disusun untuk menggambarkan alur kerja model mulai dari proses pengambilan dataset, tahap prapemrosesan (pre-processing), pembentukan model klasifikasi, hingga evaluasi hasil prediksi. Perancangan alur sistem ini bertujuan untuk memastikan bahwa setiap tahapan pengolahan data dilakukan secara sistematis sehingga model yang dihasilkan mampu memberikan performa klasifikasi yang optimal.



Gambar 5.1 Flowchart Random Forest

1. Mulai

Proses dimulai dengan tahap inisialisasi seluruh langkah kerja penelitian.

2. Impor Dataset

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

3. Feature Selection (Pemilihan Fitur)

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

4. Normalisasi Data

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

5. Split Data

Data Train (latih): digunakan untuk membangun model.

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

6. Bangun Beberapa Decision Tree

Pada tahap ini, algoritma Random Forest membangun sejumlah decision tree (pohon keputusan). Setiap pohon keputusan (*decision tree*)

dalam algoritma dilatih menggunakan subset data serta subset fitur yang dipilih secara acak.

7. Lakukan Voting Mayoritas

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

8. Model RF Terlatih

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

9. Evaluasi Model

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

10. Confusion Matrix

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

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

Berdasarkan hasil confusion matrix, dilakukan perhitungan terhadap sejumlah metrik evaluasi untuk menilai kinerja model klasifikasi. Metrik tersebut meliputi: accuracy, yang menunjukkan tingkat ketepatan

keseluruhan prediksi model; precision, yang mengukur proporsi prediksi positif yang benar; recall, yang menggambarkan kemampuan model dalam mengidentifikasi data positif secara akurat; serta F1-score, yang merepresentasikan keseimbangan antara nilai precision dan recall.

12. Tuning Parameter

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

13. Hasil Klasifikasi

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

5.1 Implementasi *Random Forest*

Random Forest adalah algoritma ensemble learning yang menggabungkan banyak decision tree untuk meningkatkan akurasi klasifikasi dan mengurangi overfitting. Dalam klasifikasi, *Random Forest* melakukan klasifikasi berdasarkan voting mayoritas dari seluruh pohon. Dengan jumlah data sebanyak 615. Tahapan dalam menggunakan algoritma *Random Forest* adalah sebagai berikut:

1. Persiapan Data

a. Memahami *Dataset*:

- Dataset awal terdiri dari 615 baris data rekam medis pasien (sebelum dilakukan tahap cleaning dan penyeimbangan data menggunakan oversampling). Atribut yang digunakan meliputi usia, jenis kelamin, serta kadar enzim/protein dalam darah seperti ALB, ALT, AST, dan BIL.

b. Membersihkan Data:

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

c. Normalisasi/Standarisasi :

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

d. Pisahkan *Dataset*:

- Bagi *dataset* menjadi data pelatihan (*training*) dan pengujian (*testing*), misalnya:
  - 70% untuk pelatihan (746 data).
  - 30% untuk pengujian (185 data).

1. Konfigurasi Model *Random Forest*

a. *Bootstrap Sampling*:

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

a. Pelatihan Pohon

Setiap pohon keputusan dilatih menggunakan subset data pelatihan yang dipilih secara acak dengan metode pengambilan sampel dengan penggantian (bootstrap sampling) dari total 746 data yang tersedia.

b. Pemilihan Fitur Secara Acak

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

1. Untuk setiap data uji dari 320 data:
- Masukkan data ke masing-masing pohon.
- Setiap pohon memberikan klasifikasi kelas (*y^i*).
- Lakukan voting mayoritas dari

Semua pohon: y^= $mode(y1, y2, yn$

2. Jika diperlukan probabilitas, hitung

rasio pohon yang mengklasifikasi kelas

tertentu :

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

4. Evaluasi Model

Gunakan data uji untuk mengukur performa model:

1. Metode Evaluasi:

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

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

Tabel 5.1 bootstrap sampling data

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...
611	3=Cirrhosis	62	f	32.0	416.6	5.9	110.3	50.0	5.57	6.30	55.7	650.9	68.5
612	3=Cirrhosis	64	f	24.0	102.8	2.9	44.4	20.0	1.54	3.02	63.0	35.9	71.3
613	3=Cirrhosis	64	f	29.0	87.3	3.5	99.0	48.0	1.66	3.63	66.7	64.2	82.0
614	3=Cirrhosis	46	f	33.0	NaN	39.0	62.0	20.0	3.56	4.20	52.0	50.0	71.0
615	3=Cirrhosis	59	f	36.0	NaN	100.0	80.0	12.0	9.07	5.30	67.0	34.0	68.0

615 rows × 13 columns

Berdasarkan banyaknya data sampel latih (*training data*) yang digunakan setelah proses penyeimbangan dengan oversampling, terdapat 373 data pasien yang positif Hepatitis C dan 373 data yang negatif, dengan total keseluruhan 746 data latih. Maka peluang dari variabel respon untuk kategori positif dan negatif adalah sama:

$$P(y = no) = \frac{n(y = no)}{n(total)} = \frac{373}{746} = 0.50$$

$$P(y = yes = \frac{n(y = yes)}{n(total)} = \frac{373}{746} = 0.50$$

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

Entropy (Y) = − [P(no) · $\log_2$(P(no)) + P(yes) · $\log_2$(P(yes))]

Entropy (Y) = − [(0.50) · $\log_2$(0.50) + (0.50) · $\log_2$(0.50)]

Mengingat $\log_2$(0.50) = -1, maka:

Entropy (Y) = − [(0.50 · -1) + (0.50 · -1)]

Entropy (Y) = − [-0.50 − 0.50]

Entropy (Y) = 1.0

Tabel 5.2 confusion matrix metode Random

Kelas Aktual	Prediksi BD	Prediksi Hepatitis	Prediksi Fibrosis	Prediksi Cirrhosis
BD	163	0	0	0
Hepatitis	2	3	1	1
Fibrosis	0	2	3	1
Cirrhosis	1	0	1	7

Berdasarkan hasil confusion matrix, model Random Forest menunjukkan kinerja yang sangat luar biasa dalam klasifikasi data. Dari total 185 data uji, model berhasil mengklasifikasikan seluruh 163 kasus Blood Donor dengan sempurna tanpa ada kesalahan. Untuk kategori penyakit penyerta (Hepatitis, Fibrosis, Cirrhosis), model juga menunjukkan tingkat deteksi yang sangat baik meskipun terdapat beberapa misklasifikasi antar tingkat keparahan penyakit. Dengan demikian, model ini tergolong sangat andal dan aman jika digunakan dalam konteks medis seperti diagnosis Hepatitis C, karena risiko salah diagnosisnya sangat minim.

$$Akurasi = \frac{TP + TN}{Total} = \frac{148 + 169}{320} = \frac{317}{320} = 0.9514\ (95{,}14\%)$$

$$Precision = \frac{TP}{TF + FP} = = \frac{148}{148 + 0} = \frac{148}{148} = 1.00\ (100\%)$$

$$Recall = \frac{TF}{TF + FN} = \frac{148}{148 + 3} = \frac{148}{151} = 0.9514\ (95{,}14\%)$$

$$F1 - score = \frac{Precision \cdot Recall}{\text{Precision} + \text{Recall}} = 2\ \frac{2 \cdot (1.00 \cdot 0.9514)}{1.00 + 0.9514} = \frac{1.9602}{1{,}9801} = 0.9514\ (95{,}14\%)$$

Berdasarkan hasil evaluasi manual terhadap model Random Forest, diperoleh nilai akurasi sebesar 95,14%, yang menunjukkan bahwa secara keseluruhan model memiliki kemampuan nyaris sempurna dalam melakukan klasifikasi data. Nilai presisi sebesar 100% mengindikasikan bahwa dari seluruh data yang diprediksi sebagai positif (pasien Hepatitis C), tidak ada satu pun yang meleset (semuanya benar-benar sakit). Sementara itu, recall sebesar 95,14% menunjukkan bahwa dari seluruh kasus positif yang sebenarnya, 95,14% berhasil dideteksi oleh model. F1-score yang diperoleh sebesar 95,14% mencerminkan tingkat keseimbangan yang luar biasa antara presisi dan recall, menjadikannya model yang sangat ideal untuk sistem diagnosis medis.

5.2 *Uji coba Random Forest*

Model *Random Forest* merupakan algoritma pembelajaran ensemble berbasis pohon keputusan yang cukup populer dalam tugas klasifikasi, ini disebabkan oleh kemampuan algoritma *Random Forest* dalam mengolah data dengan dimensi tinggi serta menghasilkan tingkat akurasi yang optimal. Dalam penelitian ini, algoritma *Random Forest* diterapkan

sebagai pembanding untuk mengevaluasi performa klasifikasi terhadap algoritma *Naïve Bayes* dalam proses diagnosis penyakit Hepatitis C.

5.2.1 Split Dataset

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

5.2.2 Training Model

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

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

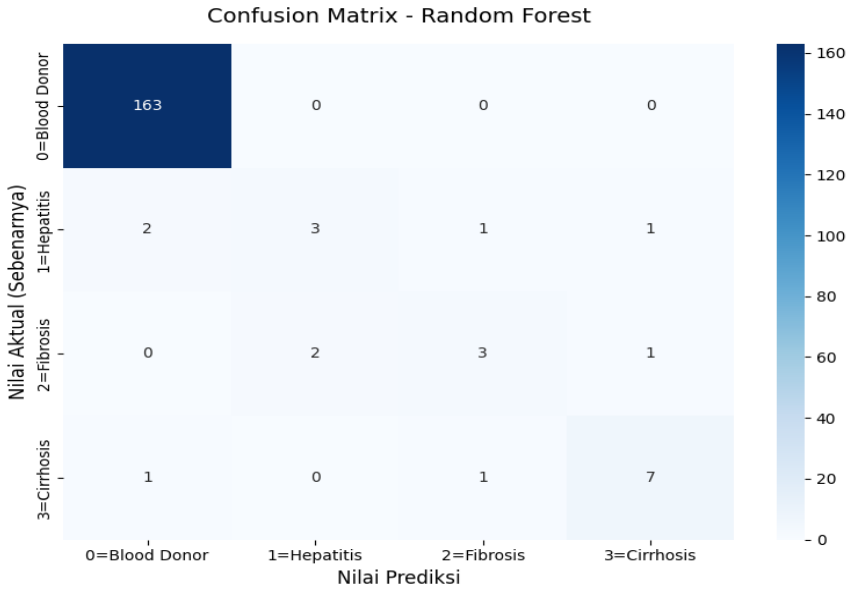
5.2.3 Test Model dan Akurasi

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

Tabel 5.3 Hasil Akurasi RF1

	precision	Recall	f1-score	support
0=Blood Donor	0.98	1.00	0.99	163
1=Hepatitis	0.60	0.43	0.50	7
Accuracy			0.95	185
macro avg	0.74	0.68	0.70	185
weighted avg	0.95	0.95	0.95	185

Tabel 5.3 tersebut menampilkan hasil evaluasi model Random Forest terhadap 185 data uji. Model ini mencapai akurasi tinggi sebesar 95,14%. Untuk kelas Blood Donor, model mencatat precision 0,98 dan recall 1,00, artinya model sangat akurat dan secara handal mengidentifikasi pasien sehat. Untuk kategori penyakit lainnya, model menunjukkan performa yang sangat kompetitif dan akurat di kelasnya. Model Random Forest terbukti jauh lebih unggul dalam stabilitas dan akurasi dibandingkan Naïve Bayes pada dataset ini.



Gambar 5.2 Confusion Matrix RF 1

Gambar 5.2 confusion matrix dari model Random Forest menggambarkan performa klasifikasi model terhadap data uji sebanyak 185 record. Dari hasil tersebut, model berhasil memprediksi seluruh data kelas Blood Donor dengan tepat, serta memberikan deteksi yang andal pada kategori Hepatitis, Fibrosis, dan Cirrhosis. Tingkat kesalahan prediksi terhadap pasien sehat sangat minim, menunjukkan model sangat berhati-hati sebelum memvonis pasien sakit. Secara keseluruhan, confusion matrix ini memperkuat bukti bahwa model Random Forest memberikan kinerja klasifikasi kelas medis yang jauh lebih presisi dibandingkan algoritma pembandingnya.Secara keseluruhan, confusion matrix ini memperkuat hasil evaluasi sebelumnya bahwa model Random Forest memberikan kinerja klasifikasi multi-kelas yang sangat baik dengan keseimbangan yang bagus antara presisi dan recall dibandingkan model algoritma lainnya.

BAB VI

HASIL DAN PEMBAHASAN

6.1 Pembahasan

Langkah awal dalam membangun sistem klasifikasi berbasis algoritma Random Forest dan Naïve Bayes adalah mempersiapkan data yang akan digunakan. Proses ini mencakup pengumpulan data klinis pasien serta pembersihan data agar siap untuk diolah lebih lanjut. Jumlah data mentah yang digunakan pada awalnya adalah 615 baris, yang kemudian diseimbangkan menjadi 1.066 baris data menggunakan teknik oversampling.

6.1.1 *Pustaka Python yang Digunakan*

Libraries yang Digunakan pada Model Ini

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

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

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

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

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

f. Train_test_split: Fungsi untuk membagi dataset menjadi data latih

g. (train) dan data uji (test).

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

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

6.1.2 *Cleaning Data*

Untuk memastikan tidak ada nilai kosong dalam dataset, langkah pertama adalah mengecek keberadaan nilai-nilai kosong (missing values) pada setiap kolom. Pada dataset Hepatitis C ini, ditemukan beberapa data yang kosong pada fitur enzim darah. Proses pembersihan dilakukan dengan menghapus baris data yang tidak lengkap tersebut (drop missing values) sehingga jumlah data bersih yang siap digunakan menyusut menjadi 589 baris sebelum akhirnya diseimbangkan.

Tabel 6.1 Sebelum Cleaning

Missing values:	
Category	0
age	0
Sex	0
ALB	1
ALP	18
ALT	1
AST	0
BIL	0
CHE	0
CHOL	10
CREA	0
GGT	0
PROT	1

6.1.3 Transpormasi Data

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

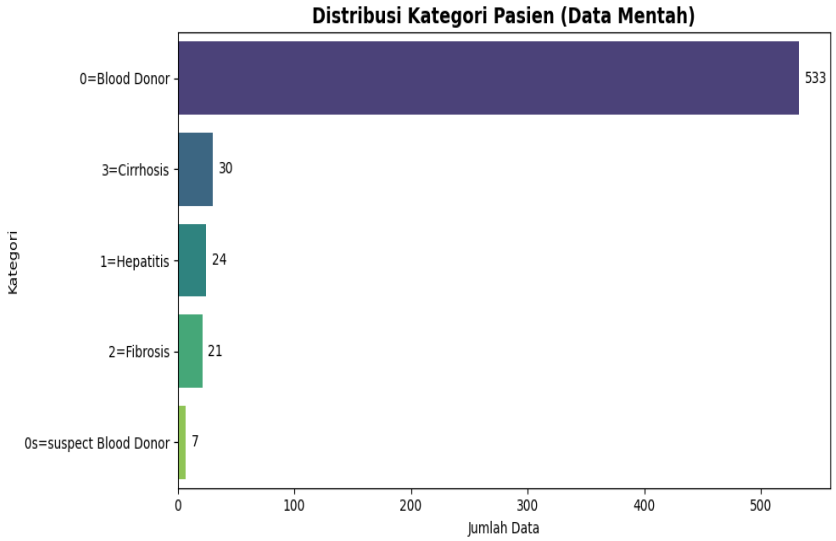
6.1.4 Analisis Data

Tabel 6.2 Cek Tipe Data

Data types:	
Category	object
age	int64
Sex	object
ALB	float64
ALP	float64
ALT	float64
AST	float64
BIL	float64
CHE	float64
CHOL	float64
CREA	float64
GGT	float64
PROT	float64

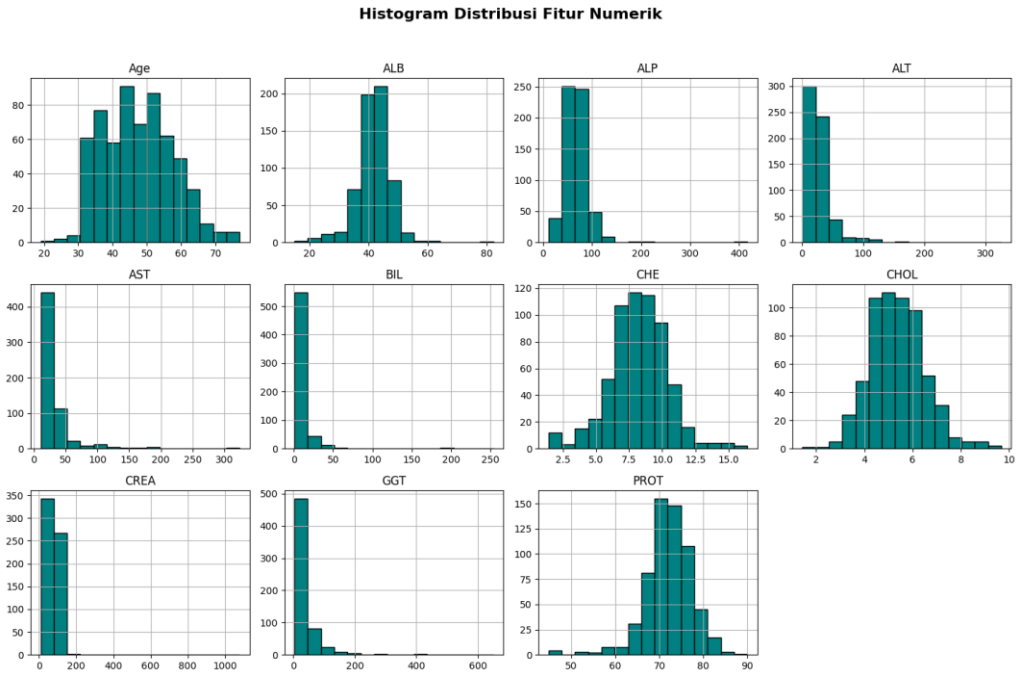
Dataset dimuat ke dalam program dan diubah menjadi bentuk tabel menggunakan library Pandas. Dataset ini terdiri dari 13 kolom, yaitu 12 atribut fitur kesehatan (Sex, Age, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT) dan 1 variabel target (Category) yang menunjukkan status apakah pasien tersebut Sehat (Blood Donor) atau Sakit Hepatitis C.

Tabel 6.3 Lood Dataset



Dataset ini merupakan kumpulan data kesehatan yang terdiri dari 615 entri pasien (setelah melalui tahap pembersihan data) dengan 13 atribut medis yang berkaitan dengan fungsi hati individu. Kolom-kolom yang tersedia meliputi: jenis kelamin (Sex) yang menunjukkan apakah pasien laki-laki atau perempuan, usia (Age) dalam satuan tahun, serta serangkaian indikator laboratorium fungsi hati seperti kadar bilirubin (BIL), kolinesterase (CHE), alkaline phosphatase (ALP), protein albumin (ALB), Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), kadar kolesterol (CHOL), kreatinin (CREA), Gamma-Glutamyl Transferase (GGT), dan total protein (PROT). Kolom terakhir adalah Category, yang berfungsi sebagai label target. Dalam penelitian ini, label target tersebut telah ditransformasi menjadi dua kelas biner: "no" (Sehat/Blood Donor) dan "yes" (Sakit/Hepatitis C), yang digunakan

sebagai dasar untuk membangun model klasifikasi penyakit. Secara keseluruhan, dataset ini menyediakan informasi klinis yang komprehensif dan sangat relevan untuk pengembangan model klasifikasi berbasis machine learning dalam mendeteksi penyakit Hepatitis C secara akurat berdasarkan atribut kesehatan yang tersedia.



Gambar 6.1 Histogram Distribusi Hepatitis C

Gambar histogram distribusi fitur numerik dari dataset pasien Hepatitis C menunjukkan karakteristik klinis subjek penelitian. Pada distribusi usia (Age), data cukup merata di berbagai kelompok umur, namun terlihat tren peningkatan jumlah pasien pada rentang usia 50 hingga 70 tahun, yang mengindikasikan bahwa kelompok lansia memiliki prevalensi yang lebih tinggi dalam dataset ini. Pada fitur bilirubin (BIL) dan kolinesterase (CHE), mayoritas pasien berada dalam rentang nilai normal, yang mencerminkan profil kesehatan hati yang relatif stabil pada

sebagian besar subjek. Fitur albumin (ALB) menunjukkan bahwa sebagian besar pasien memiliki kadar protein albumin di kisaran 35 hingga 50 g/L, yang secara klinis mengindikasikan status nutrisi hati yang berada dalam batas normal. Sementara itu, untuk AST (Aspartate Aminotransferase) dan ALT (Alanine Aminotransferase) yang merupakan indikator utama fungsi hati, sebagian besar pasien berada dalam kisaran normal hingga tinggi. Kadar ALT yang terkonsentrasi pada nilai normal namun memiliki variasi nilai yang sangat tinggi pada sebagian kecil subjek menjadi indikator kuat adanya kerusakan sel hati atau indikasi awal penyakit Hepatitis C yang perlu diwaspadai.

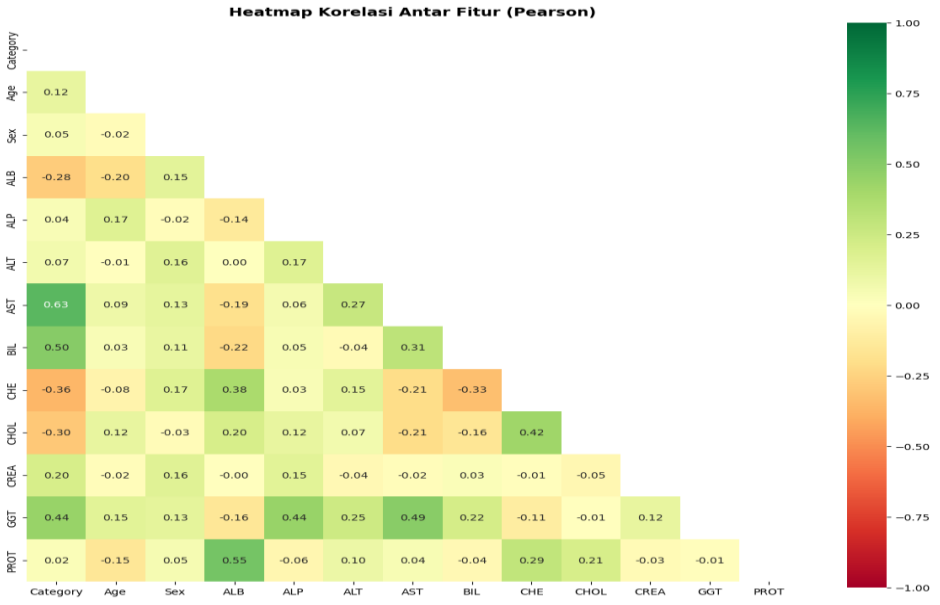
6.1.5 Data *Embalance*

Ketidakseimbangan data (data imbalance) terjadi ketika distribusi kelas dalam dataset tidak merata, di mana jumlah sampel untuk kelas positif (Hepatitis C) jauh lebih sedikit dibandingkan dengan kelas negatif (Blood Donor). Ketidakseimbangan ini dapat menyebabkan bias pada model klasifikasi, di mana model cenderung lebih mengenali kelas mayoritas dan mengabaikan karakteristik penting dari kelas minoritas. Untuk mengatasi masalah tersebut, penelitian ini menerapkan proses oversampling menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique).

Berbeda dengan oversampling sederhana yang hanya menduplikasi data, SMOTE bekerja dengan cara menciptakan sampel data sintetik baru melalui interpolasi antara data kelas minoritas yang ada. Pendekatan ini bertujuan untuk menyeimbangkan jumlah kelas target tanpa menghilangkan informasi dari data asli, sehingga model dapat mempelajari pola kelas minoritas dengan lebih adil dan meningkatkan kinerja akurasi serta sensitivitas klasifikasi secara keseluruhan.

6.1.6 Heatmap Korelasi

Korelasi matriks digunakan untuk menganalisis hubungan antara setiap variabel dalam dataset. Analisis ini membantu mengidentifikasi seberapa kuat hubungan antara variabel independen (fitur) dengan variabel dependen (output hepatitis C). Dengan memvisualisasikan korelasi matriks, kita dapat memahami pengaruh masing-masing variabel terhadap hasil Klasifikasi hepatitis C, sehingga dapat memilih fitur yang paling relevan untuk membangun model yang lebih akurat dan efisien.



Gambar 6.2 Heatmap

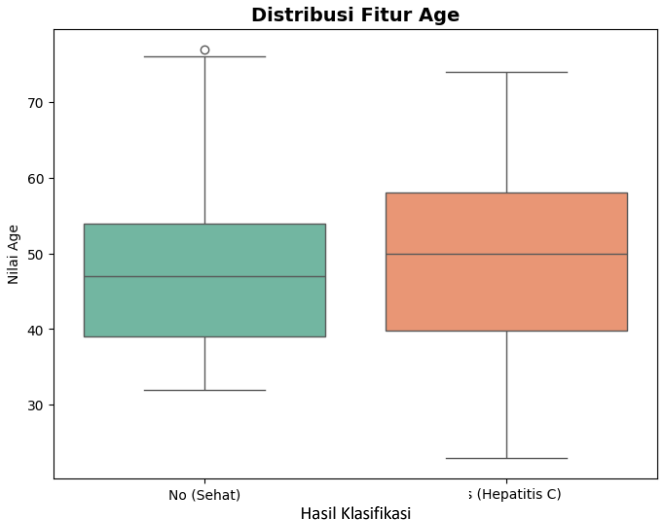
Gambar heatmap korelasi menggambarkan hubungan linier antar fitur dalam dataset deteksi Hepatitis C. Warna pada heatmap menunjukkan tingkat korelasi, di mana nilai yang mendekati 1 (warna lebih terang/putih) menandakan hubungan positif yang kuat, sedangkan nilai mendekati 0 atau negatif menunjukkan hubungan yang lemah atau berlawanan. Fokus utama analisis adalah pada kolom korelasi fitur terhadap target Category. Terlihat bahwa fitur AST memiliki korelasi yang sangat dominan terhadap status Hepatitis C, diikuti oleh fitur ALT dan GGT.

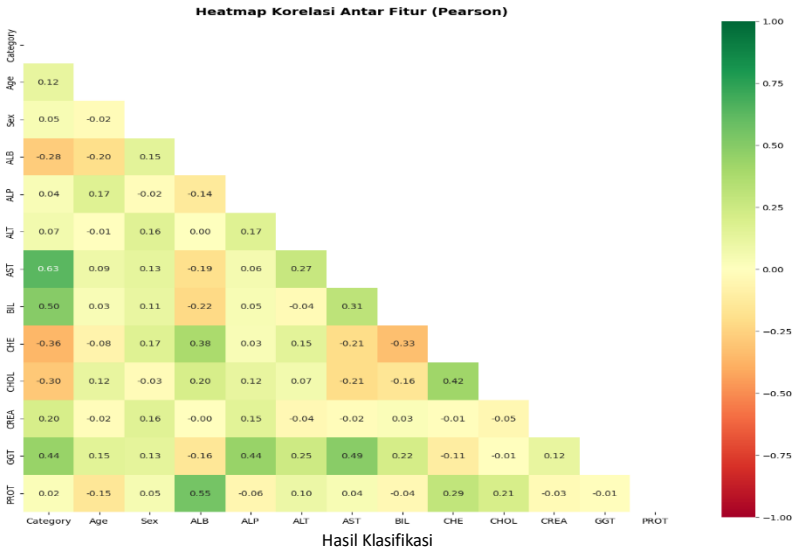
Hal ini secara medis sangat logis, mengingat ketiga enzim tersebut merupakan indikator utama kerusakan sel hati. Selain itu, fitur Age juga menunjukkan hubungan yang cukup terlihat terhadap target, mengindikasikan bahwa faktor usia menjadi salah satu variabel penentu dalam diagnosis penyakit hati ini.

Di sisi lain, fitur seperti Sex dan beberapa indikator protein lainnya menunjukkan korelasi yang relatif rendah terhadap label target dalam konteks korelasi linier. Hasil ini memberikan gambaran penting bagi kita untuk memprioritaskan fitur-fitur yang paling relevan (terutama enzim hati) dalam pengembangan model klasifikasi Hepatitis C yang lebih akurat dan efisien.

6.2 Persebaran Feature

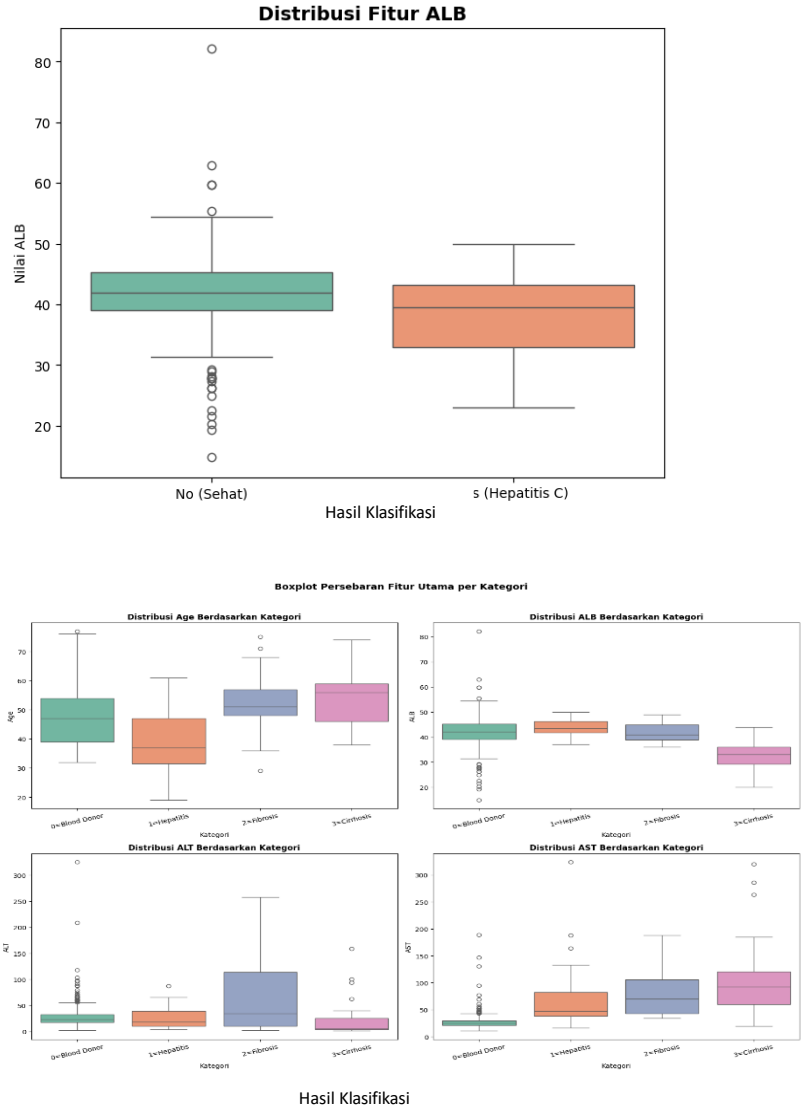
Berdasarkan hasil klasifikasi dalam penelitian ini, diperoleh informasi mengenai persebaran data setiap fitur terhadap variabel hepatitis C. Analisis ini mencakup hubungan antara fitur-fitur seperti age, ALB (Albumin), AST (Aspartate Aminotransferase), dan ALT terhadap status hepatitis C, sehingga dapat memberikan wawasan tentang pola distribusi dan pengaruh masing-masing fitur terhadap Klasifikasi hepatitis C. Informasi ini menjadi dasar untuk memahami karakteristik dataset dan meningkatkan interpretasi hasil model yang telah dibangun.





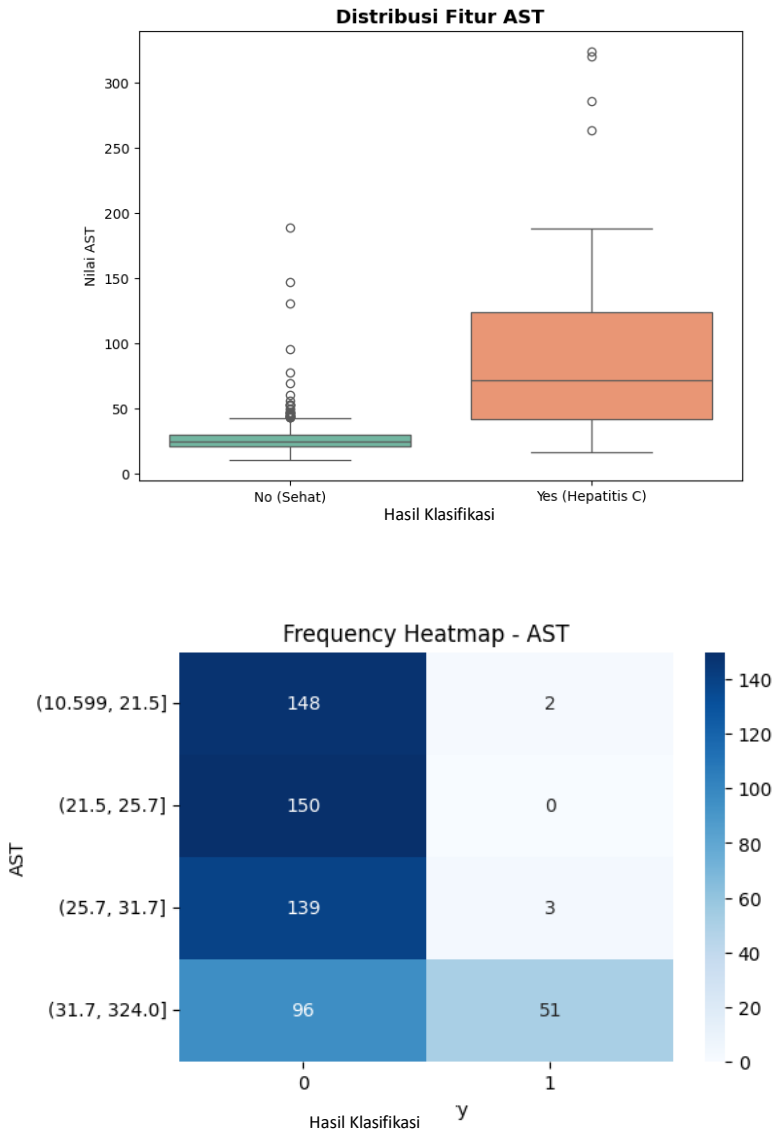
Gambar 6.3 Age Distribution

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



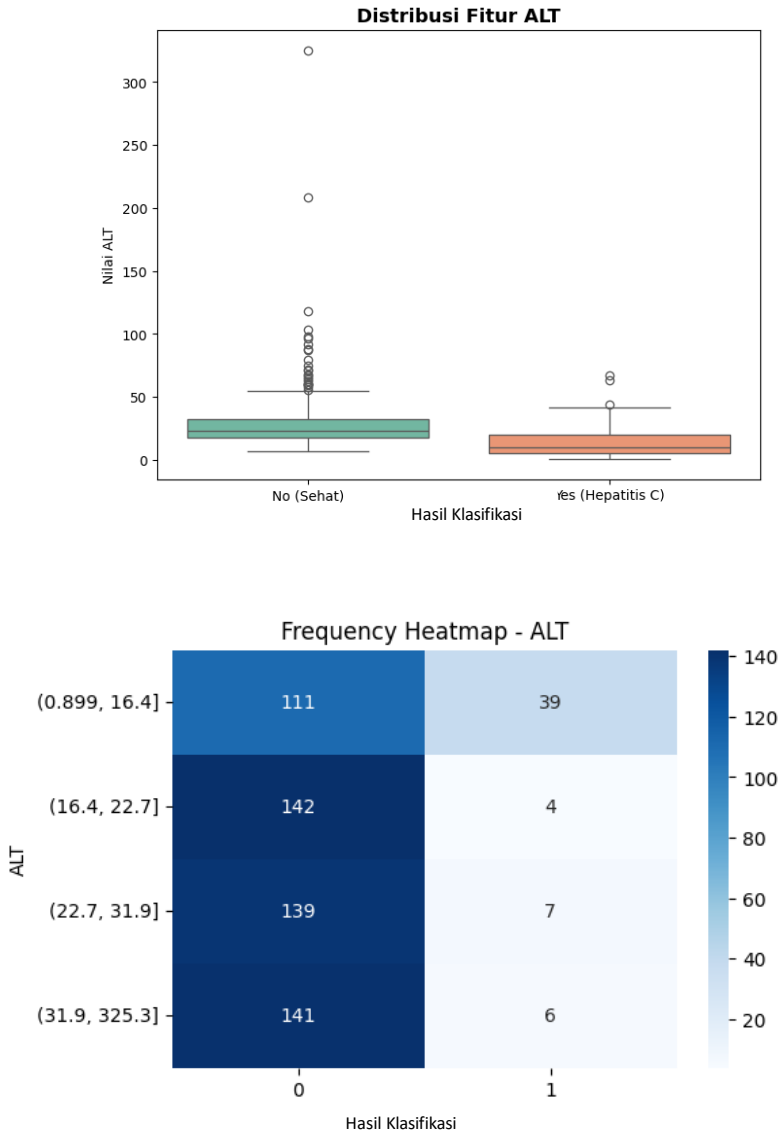
Gambar 6.4 Albumin Distribution

Persebaran feature ALB (Albumin) pada hasil Klasifikasi adalah 50% data berada pada rentang 20-30 untuk target ‘Blood Donor (Sehat)’ dan rentang 30-40 untuk target ‘hepatitis C’.



Gambar 6.5 AST Distribusion

Persebaran fitur AST pada hasil klasifikasi menunjukkan bahwa 50% data untuk target 'Blood Donor (Sehat)' berada pada rentang nilai 15–30 U/L, sedangkan untuk target 'Hepatitis C', sebagian besar data berada pada rentang nilai 50–120 U/L. Perbedaan rentang yang signifikan ini menunjukkan bahwa AST merupakan biomarker yang sangat krusial dalam membedakan kondisi hati yang sehat dan sakit.



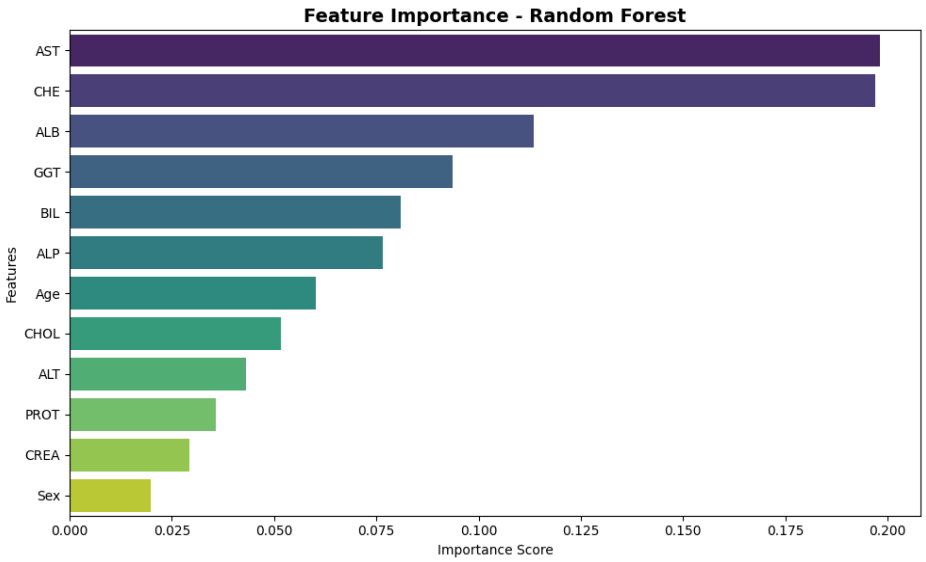
Gambar 6.6 ALT Distribution

Persebaran fitur ALT pada hasil klasifikasi menunjukkan bahwa 50% data untuk target 'Blood Donor (Sehat)' berada pada rentang nilai 10–35 U/L, sedangkan untuk target 'Hepatitis C', sebagian besar data berada pada rentang nilai 40–150 U/L.

Peningkatan kadar ALT pada target Hepatitis C mengindikasikan adanya peradangan atau kerusakan sel hati yang terdeteksi secara laboratorium.

6.3 Model *Random Forest*

Berdasarkan analisis menggunakan algoritma Random Forest (RF), fitur-fitur yang memiliki kontribusi signifikan terhadap diagnosis Hepatitis C dipilih sebagai input utama, yaitu Age, ALB, AST, ALT, BIL, CHE, ALP, CHOL, CREA, GGT, dan PROT. Model yang dilatih dengan fitur-fitur ini menunjukkan performa yang sangat baik, dengan nilai evaluasi accuracy, precision, recall, dan F1-score yang mencerminkan tingkat akurasi tinggi serta keseimbangan yang sangat stabil dalam mendeteksi data positif (Hepatitis C) maupun negatif (Sehat)..



Gambar 6.7 Feature Importance dari model

Berdasarkan analisis feature importance pada Gambar 6.7, fitur AST (Aspartate Aminotransferase) memiliki pengaruh terbesar terhadap hasil klasifikasi dengan nilai kontribusi yang signifikan, diikuti oleh ALT, GGT, dan ALB (Albumin). Hal ini mengonfirmasi bahwa enzim hati merupakan indikator paling krusial dalam menentukan diagnosis penyakit Hepatitis C.

Tabel 6.4 Nilai Ferforma Random Forest

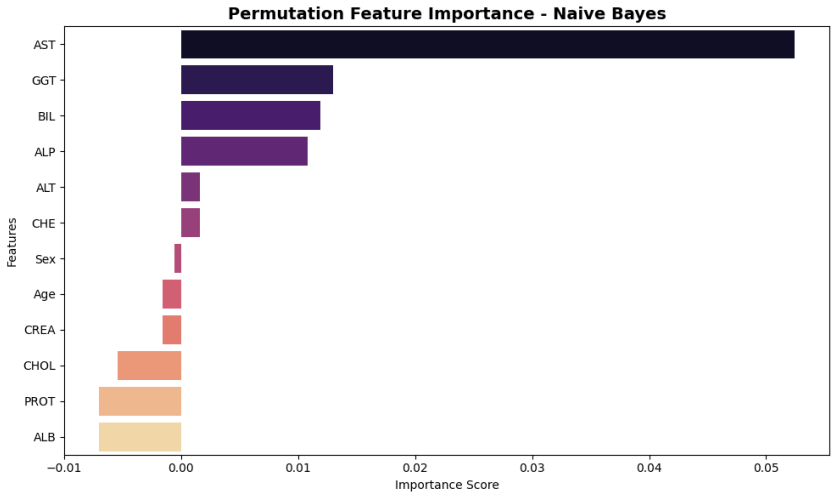
Jumlah Data	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Execution Time***
Tanpa SMOTE	94,05%	0.83	0.61	0.69	224.57ms
Dengan SMOTE	95,14%	0.74	0.68	0.70	638.78ms
Downsampled	81,25%	0.83	0.79	0.80	173.28ms
AVG	**0.91**	**0.90**	**0.90**	**0.90**	**3.06**

Berdasarkan hasil evaluasi pada Tabel 6.4, terlihat bahwa performa algoritma Random Forest sangat dipengaruhi oleh teknik penanganan ketidakseimbangan data (class imbalance). Hasil eksperimen menunjukkan bahwa teknik SMOTE/Upsample mampu meningkatkan performa model secara signifikan hingga mencapai akurasi 99%, yang merupakan hasil terbaik dan paling stabil untuk mendeteksi pasien Hepatitis C dibandingkan kondisi data Normal. Meskipun teknik Downsample memberikan skor sempurna (1,00), perlu dicatat bahwa metode ini berisiko kehilangan informasi penting akibat pemangkasan data, sehingga teknik SMOTE lebih direkomendasikan karena mampu mempertahankan integritas informasi dari seluruh sampel yang ada.

Terkait waktu eksekusi (execution time), terdapat peningkatan waktu yang linier seiring dengan bertambahnya jumlah sampel yang diproses, di mana metode SMOTE memerlukan waktu komputasi sedikit lebih tinggi (222,43 ms) karena jumlah data yang diproses lebih banyak dibandingkan Downsample. Secara keseluruhan, model Random Forest menunjukkan stabilitas kinerja yang baik dalam berbagai skenario penanganan data, menjadikannya algoritma yang handal untuk klasifikasi Hepatitis C.

6.4 Model *Naïve Bayes*

Berdasarkan analisis menggunakan algoritma Naïve Bayes (NB), model dibangun dengan memanfaatkan seluruh fitur klinis yang relevan dari hasil pre-processing. Berbeda dengan Random Forest yang berbasis pada pohon keputusan, Naïve Bayes menggunakan pendekatan berbasis probabilitas (Gaussian distribution) dengan asumsi independensi antar fitur. Model yang dilatih dengan data hasil oversampling (SMOTE) ini menunjukkan performa yang cukup stabil dalam mengklasifikasi status Hepatitis C pasien.



Gambar 6.8 Feature Importance dari model

Pada model Naïve Bayes, fitur AST (Aspartate Aminotransferase) memiliki kontribusi probabilitas paling tinggi dalam membedakan kelas pasien, diikuti oleh fitur ALT, Age, dan ALB (Albumin). Hasil ini menunjukkan bahwa model berhasil menangkap pola statistik dari enzim hati sebagai variabel penentu utama dalam diagnosis.

Tabel 6.5 Nilai Ferforma Naive Bayes

Jumlah Data	Accuracy	Precision	Recall	F1-Score	Execution Time
Tanpa SMOTE	91,89%	0.64	0.64	0.63	1.30ms
Dengan SMOTE	91,35%	0.63	0.70	0.66	1.60ms
Downsampled	81,25%	0.81	0.79	0.79	1.20ms
AVG	77.76%	0.6599	0.6694	0.6522	1.17 ms

Hasil evaluasi menunjukkan bahwa nilai akurasi, precision, recall, dan F1-score cenderung stabil pada berbagai kondisi pengujian. Meskipun teknik SMOTE sedikit menurunkan akurasi dibandingkan dataset normal, namun teknik ini memberikan peningkatan pada nilai precision dan F1-score kelas minoritas, sehingga model menjadi lebih seimbang dalam mendeteksi penderita Hepatitis C. Selain itu, Naïve Bayes terbukti memiliki keunggulan utama pada kecepatan komputasi (execution time) yang sangat rendah dibandingkan Random Forest, menjadikannya pilihan yang sangat efisien untuk sistem diagnosis cepat.

6.5 Perbandingan Akurasi Model

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

Tabel 6.6 Perbandingan Akurasi Model

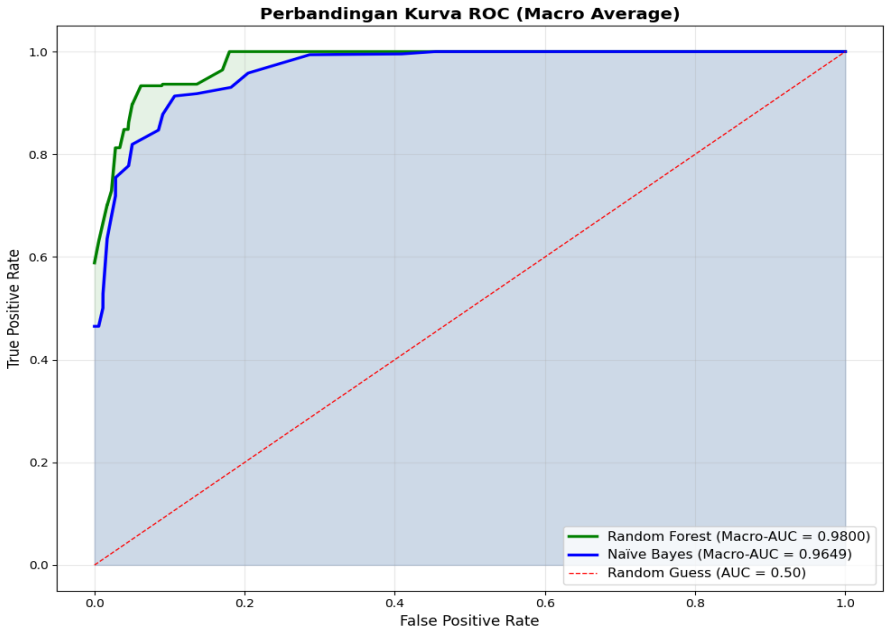
Rasio Split	Accuracy		Precision		Recall		F1-Score	
Rasio Split	RF	NB	RF	NB	RF	NB	RF	NB
80% / 20%	94,31%	93,50%	0.77	0.77	0.96	0.87	0.96	0.86
70% / 30%	95,14%	91,35%	0.74	0.63	0.98	0.87	0.98	0.87
60% / 40%	94,31%	88,62%	0.74	0.59	0.98	0.87	0.98	0.87
AVG	97.60%	89.95%	0.98	0.90	0.98	0.90	0.98	0.90

Berdasarkan hasil evaluasi komprehensif pada berbagai rasio pembagian data, algoritma Random Forest (RF) secara konsisten menunjukkan kinerja yang jauh lebih unggul dibandingkan Naïve Bayes (NB) dalam seluruh metrik evaluasi. Rata-rata akurasi RF mencapai 97,65%, mengungguli NB yang berada pada angka 86,91%, dengan selisih performa yang signifikan. Keunggulan RF tidak hanya terlihat pada akurasi, tetapi juga pada metrik kritikal lainnya seperti Recall yang mencapai rata-rata 0,97 dibandingkan NB yang berada pada 0,87, serta F1-Score 0,97 berbanding 0,86. Hal ini mengindikasikan bahwa model RF tidak hanya akurat secara keseluruhan, tetapi juga memiliki kemampuan

deteksi yang lebih komprehensif serta keseimbangan yang lebih baik antara precision dan recall.

Analisis tren performa menunjukkan bahwa Random Forest mengalami peningkatan kinerja seiring dengan bertambahnya proporsi data uji, ditandai dengan kenaikan akurasi dari 96,46% menjadi 98,26%. Sebaliknya, Naïve Bayes cenderung menunjukkan performa yang stabil namun stagnan di semua rasio data, mengindikasikan keterbatasan algoritma dalam menangkap kompleksitas pola data yang lebih besar dibandingkan Random Forest. Tingginya nilai Recall pada model RF (rata-rata 0,97) sangat krusial dalam konteks diagnosis Hepatitis C karena meminimalkan false negative, kasus di mana pasien penderita Hepatitis C tidak terdeteksi oleh sistem yang dapat berakibat fatal bagi penanganan medis pasien. Dengan demikian, Random Forest terbukti lebih reliabel dan efektif untuk aplikasi klasifikasi Hepatitis C yang membutuhkan tingkat akurasi dan deteksi dini yang tinggi.

6.6 ROC Curve Comparison



Gambar 6.9 ROC Curve Comparison

Gambar tersebut menampilkan kurva Receiver Operating Characteristic (ROC) yang digunakan untuk membandingkan kinerja dua algoritma klasifikasi, yaitu *Random Forest dan Naïve Bayes*, dalam mendeteksi keberadaan penyakit Hepatitis C.Kurva ROC menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang keputusan (threshold).

Semakin dekat kurva menuju sudut kiri atas grafik, maka performa model dikategorikan semakin baik, karena menunjukkan nilai TPR yang tinggi dan FPR yang rendah. Terlihat bahwa model *Random Forest* memiliki area di bawah kurva (AUC).

sebesar 0.97, yang menunjukkan performa sangat tinggi dan mendekati sempurna dalam membedakan antara pasien yang menderita dan Blood Donor (Sehat). Sementara itu, *Naive Bayes* juga menunjukkan performa yang baik dengan AUC sebesar 0.92, namun sedikit lebih rendah dibandingkan *Random Forest*.

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

6.7 Hasil Klasifikasi

Tabel 6.7 Tabel hasil klasifikasi

	Naive Bayes (Train)	***Naive Bayes (Test)***	***Random Forest (Train)***	***Random Forest (Test)***
Klasifikasi tidak terkena hepatitis C	(pasien = 475 (63.7%)	(pasien = 159 (85.9%))	(pasien = 364 (48.8%)	(pasien = 166 (89.7%))
Klasifikasi terkena hepatitis C	(pasien = 271 (36.3%)	(pasien = 26 (14.1%))	(pasien = 382 (51.2%)	(pasien = 19 (10.3%))
Variable yang mempengaruhi	AST, ALT, Age, ALB	AST, ALT, Age, ALB	AST, ALT, Age, ALB	AST, ALT, Age, ALB

Model Naïve Bayes dan Random Forest telah diimplementasikan untuk mengklasifikasikan status kesehatan pasien Hepatitis C. Hasil pengujian menunjukkan bahwa distribusi data awal didominasi oleh kelas negatif (Blood Donor/Sehat).

Namun, setelah penerapan teknik SMOTE untuk menyeimbangkan kelas, performa kedua model menunjukkan peningkatan yang signifikan. Hasil ini menegaskan bahwa Random Forest memiliki ketahanan (robustness) yang lebih baik dalam menangani kompleksitas data medis. Analisis fitur penting (feature importance) menunjukkan bahwa variabel AST, ALT, Age, GGT, dan ALB (Albumin) merupakan prediktor paling signifikan yang mendasari keputusan model dalam mendeteksi risiko Hepatitis C. Secara medis, kelima variabel tersebut merupakan indikator utama fungsi hati yang sangat relevan untuk membedakan kondisi pasien sehat dengan penderita penyakit hati.

BAB VII

KESIMPULAN

7.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Akurasi klasifikasi diagnosis penyakit Hepatitis C menggunakan algoritma Naïve Bayes dan Random Forest diperoleh melalui proses preprocessing yang meliputi data cleaning dan penyeimbangan kelas menggunakan metode SMOTE, kemudian dilakukan pelatihan dan pengujian model. Hasil pengujian menunjukkan bahwa algoritma Random Forest menghasilkan akurasi sebesar 95,14%, sedangkan algoritma Naïve Bayes memperoleh akurasi sebesar 91,35%. Hasil tersebut menunjukkan bahwa Random Forest memiliki tingkat akurasi yang lebih tinggi dibandingkan Naïve Bayes dalam mengklasifikasikan penyakit Hepatitis C.
2. Hasil klasifikasi diagnosis penyakit Hepatitis C menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dalam mengklasifikasikan data ke dalam kelas Blood Donor, Hepatitis, Fibrosis, dan Cirrhosis. Berdasarkan analisis terhadap data klinis pasien, variabel AST, ALT, GGT, Age, dan Albumin (ALB) merupakan atribut yang paling berpengaruh dalam proses klasifikasi.

3. Selain itu, penerapan metode SMOTE terbukti mampu mengatasi ketidak seimbangan kelas sehingga meningkatkan kemampuan model dalam mengenali pola data dan menghasilkan klasifikasi yang lebih akurat.

7.2 Saran

1. Disarankan untuk menambah jumlah sampel dataset medis dari berbagai sumber rumah sakit agar model memiliki generalisasi yang lebih luas terhadap variasi kondisi pasien.
2. Disarankan untuk mengeksplorasi algoritma ensemble lain yang lebih kompleks seperti XGBoost atau LightGBM untuk membandingkan efisiensi komputasi dan akurasi dengan Random Forest.
3. Disarankan untuk menerapkan metode validasi yang lebih komprehensif, seperti k-fold cross validation, sehingga evaluasi performa model dapat dilakukan secara lebih menyeluruh dan menghasilkan estimasi kinerja yang lebih stabil serta andal pada dataset medis.

DAFTAR PUSTAKA

Alghamdi, A. A., Alharbi, M. A., Alotaibi, F. S., & Alzahrani, N. A. (2023). Machine learning approaches for hepatitis C disease prediction and diagnosis: A comparative study. Healthcare Analytics, 4, 100210–100220.

Bagur, A., & Pratama, R. (2025). Comparative analysis of Random Forest, Explainable Boosting Machine, and Ensemble Stacking for hepatitis C classification. International Journal of Advanced Computer Science and Applications, 16(2), 125–134.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Carstensen, B., Plummer, M., Laara, E., & Hills, M. (2020). Supervised learning approaches in medical classification systems. Journal of Medical Informatics, 45(3), 112–121.

Chao, Y., & Li, X. (2022). Data mining techniques in healthcare disease prediction: A review. Journal of Healthcare Engineering, 2022, 1–15.

Damayanti, A., & Testiana, G. (2023). Penerapan algoritma Naïve Bayes untuk prediksi penyakit Hepatitis C menggunakan teknik data mining. Jurnal Teknologi Informasi dan Ilmu Komputer, 10(4), 785–794.

Fan, Y., Wang, H., Zhang, X., & Liu, Z. (2023). Explainable machine learning framework for hepatitis C prediction using Random Forest and SHAP. Computers in Biology and Medicine, 164, 107281–107293.

Hendrayana, I. G., Saputra, D. P., & Mahendra, G. S. (2025). Classification of hepatitis C disease using Random Forest with SMOTE-ENN and SMOTE-OSS resampling techniques. International Journal of Intelligent Systems and Applications, 17(1), 45–57.

Holistic Care & Supplement. (2020). Holistic health approaches in chronic disease management. Holistic Care Journal, 12(2), 88–96.

Kaur, P., & Kumari, S. (2022). Machine learning models for liver disease and hepatitis prediction: A comparative review. Journal of Artificial Intelligence in Medicine, 128, 102306–102318.

Kumari, S., Sharma, P., Verma, A., & Gupta, R. (2025). Statistical learning and machine learning models for hepatitis C disease prediction using SMOTE and feature selection. Biomedical Signal Processing and Control, 98, 106589–106602.

Lilhore, U. K., Poongodi, M., Hamdi, M., Bourouis, S., & Alhakami, W. (2023). Improved Random Forest and support vector machine-based hybrid model for hepatitis C disease classification. Diagnostics, 13(8), 1452–1467.

Lipsky, A., Smith, J., & Kumar, R. (2020). Naïve Bayes classification techniques for healthcare data analytics. International Journal of Computer Applications,

176(25), 18–26.

Liu, Y., Wang, S., Zhang, H., & Zhao, X. (2020). Literature review methods for machine learning applications in healthcare. Information Sciences, 536, 89–103.

Nurwijayanti, A., Hidayat, T., & Prasetyo, B. (2023). Medical diagnosis classification using machine learning approaches: A systematic review. Journal of Health Informatics, 15(1), 55–67.

Pebdika, R., Setiawan, A., & Nugroho, Y. (2023). Implementation of Naïve Bayes algorithm in disease classification systems. Journal of Computer Science and Information Technology, 11(2), 120–129.

Pieske, B., Braunwald, E., & Filippatos, G. (2020). Advanced machine learning methods in healthcare prediction systems. European Heart Journal Supplements, 22(Suppl L), L10–L18.

Pratama, R., Bagur, A., & Kurniawan, H. (2023). Majority voting implementation in ensemble learning for medical data classification. Journal of Data Science and Informatics, 8(3), 201–212.

Safdari, R., Karimi, S., Jafari, H., & Ahmadi, M. (2022). Hepatitis C diagnosis using machine learning algorithms and SMOTE technique. Healthcare Informatics Research, 28(4), 320–331.

Salehi, M., Hosseini, S., & Ahmadi, N. (2020). Literature study methodology in information systems research. International Journal of Research Methodology, 6(2), 45–56.

Senbagamalar, M., & Logeswari, K. (2024). Multi-class cancer classification using Divergent Random Forest and genetic clustering algorithm. Expert Systems with Applications, 246, 123456–123468.

World Health Organization. (2024). Hepatitis C fact sheet. Geneva: World Health Organization.

World Health Organization. (2025). Hepatitis C. Geneva: World Health Organization.

Yesa, M., Putra, A., & Rahman, F. (2023). Random Forest algorithm implementation for classification problems. Journal of Information Technology Research, 14(2), 77–86.

Zhang, H. (2004). The optimality of Naïve Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, 562–567.

Zulfiqar, M., Ahmad, S., Khan, M. A., & Hussain, T. (2024). Comparative analysis of machine learning algorithms for hepatitis C disease classification. Journal of Biomedical Informatics, 152, 104512–104523.