PEMODELAN DETEKSI DINI GEJALA PENYAKIT SIROSIS HATI MENGGUNAKAN MACHINE LEARNING DENGAN PENDEKATAN SUPERVISED LEARNING

Rizza Muhammad Arief [a,1,*], Divira Salsabiil Susanto [b,2]

[a,b] Sistem Informasi, Universitas Merdeka Malang, Jalan Terusan Dieng. 62-64 Klojen, Pisang Candi, Sukun, Kota Malang
[1]rizza@unmer.ac.id; [2]20083000178@student.unmer.ac.id;

* Penulis Korespondensi

Diterima: 19 Mei 2024 | Direvisi : 20 Juli 2024 | Diterbitkan : 07 Agustus 2024

ABSTRAK

Sirosis hati merupakan konsekuensi serius dari hepatitis kronis yang dapat mengakibatkan komplikasi fatal. Deteksi dini sirosis hati sangat penting untuk meningkatkan prognosis dan mengurangi risiko komplikasi. Namun, gejalanya seringkali tidak spesifik, menyulitkan diagnosis pada tahap awal. Penelitian ini menggunakan dataset dari Mayo Clinic untuk menganalisis sirosis hati dengan tiga model machine learning: K-Nearest Neighbors (KNN), Naive Bayes, dan Support Vector Machine (SVM). Hasilnya menunjukkan bahwa model KNN memiliki akurasi tertinggi (92.04%), menunjukkan kemampuan yang efektif dalam mengklasifikasikan sirosis hati. Berdasarkan confusion matrix, KNN mampu mengklasifikasikan dengan tepat pasien yang menderita sirosis hati, dengan sedikit kesalahan dalam mengidentifikasi kelas yang berbeda. Sebagai perbandingan, model Naive Bayes menunjukkan performa yang lebih rendah dengan akurasi 52.14%, sementara SVM memiliki akurasi sebesar 81.88%. Dalam konteks deteksi dini sirosis hati, model KNN menonjol sebagai pilihan terbaik karena akurasinya yang tinggi dan kemampuannya dalam mengklasifikasikan pasien dengan benar. Langkah-langkah preprocessing data, seperti normalisasi dan one-hot encoding, juga berperan penting dalam meningkatkan kinerja model. Penemuan ini memberikan landasan penting untuk pengembangan sistem deteksi dini yang lebih baik untuk sirosis hati, sehingga memungkinkan intervensi medis yang tepat waktu dan peningkatan prognosis pasien.

KATA KUNCI

Sirosis Hati
Machine Learning
KNN
Naïve Bayes
Support Vector Machine
Supervised Learning

ABSTRACT

Cirrhosis of the liver is a serious consequence of chronic hepatitis that can lead to fatal complications. Early detection of liver cirrhosis is crucial to improving prognosis and reducing the risk of complications. However, its symptoms are often nonspecific, making diagnosis difficult at an early stage. This study utilizes a dataset from the Mayo Clinic to analyze liver cirrhosis using three machine learning models: K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM). The results indicate that the KNN model has the highest accuracy (92.04%), demonstrating effective capability in classifying liver cirrhosis. Based on the confusion matrix, KNN accurately classifies patients with liver cirrhosis, with few errors in identifying different classes. In comparison, the Naive Bayes model shows lower performance with an accuracy of 52.14%, while SVM has an accuracy of 81.88%. In the context of early detection of liver cirrhosis, the KNN model stands out as the best choice due to its high accuracy and ability to correctly classify patients. Preprocessing steps such as normalization and one-hot encoding also play a crucial role in improving model performance. These findings provide an important foundation for the development of better early detection systems for liver cirrhosis, enabling timely medical interventions and improving patient prognosis.

KEYWORD

Liver cirrhosis
Machine Learning
KNN
Naïve Bayes
Support Vector Machine
Supervised Learning

Jurnal Restikom : Riset Teknik Informatika dan Komputer
Vol. 6, No. 2, Agustus 2024, pp. 223-239

1. Pendahuluan

Setiap tanggal 28 Juli diperingati sebagai Hari Hepatitis Sedunia, dimana upaya kampanye dilakukan untuk meningkatkan kesadaran akan bahaya yang dihadapi oleh masyarakat terkait lima jenis virus hepatitis: tipe A, B, C, D, dan E. Tantangan besar yang dihadapi adalah kesulitan dalam mendeteksi penyakit ini, karena biasanya pasien baru teridentifikasi saat kondisinya sudah mencapai tahap lanjut. Kondisi ini menyoroti pentingnya deteksi dini penyakit hepatitis serta komplikasi serius yang dapat muncul, seperti sirosis hati. Sirosis hati merupakan salah satu konsekuensi paling serius dari hepatitis kronis, yang ditandai dengan kerusakan parah pada jaringan hati dan pembentukan jaringan parut. Sebagai akibatnya, fungsi hati menjadi terganggu secara bertahap, meningkatkan risiko terjadinya komplikasi seperti hipertensi portal, perdarahan varises esofagus, ensefalopati hepatik, dan karsinoma hepatoseluler[1], [2], [3], [4].

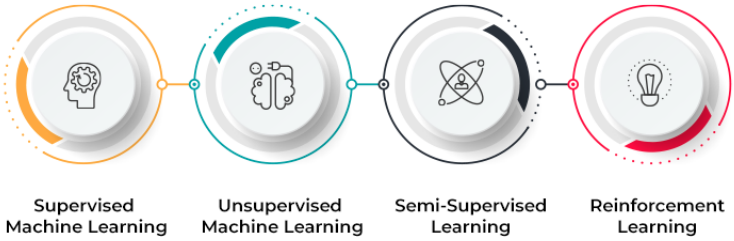
Gejala awal sirosis hati seringkali tidak spesifik dan dapat meliputi kelelahan, penurunan berat badan, perubahan pada kulit dan mata, serta perdarahan gusi. Namun, deteksi dini gejala penyakit ini sangat penting untuk meningkatkan prognosis dan mengurangi risiko komplikasi serius. Dengan mengidentifikasi sirosis hati pada tahap awal, intervensi medis dapat dilakukan lebih awal untuk memperlambat perkembangan penyakit, mengurangi risiko komplikasi, dan meningkatkan kualitas hidup pasien. Pentingnya deteksi dini sirosis hati diakui secara luas, masih terdapat tantangan dalam menerapkannya secara efektif dalam praktik klinis. Salah satu tantangan utama adalah kurangnya alat diagnostik yang sensitif dan spesifik untuk mendeteksi sirosis hati pada tahap awal dengan bantuan dataset yang kami peroleh dari sumber *Mayo Clinic study on primary biliary cirrhosis (PBC)* [5].

Dataset yang kami analisis mengandung informasi yang penting untuk memahami gejala dan karakteristik pasien dengan sirosis hati, yang merupakan salah satu komplikasi serius dari hepatitis kronis. Sebagai salah satu penyakit prioritas menurut resolusi WHO pada tahun 2020, hepatitis menunjukkan dampak yang signifikan dalam kesehatan global, dengan jumlah orang yang hidup dengan hepatitis B atau C mencapai 354 juta di seluruh dunia dan angka kematian akibat hepatitis yang mencapai satu juta setiap tahunnya. Di Indonesia, prevalensi tinggi terjadi pada kasus hepatitis B, yang menyebabkan dampak serius seperti sirosis hati. Menurut data dari CDA Foundation, angka kematian akibat hepatitis B mencapai 51.100 setiap tahun, sedangkan kematian akibat hepatitis C mencapai 5.942 tiap tahun pada tahun 2016. Sirosis hati ditandai dengan kerusakan parah pada jaringan hati dan pembentukan jaringan parut, yang dapat menyebabkan gangguan fungsi hati secara bertahap dan meningkatkan risiko terjadinya komplikasi serius. Meskipun gejalanya seringkali tidak spesifik, deteksi dini gejala sirosis hati sangat penting untuk meningkatkan prognosis dan mengurangi risiko komplikasi. Dengan mengidentifikasi sirosis hati pada tahap awal, intervensi medis dapat dilakukan lebih awal untuk memperlambat perkembangan penyakit, mengurangi risiko komplikasi, dan meningkatkan kualitas hidup pasien. Namun, tantangan dalam deteksi dini sirosis hati masih ada, terutama karena kurangnya alat diagnostik yang sensitif dan spesifik untuk mendeteksi sirosis hati pada tahap awal. Gejala awal sirosis hati seringkali tidak spesifik dan dapat meliputi kelelahan, penurunan berat badan, perubahan pada kulit dan mata, serta perdarahan gusi. Dalam upaya untuk memahami dan mendeteksi gejala awal penyakit ini, dataset yang digunakan dalam penelitian ini memberikan informasi penting tentang berbagai parameter klinis dan biokimia pasien. Mulai dari usia dan jenis kelamin pasien hingga kehadiran gejala fisik seperti asites, hepatomegali, spider nevi, dan edema, serta parameter laboratorium seperti kadar bilirubin, kolesterol, albumin, tembaga, alkaline phosphatase, dan lain-lain. Dengan menggunakan dataset ini, penelitian ini bertujuan untuk memodelkan deteksi dini gejala penyakit sirosis hati menggunakan pendekatan supervised learning dalam pembelajaran mesin[6], [7]. Diharapkan bahwa analisis data yang cermat akan mengungkap pola-pola yang berkaitan dengan tahap sirosis hati pada pasien, sehingga memungkinkan identifikasi dini pasien yang berisiko tinggi dan intervensi medis yang tepat waktu.

2. Tinjauan Pustaka

Pemodelan deteksi dini gejala penyakit sirosis hati menggunakan machine learning dengan pendekatan supervised learning merupakan suatu pendekatan yang menjanjikan dalam upaya meningkatkan diagnosis dini dan manajemen penyakit yang berkaitan dengan kerusakan hati. Dalam konteks ini, tinjauan pustaka akan menyelidiki kontribusi machine learning, khususnya supervised learning, dalam mengidentifikasi tanda-tanda awal sirosis hati serta tantangan dan solusi yang terlibat dalam penerapannya.

2.1. Pengenalan Machine Learning dalam Deteksi Dini Sirosis Hati



Gambar 1. Konsep *Machine Learning*

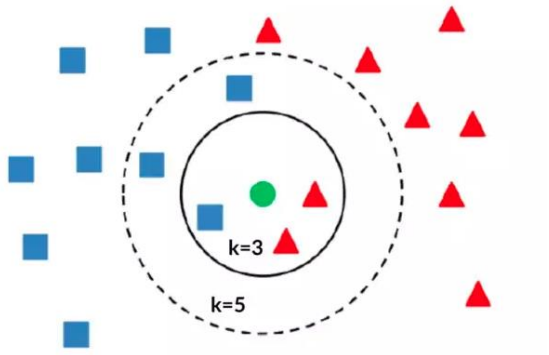
Penggunaan machine learning untuk mengklasifikasi dataset deteksi dini penyakit sirosis hati merupakan aspek penting dalam upaya pencegahan dan penanganan penyakit yang serius ini. Dengan menggunakan konsep dasar machine learning, kita dapat mengembangkan model prediktif yang mampu mengidentifikasi tanda-tanda awal sirosis hati berdasarkan pada data klinis dan biokimia pasien. Gambar 1. menunjukkan beragam pendekatan machine learning yang dapat diterapkan dalam konteks deteksi dini penyakit sirosis hati, termasuk supervised learning, unsupervised learning, semi-supervised learning, dan reinforcement learning. Dalam supervised learning, model dilatih menggunakan data yang berlabel, di mana output yang diinginkan sudah diketahui, seperti tahap sirosis hati pada pasien. Dengan menggunakan data tersebut, model dapat belajar untuk mengenali pola-pola yang berkaitan dengan tahap penyakit dan membuat prediksi berdasarkan input baru. Di sisi lain, unsupervised learning memungkinkan model untuk mengekstraksi struktur atau pola dari data tanpa memerlukan label yang telah ditentukan sebelumnya. Metode ini dapat digunakan untuk mengidentifikasi kelompok pasien dengan karakteristik serupa tanpa adanya informasi tahap penyakit yang telah ditentukan. Sementara itu, semi-supervised learning mencoba memanfaatkan sebagian kecil data yang berlabel bersama dengan sejumlah besar data yang tidak berlabel untuk meningkatkan kinerja model. Terakhir, reinforcement learning memungkinkan model untuk belajar dari interaksi dengan lingkungan, di mana model menerima umpan balik positif atau negatif sebagai hasil dari tindakan yang diambil. Dengan memanfaatkan berbagai teknik machine learning ini, kita dapat memperkuat upaya deteksi dini penyakit sirosis hati, yang merupakan langkah kritis dalam meningkatkan prognosis dan pengelolaan penyakit ini.[8]

2.2. K-NN, Naïve Bayes, dan Support Vector Machine

Dalam konteks deteksi dini sirosis hati menggunakan model supervised learning, beberapa algoritma yang umum digunakan termasuk K-Nearest Neighbors (K-NN), Naïve Bayes, dan Support Vector Machine (SVM).

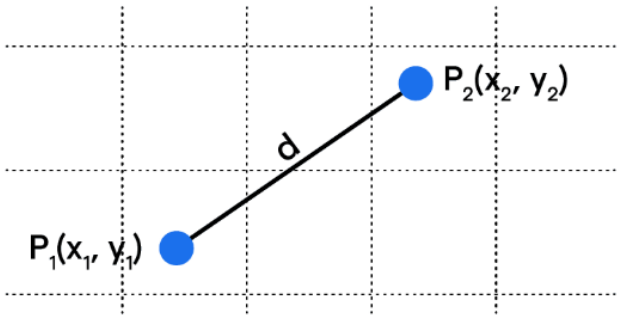
Jurnal Restikom : Riset Teknik Informatika dan Komputer
Vol. 6, No. 2, Agustus 2024, pp. 223-239

a. K-NN



Gambar 2. K-NN *scatter plot*

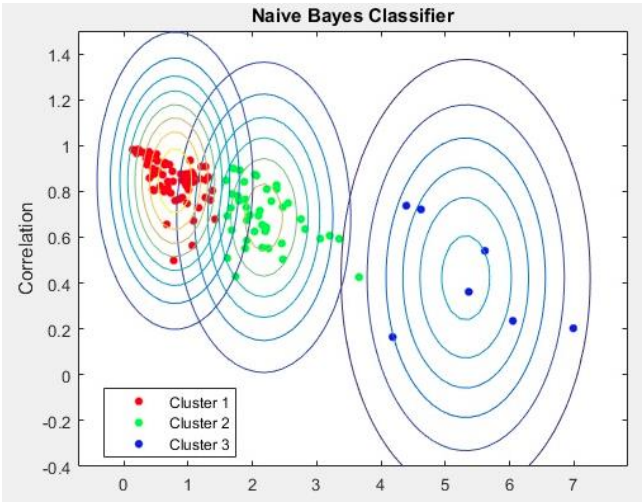
K-NN adalah algoritma yang sederhana dan intuitif. Konsep dasarnya adalah memprediksi kelas suatu data baru berdasarkan mayoritas kelas dari k data terdekat (tetangga terdekat) dalam ruang fitur menggunakan *euclidean distance*.



Gambar 3. Visualisasi *Ecluidean Distance*

Euclidean distance digunakan untuk mengukur jarak antara dua titik dalam ruang fitur. Ini dihitung sebagai jarak lurus antara dua titik dalam ruang multidimensi. Dalam konteks K-NN, Euclidean distance digunakan untuk menghitung jarak antara titik data yang akan diprediksi dengan setiap titik data dalam dataset pelatihan. Kemudian, k titik data terdekat yang memiliki jarak terpendek dengan data yang akan diprediksi dipilih. Setelah tetangga terdekat dipilih, kelas mayoritas dari tetangga-tetangga ini ditentukan. Data baru kemudian diberi label dengan kelas mayoritas ini. Misalnya, jika mayoritas tetangga terdekat adalah kelas "sirosis hati stadium 2", maka data baru akan diprediksi sebagai "sirosis hati stadium 2". Visualisasi Euclidean distance dapat membantu untuk memahami konsep ini dengan lebih baik. Gambar 3 menunjukkan visualisasi Euclidean distance dalam ruang dua dimensi, di mana jarak antara dua titik diplot sebagai garis lurus. Ini membantu untuk memahami bagaimana jarak antara titik data dihitung dan bagaimana titik-titik terdekat dipilih dalam algoritma K-NN. Dalam konteks deteksi dini sirosis hati, K-NN dapat digunakan untuk memprediksi tahap sirosis berdasarkan atribut klinis dan biokimia pasien. Algoritma ini cocok untuk dataset dengan jumlah sampel yang relatif kecil[6], [7], [9].

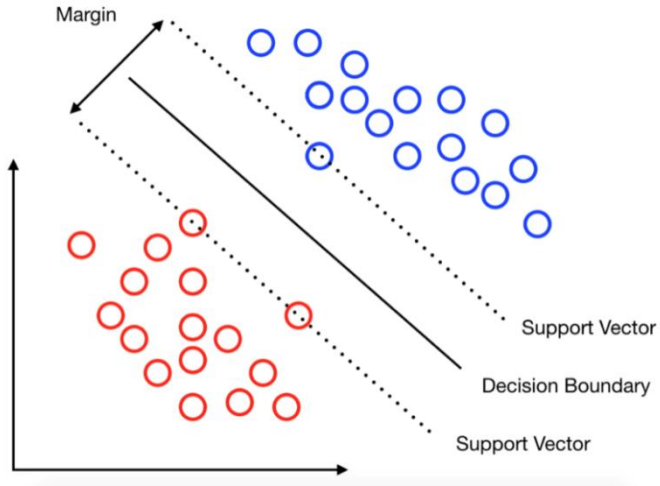
b. Naïve Bayes



Gambar 4. Visualisasi Naïve Bayes

Naïve Bayes adalah algoritma klasifikasi probabilistik yang didasarkan pada teorema Bayes dengan asumsi sederhana yaitu setiap fitur independen satu sama lain, meskipun korelasi antara fitur mungkin ada. Algoritma ini sering digunakan dalam klasifikasi teks dan klasifikasi dokumen. Konsep dasar Naïve Bayes adalah menggunakan teorema Bayes untuk menghitung probabilitas kelas atau label berdasarkan distribusi probabilitas dari fitur-fitur yang diamati. Secara sederhana, Naïve Bayes menghitung probabilitas bahwa suatu data tertentu termasuk dalam suatu kelas berdasarkan fitur-fiturnya. Kemudian, kelas dengan probabilitas tertinggi dipilih sebagai prediksi. Visualisasi Naïve Bayes dapat membantu untuk memahami konsep ini lebih baik. Gambar 4 menunjukkan contoh visualisasi sederhana dari algoritma Naïve Bayes. Dalam gambar ini, ada dua fitur (misalnya, x1 dan x2) dan dua kelas (misalnya, kelas A dan kelas B). Garis putus-putus menunjukkan garis keputusan yang dibuat oleh algoritma Naïve Bayes berdasarkan probabilitas fitur-fitur terhadap kelas-kelas yang diamati dalam dataset. Dengan memahami garis keputusan ini, kita dapat melihat bagaimana algoritma Naïve Bayes membuat prediksi kelas untuk data baru berdasarkan distribusi probabilitas fitur-fiturnya.[10], [11], [12]

c. Support Vector Machine



Gambar 5. Visualisasi Support Vector Machine

Support Vector Machine (SVM) adalah algoritma klasifikasi yang digunakan untuk memisahkan dua kelas dengan mencari hyperplane optimal yang memiliki margin terbesar antara dua kelas. SVM bekerja dengan mencari hyperplane yang memaksimalkan margin, yaitu jarak terdekat antara hyperplane dan titik-titik data dari masing-masing kelas, yang disebut support vectors. Konsep dasar dari SVM adalah untuk mencari hyperplane yang dapat memisahkan dua kelas dengan margin maksimal. Hyperplane ini dipilih sedemikian rupa sehingga jarak terdekat antara hyperplane dan titik-titik data dari masing-masing kelas adalah maksimal. SVM juga dapat digunakan untuk masalah klasifikasi non-linear dengan menggunakan kernel untuk mentransformasi data ke dimensi yang lebih tinggi. Visualisasi Support Vector Machine dapat membantu untuk memahami konsep ini lebih baik. Gambar 5 menunjukkan contoh visualisasi sederhana dari algoritma SVM. Dalam gambar ini, titik-titik data dari dua kelas (misalnya, kelas A dan kelas B) diplot dalam ruang dua dimensi. Garis solid menunjukkan hyperplane yang memisahkan dua kelas, sedangkan garis putus-putus menunjukkan margin, yaitu jarak terdekat antara hyperplane dan titik-titik data dari masing-masing kelas. Dengan memahami konsep margin dan hyperplane dalam SVM, kita dapat melihat bagaimana algoritma ini bekerja untuk memisahkan dua kelas dengan margin maksimal.

d. Dataset

Dataset ini merupakan kumpulan data yang sangat berharga yang memuat informasi tentang berbagai fitur klinis dan biokimia yang terkait dengan pasien yang menderita sirosis hati. Setiap entri dalam dataset ini mencakup informasi tentang berbagai aspek penting, mulai dari jumlah hari sejak awal pengamatan pasien hingga tahap sirosis hati yang dialaminya. Selain itu, dataset ini juga mencatat status pasien, jenis obat yang diberikan, usia, jenis kelamin, serta kehadiran beberapa gejala fisik seperti asites, hepatomegali, spider nevi, dan edema. Parameter-parameter laboratorium seperti kadar bilirubin, kolesterol, albumin, tembaga, dan lain-lain juga dicatat dalam dataset ini. Dengan data yang kaya ini, penelitian dan analisis lebih lanjut dapat dilakukan untuk mengidentifikasi pola-pola yang bermanfaat dalam memahami perkembangan penyakit ini, memperkirakan prognosis pasien, dan mengembangkan metode deteksi dini yang lebih efektif. Dataset ini terdiri dari 25.000 entri dan mencakup informasi tentang pasien-pasien dengan sirosis hati melalui 12 fitur yang berbeda. Berikut adalah deskripsi lengkap dari setiap fitur dalam dataset:

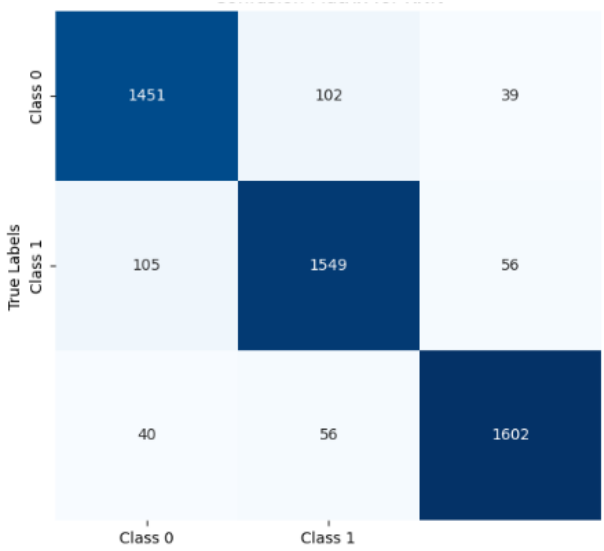
1. N_Days: Merupakan jumlah hari sejak awal pengamatan terhadap pasien.
2. Age: Merupakan usia pasien pada saat pengamatan, yang berkisar antara 9598 tahun hingga 28.650 tahun. Rata-rata usia pasien adalah sekitar 18.495,88 tahun, dengan standar deviasi sebesar 3.737,60 tahun.
3. Bilirubin: Menyatakan kadar bilirubin dalam darah pasien. Nilai-nilai berkisar antara 0,3 hingga 28,0, dengan rata-rata sekitar 3,40 dan standar deviasi sebesar 4,71.
4. Cholesterol: Merupakan kadar kolesterol dalam darah pasien, dengan nilai minimum 120, maksimum 1775, rata-rata sekitar 372,33, dan standar deviasi sekitar 193,67.
5. Albumin: Menggambarkan kadar albumin dalam darah pasien, yang memiliki rentang nilai antara 1,96 hingga 4,64, dengan rata-rata sekitar 3,49 dan standar deviasi sebesar 0,38.
6. Copper: Menunjukkan kadar tembaga dalam darah pasien. Rentang nilainya bervariasi dari 4 hingga 588, dengan rata-rata sekitar 100,18 dan standar deviasi sekitar 73,18.
7. Alk_Phos: Mengindikasikan kadar alkaline phosphatase dalam darah pasien, dengan nilai minimum 289, maksimum 13.862,4, rata-rata sekitar 1995,68, dan standar deviasi sekitar 1798,89.

8. SGOT: Menyatakan kadar serum glutamic-oxaloacetic transaminase dalam darah pasien, yang memiliki rentang nilai antara 26,35 hingga 457,25, dengan rata-rata sekitar 123,17 dan standar deviasi sebesar 47,75.
9. Tryglicerides: Merupakan kadar trigliserida dalam darah pasien, dengan nilai minimum 33, maksimum 598, rata-rata sekitar 123,82, dan standar deviasi sebesar 52,79.
10. Platelets: Menggambarkan jumlah trombosit dalam darah pasien, yang berkisar antara 62 hingga 721, dengan rata-rata sekitar 256,01 dan standar deviasi sekitar 98,68.
11. Prothrombin: Menunjukkan waktu protrombin pasien, dengan nilai minimum 9, maksimum 18, rata-rata sekitar 10,73, dan standar deviasi sebesar 0,90.
12. Stage: Merupakan tahap sirosis hati pada pasien, yang dapat bernilai 1, 2, atau 3.

Deskripsi statistik ini memberikan gambaran yang komprehensif tentang distribusi fitur-fitur dalam dataset, yang dapat digunakan untuk analisis lebih lanjut serta pengembangan model deteksi dini penyakit sirosis hati menggunakan machine learning.

e. Metode Evaluasi Kinerja Model

Confusion Matrix adalah salah satu alat evaluasi yang sangat penting dalam mengevaluasi kinerja model klasifikasi. Tabel ini memberikan gambaran komprehensif tentang seberapa baik model dapat membedakan antara kelas yang berbeda. Terdiri dari empat entri utama: true positives (TP), true negatives (TN), false positives (FP), dan false negatives (FN), confusion matrix memberikan wawasan yang mendalam tentang kekuatan dan kelemahan model.



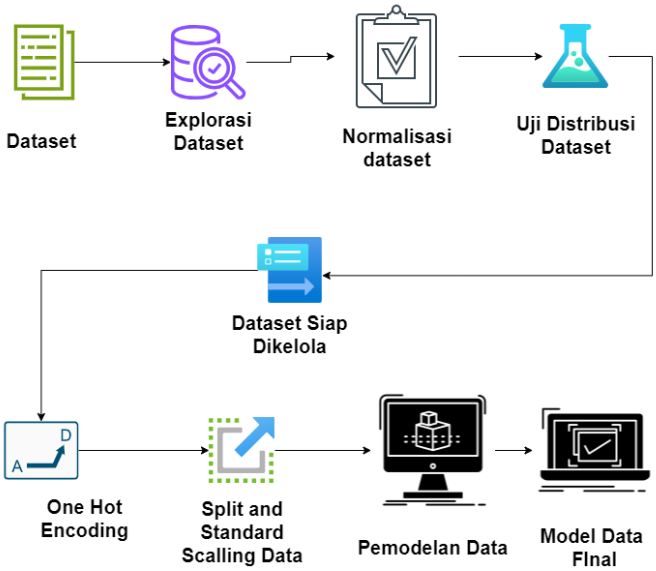
Gambar 6. Visualisasi Confusion Matrix

True positives (TP) adalah jumlah kasus positif yang diprediksi dengan benar oleh model, sementara true negatives (TN) adalah jumlah kasus negatif yang diprediksi dengan benar. Di sisi lain, false positives (FP) adalah jumlah kasus negatif yang salah diprediksi sebagai positif, dan false negatives (FN) adalah jumlah kasus positif yang salah diprediksi sebagai negatif. Dengan memahami kombinasi dari empat entri ini, kita dapat menghitung berbagai metrik evaluasi untuk mengevaluasi kinerja model. Salah satu metrik evaluasi yang umum digunakan adalah akurasi (accuracy), yang mengukur sejauh mana model dapat memprediksi dengan benar kelas dari semua instance. Akurasi dihitung dengan membagi jumlah prediksi yang benar (TP dan TN) dengan jumlah total instance.

Meskipun akurasi memberikan gambaran umum tentang kinerja model, itu mungkin tidak cukup informatif ketika kelas target tidak seimbang secara signifikan. Untuk kasus di mana kelas target tidak seimbang, metrik evaluasi seperti presisi (precision), recall (sensitivity atau true positive rate), dan F1-score menjadi lebih relevan. Presisi mengukur sejauh mana model dapat mengidentifikasi dengan benar instance positif dari semua instance yang diprediksi positif. Recall, di sisi lain, mengukur sejauh mana model dapat mendeteksi dengan benar instance positif dari semua instance yang sebenarnya positif. F1-score adalah nilai rata-rata harmonik dari presisi dan recall, memberikan keseimbangan antara kedua metrik ini.

3. Metodologi Penelitian

Metodologi penelitian ini dirancang untuk menguraikan langkah-langkah yang akan diambil dalam melakukan analisis deteksi dini penyakit sirosis hati menggunakan pendekatan supervised learning dengan memanfaatkan algoritma K-NN, Naïve Bayes, dan Support Vector Machine. Metodologi ini mencakup tahap persiapan data, pelatihan model, evaluasi kinerja model, dan interpretasi hasil.



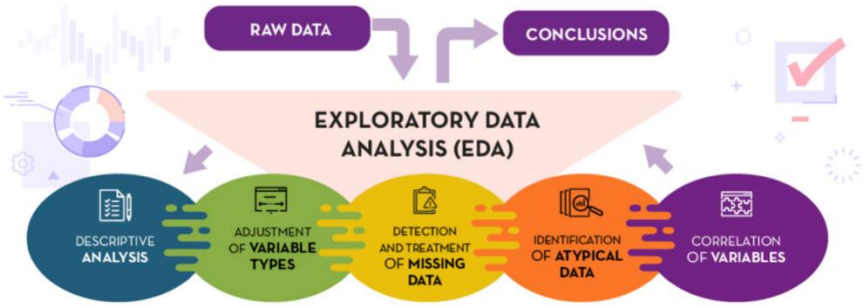
Gambar 7. Methodologi Pemodelan Deteksi Dini Penyakit Sirosis Hati

3.1. Persiapan Data

Langkah pertama dalam penelitian ini adalah persiapan data. Dataset yang digunakan akan dimuat dan dipersiapkan untuk analisis. Ini termasuk langkah-langkah seperti membersihkan data dari nilai yang hilang atau tidak valid, melakukan normalisasi jika diperlukan, dan membagi dataset menjadi subset pelatihan dan pengujian.

3.1.1. Eksplorasi Data

Dalam tahap awal eksplorasi data, fokus pada penelitian ini adalah pada pemahaman yang mendalam tentang dataset sirosis hati yang kami tangani. Pertama-tama, kami memuat dataset dan melakukan peninjauan awal terhadap struktur datanya, mengidentifikasi nama kolom, jenis data, serta potensi keberadaan nilai yang hilang atau tidak valid yang mungkin memerlukan perbaikan lebih lanjut. Selanjutnya, kami melakukan analisis deskriptif yang komprehensif terhadap setiap variabel dalam dataset. Ini mencakup statistik ringkasan seperti rata-rata, median, kuartil, dan rentang untuk variabel numerik seperti usia, kadar bilirubin, kolesterol, albumin, dan lainnya.

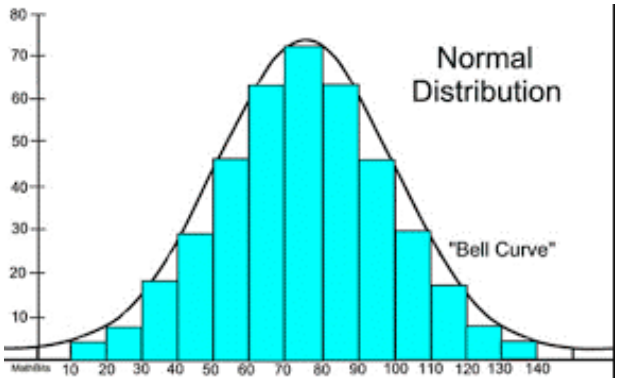


Gambar 8. Kerangka *Exploratory Data Analysis (EDA)*

Untuk variabel kategorikal seperti jenis kelamin, kehadiran asites, hepatomegali, spider nevi, dan edema, kami mengevaluasi distribusi frekuensi dan proporsi masing-masing kategori. Selain analisis deskriptif, kami menggunakan visualisasi data yang mendalam untuk menggambarkan berbagai aspek dari dataset. Kami menciptakan histogram untuk variabel numerik guna mengeksplorasi distribusi dan pola frekuensi, serta box plot untuk menangkap adanya outlier yang potensial. Diagram batang digunakan untuk menganalisis proporsi dan distribusi variabel kategorikal. Selanjutnya, kami menerapkan matriks korelasi untuk melihat hubungan antara berbagai variabel numerik dalam dataset, yang membantu kami memahami tingkat asosiasi antar-fitur. Proses eksplorasi fitur terhadap variabel target, yaitu tahap sirosis hati, kami memperdalam pemahaman kami dengan melakukan analisis perbandingan antara fitur-fitur untuk setiap kelas target. Kita menggunakan berbagai teknik visualisasi seperti scatter plot, heatmap, dan diagram yang membandingkan distribusi fitur-fitur terhadap tahap sirosis hati untuk mengidentifikasi pola yang signifikan. Langkah-langkah eksplorasi ini memberikan wawasan yang komprehensif tentang dataset sirosis hati kami, mempersiapkan fondasi yang kokoh untuk pengembangan model prediktif yang lebih lanjut.

3.1.2. Uji Distribusi Data

Memeriksa distribusi data apakah memenuhi kondisi distribusi normal atau tidak. Ini penting karena beberapa algoritma machine learning mengasumsikan bahwa data memiliki distribusi normal. Jika data tidak memenuhi syarat ini, kami mungkin perlu menerapkan transformasi data seperti log-transform atau Box-Cox untuk membuat distribusi data menjadi lebih normal. Uji distribusi membantu kami menentukan apakah transformasi semacam itu diperlukan sebelum memulai analisis lanjutan.



Gambar 9. Uji Distribusi Data Menggunakan *Bell Curve*

3.1.3. *One Hot Encoding Data*

Kami juga menghadapi variabel kategorikal dalam dataset kami. Untuk menggunakan data kategorikal dalam pemodelan, kami menerapkan teknik one hot encoding untuk mengubahnya menjadi format numerik. Ini memungkinkan kami untuk menghindari bias yang mungkin timbul dari peringkat atau urutan dalam data kategorikal. Dengan menggunakan one hot encoding, kami memperlakukan setiap

kategori sebagai variabel biner yang terpisah, menjadikannya lebih mudah diinterpretasikan oleh algoritma machine learning.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Gambar 10. One-Hot Encoding Machine Learning

Sebagai contoh, dalam dataset sirosis hati yang kami gunakan, variabel seperti jenis kelamin (Sex), keberadaan asites (Ascites), hepatomegali (Hepatomegaly), spider nevi (Spiders), dan edema (Edema) semuanya merupakan variabel kategorikal. Dengan one hot encoding, kami mengubah setiap kategori dalam variabel tersebut menjadi kolom biner yang terpisah, di mana nilai '1' menunjukkan kehadiran fitur tersebut, dan nilai '0' menunjukkan ketidakhadirannya. Misalnya, untuk variabel jenis kelamin, kami memiliki kolom 'Sex_M' dan 'Sex_F', yang menunjukkan apakah pasien adalah laki-laki atau perempuan. Demikian pula, untuk variabel seperti asites dan hepatomegali, kami memiliki kolom 'Ascites_Y', 'Ascites_N', 'Hepatomegaly_Y', dan 'Hepatomegaly_N'. Dengan cara ini, model machine learning dapat menangani variabel kategorikal dengan lebih efektif, karena semua fitur sekarang berada dalam format numerik yang sesuai untuk analisis dan pemodelan. Gambar 10. menunjukkan bagaimana one hot encoding diterapkan pada dataset sirosis hati, mengubah fitur kategorikal menjadi variabel biner yang dapat digunakan dalam pemodelan machine learning. Ini adalah langkah penting dalam preprocessing data yang memastikan bahwa semua variabel dapat diproses dan dianalisis secara akurat oleh algoritma prediksi yang akan kami gunakan.

3.1.4. *Split and Scalling Data*

Sebelum melatih model, kami membagi dataset menjadi dua bagian, yaitu data latih (training) dan data uji (testing). Kami menggunakan teknik seperti train-test split untuk memastikan bahwa model kami dinilai secara objektif. Selain itu, kami melakukan standarisasi skala data numerik untuk memastikan bahwa setiap fitur memiliki pengaruh yang sebanding dalam analisis. Dengan melakukan standarisasi, kami memperbaiki masalah skala dan membantu algoritma machine learning konvergen dengan lebih cepat dan menghasilkan hasil yang lebih baik.

3.1.5 Pemodelan Data

Setelah model dibuat, kami menyimpannya untuk penggunaan dan evaluasi selanjutnya. Ini mencakup tahapan validasi dan fine-tuning model untuk memastikan kinerja yang optimal. Dengan menyimpan model, kami dapat menggunakannya secara langsung untuk prediksi data baru atau mengembangkannya lebih lanjut dalam penelitian mendatang. Akhirnya, kami memiliki dataset final yang telah dipersiapkan dan dikombinasikan dari seluruh langkah-langkah di atas. Dataset ini siap digunakan untuk analisis lebih lanjut dan implementasi solusi AI yang telah dikembangkan.

4. Hasil Kegiatan Riset

Pada bab ini, kami memaparkan hasil dari kegiatan riset yang telah dilakukan. Hasil-hasil ini mencakup berbagai analisis dan pemodelan yang diterapkan pada dataset sirosis hati, serta evaluasi kinerja model machine learning yang digunakan. Kami memulai dengan eksplorasi data untuk memahami distribusi dan karakteristik data yang ada, kemudian melanjutkan dengan penerapan teknik-teknik preprocessing seperti normalisasi dan one hot encoding. Setelah itu, kami membagi data menjadi set pelatihan dan pengujian untuk memastikan validitas model yang dibangun.

4.1 Normalisasi Dataset

Normalisasi dataset merupakan langkah penting dalam preprocessing data untuk memastikan setiap fitur memiliki skala yang sama. Ini membantu dalam menghindari bias dalam analisis dan pemodelan, karena algoritma machine learning sensitif terhadap skala fitur. Normalisasi dilakukan dengan mengubah skala variabel numerik menjadi rentang yang seragam, biasanya antara 0 dan 1, atau menstandarisasi variabel menjadi distribusi dengan rata-rata 0 dan standar deviasi 1. Pada tahap awal, kami memuat dataset dan memeriksa struktur datanya. Proses ini mencakup identifikasi jumlah baris dan kolom, jenis data masing-masing kolom, serta adanya nilai yang hilang atau tidak valid. Setelah itu, kami melakukan analisis statistik deskriptif untuk memahami karakteristik dasar dataset, seperti rata-rata, median, kuartil, dan rentang untuk setiap fitur numerik.

Dalam proses normalisasi, setiap fitur numerik dalam dataset kami diubah ke skala yang seragam. Teknik yang kami gunakan adalah Min-Max Scaling, yang mengubah nilai fitur sehingga berada dalam rentang [0, 1]. Formula yang digunakan adalah:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Di mana X adalah nilai asli, Xmin adalah nilai minimum, dan Xmax adalah nilai maksimum dari fitur tersebut.

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
c	Code cell output actions	0.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.00000
mean	1887.117040	18495.877080	3.402644	372.331471	3.486578	100.184663	1995.675597	123.166345	123.822548	256.007337	10.734549	2.00116
std	1091.690918	3737.596616	4.707491	193.668452	0.380488	73.184840	1798.885660	47.747616	52.786350	98.679006	0.904436	0.81387
min	41.000000	9598.000000	0.300000	120.000000	1.960000	4.000000	289.000000	26.350000	33.000000	62.000000	9.000000	1.00000
25%	1080.000000	15694.000000	0.800000	275.000000	3.290000	52.000000	1032.000000	92.000000	92.000000	189.000000	10.000000	1.00000
50%	1680.000000	18499.000000	1.300000	369.510563	3.510000	97.648387	1828.000000	122.556346	124.702128	251.000000	10.600000	2.00000
75%	2576.000000	20955.000000	3.400000	369.510563	3.750000	107.000000	1982.655769	134.850000	127.000000	311.000000	11.100000	3.00000
max	4795.000000	28650.000000	28.000000	1775.000000	4.640000	588.000000	13862.400000	457.250000	598.000000	721.000000	18.000000	3.00000

Gambar 11. Metode Min-Max pada Dataset Sirosis Hati

Pada Gambar 11, kita dapat melihat bagaimana metode Min-Max Scaling diterapkan pada dataset sirosis hati. Setiap fitur numerik, seperti usia, kadar bilirubin, dan kadar albumin, telah diubah ke dalam rentang 0 hingga 1. Hal ini memungkinkan setiap fitur memiliki skala yang sama, sehingga tidak ada satu fitur pun yang mendominasi atau mempengaruhi hasil analisis secara tidak proporsional. Proses ini sangat penting dalam model machine learning karena beberapa algoritma, seperti k-NN dan SVM, sangat dipengaruhi oleh jarak antar data poin. Jika fitur memiliki skala yang berbeda, fitur dengan rentang yang lebih besar akan memiliki pengaruh yang lebih besar dalam menentukan jarak ini, yang dapat menyebabkan bias dalam model. Dengan menggunakan Min-Max Scaling, kita memastikan bahwa semua fitur berkontribusi secara merata terhadap hasil akhir.

Dengan menerapkan Min-Max Scaling menggunakan formula min – max scalling, kita mentransformasikan setiap nilai fitur menjadi nilai antara 0 dan 1. Contoh normalisasi untuk beberapa fitur:

1. N_Days:
   Original Range: 41 – 4795
   Normalized Range: 0 – 1
2. Age:
   Original Range: 9598 – 28650
   Normalized Range: 0 – 1

3. Bilirubin:
   Original Range: 0.3 – 28
   Normalized Range: 0 - 1
4. Cholesterol:
   Original Range: 120 – 1775
   Normalized Range: 0 - 1

Setelah normalisasi, mean dan standar deviasi setiap fitur akan berada dalam skala yang serupa, yang membuat perbandingan antar fitur lebih intuitif. Kuartil (25%, 50%, dan 75%) akan menunjukkan distribusi data dalam rentang yang distandardisasi. Normalisasi dengan metode Min-Max menghasilkan dataset yang lebih seragam dalam hal skala, memungkinkan algoritma machine learning untuk melakukan analisis dengan lebih adil dan akurat. Gambar 11 menunjukkan hasil normalisasi yang memvisualisasikan bagaimana setiap fitur numerik telah disesuaikan ke dalam rentang [0, 1], sehingga siap untuk langkah-langkah pemodelan lebih lanjut.

4.2 One – Hot Encoding

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	\
0	2221	18499	0.5	149.0	4.04	227.0	598.0	52.70	
1	1230	19724	0.5	219.0	3.93	22.0	663.0	45.00	
2	4184	11839	0.5	320.0	3.54	51.0	1243.0	122.45	
3	2090	16467	0.7	255.0	3.74	23.0	1024.0	77.50	
4	2105	21699	1.9	486.0	3.54	74.0	1052.0	108.50	

	Tryglicerides	Platelets	...	Sex_M	Ascites_N	Ascites_Y	Hepatomegaly_N	\
0	57.0	256.0	...	0.0	1.0	0.0	0.0	
1	75.0	220.0	...	1.0	0.0	1.0	1.0	
2	80.0	225.0	...	0.0	1.0	0.0	1.0	
3	58.0	151.0	...	0.0	1.0	0.0	1.0	
4	109.0	151.0	...	0.0	1.0	0.0	0.0	

	Hepatomegaly_Y	Spiders_N	Spiders_Y	Edema_N	Edema_S	Edema_Y
0	1.0	1.0	0.0	1.0	0.0	0.0
1	0.0	0.0	1.0	1.0	0.0	0.0
2	0.0	1.0	0.0	1.0	0.0	0.0
3	0.0	1.0	0.0	1.0	0.0	0.0
4	1.0	1.0	0.0	1.0	0.0	0.0

Gambar 12. Hasil One – Hot encoding Dataset

One-Hot Encoding adalah teknik yang digunakan untuk mengubah variabel kategorikal menjadi format yang dapat digunakan dalam pemodelan machine learning. Variabel kategorikal tidak dapat langsung digunakan dalam model numerik karena algoritma machine learning sensitif terhadap skala dan nilai. Oleh karena itu, variabel kategorikal harus diubah menjadi format biner yang lebih mudah diinterpretasikan oleh algoritma. Pada Gambar 12. di atas, kami menerapkan one-hot encoding pada dataset sirosis hati. Berikut penjelasan mengenai setiap variabel dan hasil dari proses one-hot encoding yang telah dilakukan:

- N_Days, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Tryglicerides, Platelets, Prothrombin, Stage: Variabel numerik yang tidak mengalami perubahan dalam proses one-hot encoding. Variabel ini tetap dalam bentuk aslinya karena sudah dalam format numerik yang dapat digunakan langsung dalam pemodelan.
- Sex (M/F): Variabel kategorikal yang menunjukkan jenis kelamin pasien. Dalam tabel, "Sex_M" adalah variabel biner yang menunjukkan apakah pasien laki-laki (1) atau bukan (0).
- Ascites (Y/N): Variabel kategorikal yang menunjukkan adanya asites pada pasien. Dalam tabel, "Ascites_N" menunjukkan tidak ada asites (1 jika tidak ada, 0 jika ada) dan "Ascites_Y" menunjukkan adanya asites (1 jika ada, 0 jika tidak ada).
- Hepatomegaly (Y/N): Variabel kategorikal yang menunjukkan adanya hepatomegali pada pasien. Dalam tabel, "Hepatomegaly_N" menunjukkan tidak ada hepatomegali (1 jika tidak ada, 0 jika ada) dan "Hepatomegaly_Y" menunjukkan adanya hepatomegali (1 jika ada, 0 jika tidak ada).

- Spiders (Y/N): Variabel kategorikal yang menunjukkan adanya spider angiomata pada pasien. Dalam tabel, "Spiders_N" menunjukkan tidak ada spider angiomata (1 jika tidak ada, 0 jika ada) dan "Spiders_Y" menunjukkan adanya spider angiomata (1 jika ada, 0 jika tidak ada).
- Edema (N/S/Y): Variabel kategorikal dengan tiga kemungkinan nilai (N: tidak ada edema, S: edema dengan diuretik, Y: edema tanpa diuretik). Dalam tabel, "Edema_N" menunjukkan tidak ada edema (1 jika tidak ada, 0 jika ada), "Edema_S" menunjukkan adanya edema dengan diuretik (1 jika ada, 0 jika tidak ada), dan "Edema_Y" menunjukkan adanya edema tanpa diuretik (1 jika ada, 0 jika tidak ada).

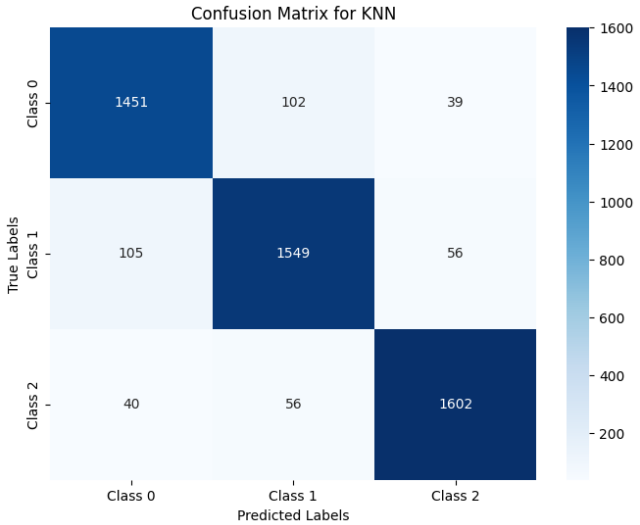
Proses one-hot encoding ini memastikan bahwa variabel kategorikal diubah menjadi beberapa variabel biner yang dapat digunakan dalam algoritma machine learning tanpa memperkenalkan bias yang mungkin timbul dari peringkat atau urutan dalam data kategorikal. Dengan menggunakan one-hot encoding, setiap kategori diperlakukan sebagai variabel yang terpisah, yang membuat interpretasi oleh algoritma menjadi lebih tepat dan efektif. Tabel hasil one-hot encoding menunjukkan bagaimana variabel-variabel kategorikal tersebut telah diubah menjadi bentuk biner yang siap untuk digunakan dalam model prediksi. Ini adalah langkah penting dalam preprocessing data yang memastikan bahwa semua fitur, baik numerik maupun kategorikal, diintegrasikan dengan benar ke dalam analisis dan pemodelan selanjutnya.

4.3 Hasil Pemodelan Menggunakan Supervised Learning

Dalam penelitian ini, kami menerapkan tiga algoritma machine learning yang berbeda untuk membangun model prediksi berdasarkan dataset sirosis hati. Algoritma yang digunakan adalah K-Nearest Neighbors (KNN), Naive Bayes, dan Support Vector Machine (SVM). Hasil evaluasi kinerja dari masing-masing model, termasuk metrik akurasi dan confusion matrix, disajikan sebagai berikut:

4.3.1 K – Nearest Neighbor

Model KNN bekerja dengan cara membandingkan setiap instance baru dengan k instance terdekat dalam dataset pelatihan. Hasil dari model KNN menunjukkan akurasi yang cukup tinggi, yaitu 92.04%. Berikut adalah hasil detail dari model KNN:



Gambar 13. Confusion Matrix untuk Pemodelan Menggunakan K-NN

Confusion matrix pada Gambar 13. yang diberikan menunjukkan detail dari hasil klasifikasi:

- Kelas pertama (Class 1): Terdapat 1451 contoh yang diklasifikasikan dengan benar, 102 contoh salah diklasifikasikan sebagai Kelas 2, dan 39 contoh salah diklasifikasikan sebagai Kelas 3.

- Kelas kedua (Class 2): Terdapat 1549 contoh yang diklasifikasikan dengan benar, 105 contoh salah diklasifikasikan sebagai Kelas 1, dan 56 contoh salah diklasifikasikan sebagai Kelas 3.
- Kelas ketiga (Class 3): Terdapat 1602 contoh yang diklasifikasikan dengan benar, 40 contoh salah diklasifikasikan sebagai Kelas 1, dan 56 contoh salah diklasifikasikan sebagai Kelas 2.

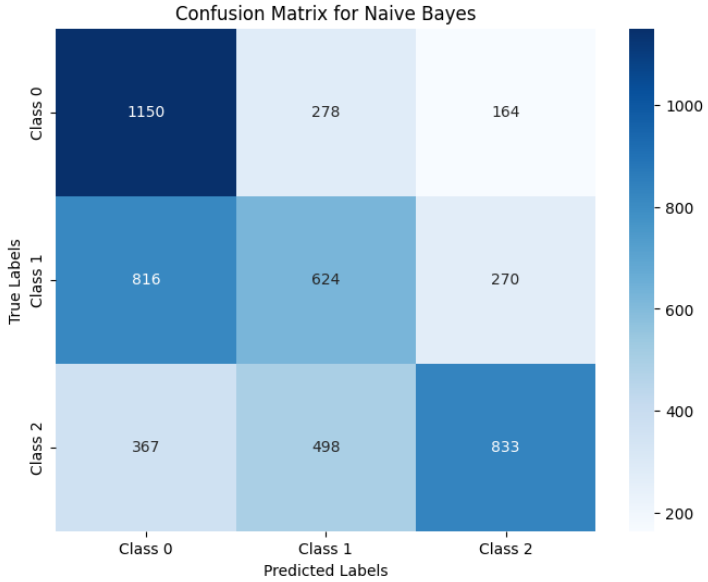
	precision	recall	f1-score	support
1	0.91	0.91	0.91	1596
2	0.91	0.91	0.91	1707
3	0.94	0.94	0.94	1697
accuracy			0.92	5000
macro avg	0.92	0.92	0.92	5000
weighted avg	0.92	0.92	0.92	5000

Gambar 14. Classification Report untuk Pemodelan Menggunakan K-NN

Dari laporan evaluasi klasifikasi pada Gambar 14., kita dapat melihat bahwa: Untuk Kelas 1, presisi (precision) sebesar 91% menunjukkan bahwa dari semua prediksi yang dilakukan sebagai Kelas 1, 91% di antaranya benar-benar merupakan Kelas 1. Recall sebesar 91% menunjukkan bahwa dari semua contoh Kelas 1 yang sebenarnya, model berhasil mengidentifikasi 91% dari mereka. Nilai F1-score sebesar 91% mencerminkan harmonic mean antara presisi dan recall. Dengan dukungan (support) sebanyak 1596, akurasi model untuk kelas ini adalah 92%. Untuk Kelas 2, presisi, recall, dan F1-score juga mencapai 91%, menunjukkan konsistensi performa model dalam mengklasifikasikan kelas ini. Akurasi model untuk Kelas 2 juga mencapai 92%. Untuk Kelas 3, nilai presisi sebesar 94%, recall sebesar 94%, dan F1-score sebesar 94%. Ini menunjukkan bahwa model memiliki kinerja yang sangat baik dalam mengidentifikasi Kelas 3. Dukungan untuk kelas ini adalah 1697, dan akurasi model adalah 92%.

4.3.2 Naïve Bayes

Naive Bayes adalah algoritma berbasis probabilistik yang mengasumsikan independensi antar fitur. Hasil dari model ini menunjukkan akurasi yang lebih rendah dibandingkan dengan KNN, yaitu 52.14%. Berikut adalah hasil detail dari model Naive Bayes:



Gambar 15. Confusion Matrix untuk Pemodelan Menggunakan Naïve Bayes

Hasil pemodelan menggunakan algoritma Naive Bayes menunjukkan performa yang menengah. Confusion matrix yang diberikan menunjukkan detail dari hasil klasifikasi:

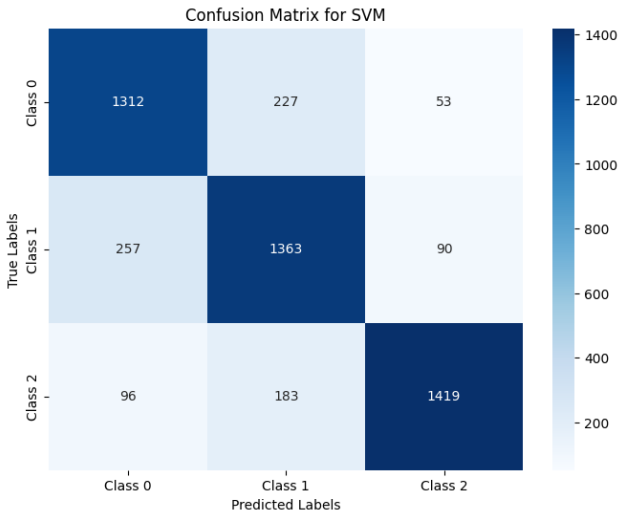
- Kelas pertama (Class 1): Terdapat 1150 contoh yang diklasifikasikan dengan benar, 278 contoh salah diklasifikasikan sebagai Kelas 2, dan 164 contoh salah diklasifikasikan sebagai Kelas 3.
- Kelas kedua (Class 2): Terdapat 624 contoh yang diklasifikasikan dengan benar, 816 contoh salah diklasifikasikan sebagai Kelas 1, dan 270 contoh salah diklasifikasikan sebagai Kelas 3.
- Kelas ketiga (Class 3): Terdapat 833 contoh yang diklasifikasikan dengan benar, 367 contoh salah diklasifikasikan sebagai Kelas 1, dan 498 contoh salah diklasifikasikan sebagai Kelas 2.

	precision	recall	f1-score	support
1	0.72	0.49	0.59	2333
2	0.36	0.45	0.40	1400
3	0.49	0.66	0.56	1267
accuracy			0.52	5000
macro avg	0.53	0.53	0.52	5000
weighted avg	0.56	0.52	0.53	5000

Gambar 16. Classification Report untuk Pemodelan Menggunakan Naïve Bayes

Dari laporan evaluasi klasifikasi pada Gambar 16, kita dapat melihat bahwa: Untuk Kelas 1, presisi (precision) sebesar 72% menunjukkan bahwa dari semua prediksi yang dilakukan sebagai Kelas 1, 72% di antaranya benar-benar merupakan Kelas 1. Recall sebesar 49% menunjukkan bahwa dari semua contoh Kelas 1 yang sebenarnya, model berhasil mengidentifikasi 49% dari mereka. Nilai F1-score sebesar 59% mencerminkan harmonic mean antara presisi dan recall. Dengan dukungan (support) sebanyak 2333, akurasi model untuk kelas ini adalah 52%. Untuk Kelas 2, presisi, recall, dan F1-score juga cukup rendah, masing-masing sebesar 36%, 45%, dan 40%. Akurasi model untuk Kelas 2 adalah 52%. Untuk Kelas 3, nilai presisi sebesar 49%, recall sebesar 66%, dan F1-score sebesar 56%. Ini menunjukkan bahwa model memiliki performa sedikit lebih baik dalam mengidentifikasi Kelas 3. Dukungan untuk kelas ini adalah 1267, dan akurasi model adalah 52%. Dari confusion matrix, kita dapat melihat bahwa model Naive Bayes memiliki performa yang lebih rendah dibandingkan dengan model KNN, dengan tingkat kesalahan yang lebih tinggi dalam mengklasifikasikan setiap kelas.

4.3.3 Support Vector Machine

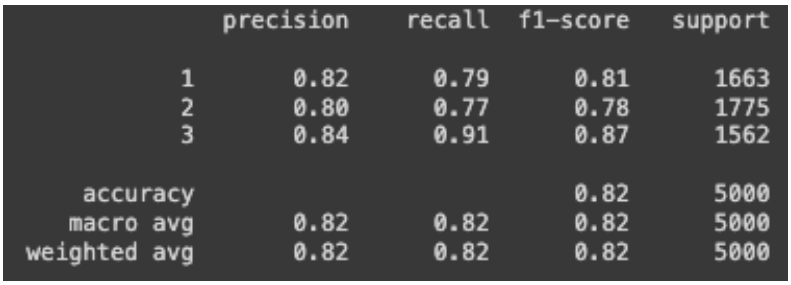


Gambar 17. Confusion Matrix untuk Pemodelan Menggunakan Support Vector Machine

Model SVM bekerja dengan mencari hyperplane yang optimal untuk memisahkan kelas-kelas dalam ruang fitur yang lebih tinggi. Hasil dari model SVM menunjukkan akurasi yang baik, yaitu 81.88%. Berikut adalah hasil detail dari model SVM:

Hasil pemodelan menggunakan algoritma Support Vector Machine (SVM) menunjukkan performa yang baik. Confusion matrix yang diberikan menunjukkan detail dari hasil klasifikasi:

- Kelas pertama (Class 1): Terdapat 1312 contoh yang diklasifikasikan dengan benar, 227 contoh salah diklasifikasikan sebagai Kelas 2, dan 53 contoh salah diklasifikasikan sebagai Kelas 3.
- Kelas kedua (Class 2): Terdapat 1363 contoh yang diklasifikasikan dengan benar, 257 contoh salah diklasifikasikan sebagai Kelas 1, dan 90 contoh salah diklasifikasikan sebagai Kelas 3.
- Kelas ketiga (Class 3): Terdapat 1419 contoh yang diklasifikasikan dengan benar, 96 contoh salah diklasifikasikan sebagai Kelas 1, dan 183 contoh salah diklasifikasikan sebagai Kelas 2.



	precision	recall	f1-score	support
1	0.82	0.79	0.81	1663
2	0.80	0.77	0.78	1775
3	0.84	0.91	0.87	1562
accuracy			0.82	5000
macro avg	0.82	0.82	0.82	5000
weighted avg	0.82	0.82	0.82	5000

Gambar 18. Classification Report untuk Pemodelan Menggunakan Support Vector Machine

Dari laporan evaluasi klasifikasi, kita dapat melihat bahwa: Untuk Kelas 1, presisi (precision) sebesar 82% menunjukkan bahwa dari semua prediksi yang dilakukan sebagai Kelas 1, 82% di antaranya benar-benar merupakan Kelas 1. Recall sebesar 79% menunjukkan bahwa dari semua contoh Kelas 1 yang sebenarnya, model berhasil mengidentifikasi 79% dari mereka. Nilai F1-score sebesar 81% mencerminkan harmonic mean antara presisi dan recall. Dengan dukungan (support) sebanyak 1663, akurasi model untuk kelas ini adalah 82%. Untuk Kelas 2, presisi, recall, dan F1-score juga cukup baik, masing-masing sebesar 80%, 77%, dan 78%. Akurasi model untuk Kelas 2 adalah 82%. Untuk Kelas 3, nilai presisi sebesar 84%, recall sebesar 91%, dan F1-score sebesar 87%. Ini menunjukkan bahwa model memiliki kinerja yang sangat baik dalam mengidentifikasi Kelas 3. Dukungan untuk kelas ini adalah 1562, dan akurasi model adalah 82%.

5. KESIMPULAN

Dalam penelitian ini, kami memproses dataset sirosis hati dengan normalisasi untuk memastikan konsistensi skala fitur. Kemudian, kami menerapkan teknik one-hot encoding untuk variabel kategorikal. Selanjutnya, kami membandingkan tiga model machine learning: K-Nearest Neighbors (KNN), Naive Bayes, dan Support Vector Machine (SVM).

Hasilnya menunjukkan bahwa:

- KNN memiliki akurasi tertinggi (92.04%), menunjukkan kemampuan efektif dalam menangkap pola data.
- Naive Bayes menunjukkan performa yang lebih rendah (52.14%), dengan banyak kesalahan klasifikasi.
- SVM memiliki akurasi yang cukup baik (81.88%) dan seimbang antara ketepatan dan recall.

Dengan demikian, untuk prediksi sirosis hati, kami merekomendasikan penggunaan model KNN karena akurasinya yang tinggi. Langkah-langkah preprocessing data, seperti normalisasi dan one-hot encoding, juga berperan penting dalam meningkatkan kinerja model.

DAFTAR PUSTAKA

[1] D. Sartika, M. A. Yuswar, and R. Susanti, "POLA PENGGUNAAN ANTIBITIK PADA PASIEN SIROSIS HATI DENGAN HEMATEMESIS MELENA DI RUMAH SAKIT UNIVERSITAS TANJUNGPURA KOTA ...," *Jurnal Mahasiswa Farmasi Fakultas …*, [Online]. Available: https://jurnal.untan.ac.id/index.php/jmfarmasi/article/view/44518

[2] M. Nurrofikoh, A. Fatima, H. Hastuti, and ..., "Cegah dan Kenali Kondisi Hati (CEK SI HATI) sebagai Upaya Pendidikan Kesehatan terkait Sirosis Hati Kepada Masyarakat," *Jurnal Kreativitas …*, 2023, [Online]. Available: https://karya.brin.go.id/33224/1/2615-0921_6_7_2023-37.pdf

[3] P. Mondrowinduro, I. Hasan, and ..., "Disfungsi Diastolik Ventrikel Kiri pada Pasien Sirosis Hati: Proporsi, Korelasi, dan Hubungan Parameter Fungsi Diastolik dengan Derajat Disfungsi Hati," *Jurnal …*, 2018, [Online]. Available: http://download.garuda.kemdikbud.go.id/article.php?article=753205&val=10415&title=Disfungsi%20Diastolik%20Ventrikel%20Kiri%20pada%20Pasien%20Sirosis%20Hati%20Proporsi%20Korelasi%20dan%20Hubungan%20Parameter%20Fungsi%20Diastolik%20dengan%20Derajat%20Disfungsi%20Hati

[4] O. K. Hernomo, "Sirosis Hati," *… Noer,(editors). Ilmu Penyakit Hati. Edisi I. Jakarta …*, 2007.

[5] B. P. Silaban, F. Lumongga, and ..., "Karakteristik Penderita Sirosis Hati," *Jurnal Kedokteran …*, 2020, [Online]. Available: https://ejurnal.methodist.ac.id/index.php/jkm/article/view/1322

[6] K. Keykhosravi, A. Hamednia, H. Rastegarfar, and E. Agrell, "Data preprocessing for machine-learning-based adaptive data center transmission," *ICT Express*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959522000212

[7] M. Frye, J. Mohren, and R. H. Schmitt, "Benchmarking of Data Preprocessing Methods for Machine Learning-Applications in Production," *Procedia CIRP*, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212827121009070

[8] A. R. Jatmiko, N. D. Hendrawan, R. M. Arief, F. Putra, and ..., "Signifikansi Pengaruh Akses Teknologi Informasi terhadap Indeks Pembangunan Manusia di Indonesia," …, 2023, [Online]. Available: https://www.jurnal.harapan.ac.id/index.php/Jitekh/article/view/780

[9] T. S. Ustun, S. M. S. Hussain, A. Ulutas, A. Onen, M. M. Roomi, and ..., "Machine learning-based intrusion detection for achieving cybersecurity in smart grids using IEC 61850 GOOSE messages," *Symmetry (Basel)*, 2021, [Online]. Available: https://www.mdpi.com/1101618

[10] D. Hendrycks, M. Mazeika, and ..., "Using self-supervised learning can improve model robustness and uncertainty," *Advances in neural …*, 2019, [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/a2b15837edac15df90721968986f7f8e-Abstract.html

[11] G. N. W. Paramartha, D. E. Ratnawati, and ..., "Analisis Perbandingan Metode K-Means Dengan Improved Semi-Supervised K-Means Pada Data Indeks Pembangunan Manusia (IPM)," *… Teknologi Informasi dan …*, 2017, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/237

[12] C. Liu and K. Gryllias, "A semi-supervised Support Vector Data Description-based fault detection method for rolling element bearings based on cyclic spectral analysis," *Mech Syst Signal Process*, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020300686