

Supervised Machine Learning Model untuk Prediksi Penyakit Hepatitis**Andriyan Dwi Putra, Dwi Nurani, Melany Mustika Dewi, Alfie Nur Rahmi, Supriatin**andriyan@amikom.ac.id, dwinurani@amikom.ac.id, melany@amikom.ac.id,
alfienurrahmi@amikom.ac.id, supriatin@amikom.ac.id

Universitas Amikom Yogyakarta

Informasi ArtikelDiterima : 8 Mar 2024
Direview : 6 Apr 2024
Disetujui : 30 Apr 2024**Kata Kunci**Hepatitis, Klasifikasi,
Supervised Machine Learning, *Naïve Bayes*, *K-Nearest Neighbor*.**Abstrak**

Hepatitis menjadi salah satu penyakit mematikan yang diakibatkan karena peradangan yang terjadi pada organ hati manusia. Hepatitis seringkali disebabkan karena infeksi virus dan gaya hidup yang tidak sehat. Hepatitis bahkan bisa menular apabila dikaitkan dengan infeksi dari adanya virus tertentu. Hepatitis perlu dideteksi secara dini dan diantisipasi sedini mungkin sehingga tidak mengakibatkan adanya penyakit komplikasi yang lebih serius yang bahkan dapat mengakibatkan terjadinya kematian. Perkembangan teknologi informasi dan komunikasi yang terus berkembang hingga saat ini memungkinkan penyakit hepatitis untuk dapat dikenali dan diprediksi. Salah satunya menggunakan teknologi pembelajaran mesin. Pada penelitian ini, metode *supervised learning* yang menerapkan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* digunakan untuk memprediksi adanya penyakit hepatitis. Dengan menggunakan dataset yang diunduh secara langsung dari halaman *website UCI Machine Learning Repository*, *Naïve Bayes* menghasilkan nilai akurasi sebesar 91.67% dengan nilai presisi dan *recall* mencapai 95%, Sedangkan penggunaan *K-Nearest Neighbor* menghasilkan nilai akurasi sebesar 95.8%, dengan adanya perbedaan nilai presisi dan *recall* sebesar 1%, menunjukkan bahwa penggunaan *supervised machine learning model* berdasarkan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* memiliki potensi untuk digunakan dalam pengembangan berbagai sistem terutama untuk prediksi penyakit hepatitis.

Keywords*Hepatitis, Classification, Supervised Machine Learning, Naïve Bayes, K-Nearest Neighbor.***Abstrak**

Hepatitis is a deadly disease caused by inflammation that occurs in the human liver. Hepatitis is often caused by viral infections and unhealthy lifestyles. Hepatitis can even be contagious if it is associated with infection from certain viruses. Hepatitis needs to be detected early and anticipated as early as possible so that it does not result in more serious complications which can even result in death. The development of information and communication technology that continues to grow to date allows hepatitis to be recognized and predicted. One of them uses machine learning technology. In this study, the supervised learning method that applies the Naïve Bayes and K-Nearest Neighbor algorithms will be used to predict the presence of hepatitis. By using a dataset downloaded directly from the UCI Machine Learning website page, Naïve Bayes algorithm produces an accuracy of 91.67% with a precision and recall value of up to 95%, while K-Nearest Neighbor produces an accuracy value of 95.8%, with a difference in precision and recall values of 1%, it shows that supervised machine learning models based on the Naïve Bayes and K-Nearest Neighbor algorithms shows the potential to be used in the development of various systems, especially for the prediction of hepatitis disease.

A. Pendahuluan

Hepatitis menjadi satu dari sekian banyaknya masalah kesehatan dengan jumlah kematian yang signifikan hampir di seluruh dunia [1]. Hepatitis menjadi penyakit yang menyebabkan munculnya peradangan pada organ dalam manusia seperti Hati dan Liver [2], [3], [4]. Hepatitis umumnya dipengaruhi oleh berbagai faktor biotik seperti adanya infeksi virus, bakteri, dan parasit. Akan tetapi, penggunaan obat tradisional dengan sifat yang *toxic*, serta penggunaan obat tradisional yang terinteraksi dengan obat lain, dan autoimun juga dapat menjadi faktor yang menyebabkan seseorang dapat terserang penyakit hepatitis. Sampai dengan saat ini, penyakit hepatitis memiliki 5 varian yakni hepatitis A, B, C, D, dan E. Untuk Hepatitis A, B, D, dan E, merupakan jenis hepatitis yang disebabkan oleh virus, sedangkan Hepatitis C merupakan jenis hepatitis akibat dari adanya *sirosis*, virus HCV, dan kanker hati [5], [6]. Di Indonesia, menurut data yang dihimpun dari dan dirilis oleh kementerian Kesehatan Indonesia, penyakit hepatitis ini termasuk dalam lima besar penyakit menular dan memerlukan perhatian secara serius [3], [7]. Hingga tahun 2022, tercatat sebanyak 7.1% atau sekitar 18 juta Masyarakat Indonesia terdeteksi memiliki infeksi hepatitis B, dimana 50% diantaranya bahkan memiliki resiko yang tinggi dan berpotensi menjadi penyakit kronis yang mengarah pada peradangan dan kanker hati. Bahkan, hingga kini, tercatat sebanyak 35.757 bayi yang baru lahir dari ibu yang telah terinfeksi hepatitis juga positif tertular hepatitis meskipun telah mendapatkan imunisasi *Hb0* dan *Hbg* dalam waktu kurang dari 24 jam [2]. Melihat adanya bahaya tersebut, Indonesia dalam sidang Majelis Kesehatan dunia (*World Health Assembly*) menjadi salah satu negara yang menyerukan serta memprakarsai agar hepatitis menjadi perhatian dunia sehingga penanggulangan, pencegahan, hingga pengobatan dapat dilakukan secara lebih komprehensif [5].

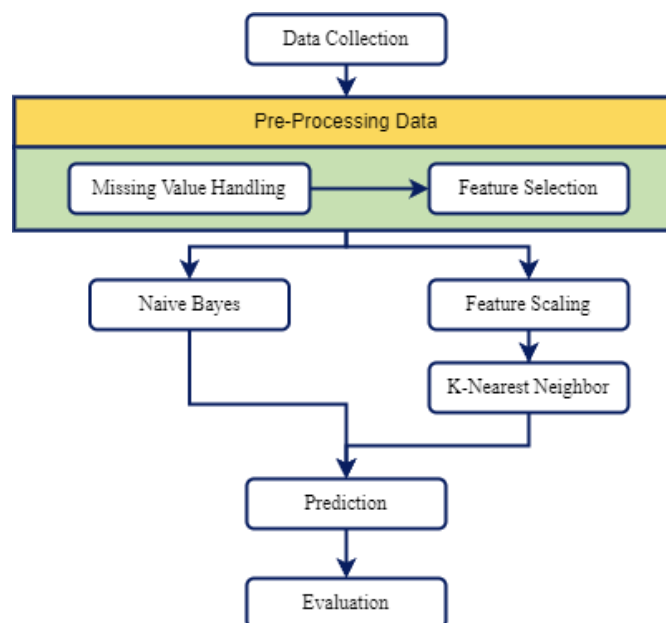
Penanggulangan dan pencegahan hepatitis dapat dilakukan dengan cepat dan tepat apabila indikasi dari adanya penyakit hepatitis dapat dikenali sedini mungkin [5], [7]. Oleh karena itu perlu adanya dukungan teknologi maupun sistem yang memungkinkan untuk dapat mengenali adanya penyakit hepatitis sebagai langkah awal pencegahan dan penanggulangan hepatitis. Di era teknologi yang berkembang pesat seperti sekarang ini, pengembangan berbagai sistem terutama yang spesifik untuk mengidentifikasi hepatitis telah banyak dilakukan oleh para peneliti sebelumnya, dimana salah satunya menggunakan pendekatan metode pembelajaran mesin (*machine learning*) [2], [3], [7]. Sebagai contoh pada penelitian [2], Implementasi pembelajaran mesin dengan algoritma *Naïve Bayes* yang diproses menggunakan aplikasi *rapidminer* menghasilkan nilai akurasi sebesar 76.77%. Algoritma *Naïve Bayes* merupakan salah satu algoritma model pembelajaran mesin yang populer dalam menyelesaikan berbagai permasalahan. Pernyataan tersebut didukung oleh asumsi bahwa setiap variabel pada algoritma *Naïve Bayes* memiliki independensi kuat dan mempertimbangkan nilai antar fitur dengan kelas variabel. Selain itu *Naïve Bayes* memiliki nilai kerja yang positif dengan penerapan skema sederhana, tidak terdampak *noise* data, atribut yang tidak relevan dan dalam prosesnya tidak membutuhkan estimasi parameter iteratif yang rumit [8], [9], [10]. Namun asumsi atau fitur yang bersifat independen pada algoritma *Naïve Bayes* seringkali tidak berlaku karena bentuk data yang memiliki ketergantungan antar fiturnya dapat mengakibatkan kesalahan dalam proses klasifikasi [9].

Algoritma populer lainnya yang digunakan sebagai model pembelajaran mesin adalah algoritma *K-Nearest Neighbor* (KNN). Kinerja dari algoritma KNN bergantung dari nilai k serta penentuan ukuran jarak (*distance*) [8], [9], [11]. Penelitian [11] menunjukkan perbedaan kinerja berdasarkan proses evaluasi yang menekankan perubahan faktor pengukuran dengan nilai ' k ' yang konstan terutama pencarian nilai ' k ' yang optimal. Penggunaan algoritma KNN menunjukkan nilai akurasi hingga 83% dalam klasifikasi *email spam* berbasis teks dan *image*. Dalam konteks pengembangan sistem deteksi hepatitis, penelitian [3] menunjukkan bahwa penggunaan KNN dalam memprediksi adanya hepatitis C menghasilkan nilai akurasi terbaik sebesar 94.40% dibandingkan *Decision Tree*, *Logistic Regression* maupun *Gradient Boosting*.

Dengan adanya berbagai referensi tersebut, maka pada penelitian ini akan dilakukan eksperimen menggunakan pembelajaran mesin terawasi (*supervised machine learning*) menggunakan algoritma *Naïve Bayes* dan KNN yang kemudian akan dievaluasi dan dibandingkan kinerjanya untuk mengetahui kinerja dari algoritma tersebut dalam melakukan prediksi terhadap penyakit hepatitis. Evaluasi model akan diujicoba pada benchmark hepatitis dataset yang dapat diunduh secara gratis pada repositori UCI Machine Learning yang dapat diakses melalui URL <https://archive.ics.uci.edu/ml/datasets/Hepatitis> [12].

B. Metode Penelitian

Penelitian ini menekankan pada aspek evaluasi dari *supervised machine learning model* dengan penerapan metode *Naïve Bayes* dan *K-Nearest Neighbor* (KNN) dalam memprediksi penyakit hepatitis. Pada proses eksperimennya, penelitian ini melakukan berbagai pemrosesan secara berurutan yang dimulai dari tahapan untuk pengambilan data, pra-pemrosesan data, pemilihan fitur, pemodelan, dan evaluasi sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

Pada Gambar 1, alur penelitian diawali dengan tahapan *data collection*, dilanjutkan dengan tahapan *pre-processing* yang meliputi tahapan penanganan terhadap *missing value (missing value handling)* dan tahapan untuk pemilihan fitur yang akan digunakan dalam penelitian. Hasil dari *pre-processing* data berikutnya akan dilakukan eksperimen untuk dimodelkan secara langsung menggunakan *Naïve Bayes* dan *K-Nearest Neighbor*. Namun, pada proses pemodelan yang dilakukan menggunakan KNN akan melewati proses *feature scaling* yang berperan meminimalisir waktu proses pemodelan. Hasil dari pemodelan kemudian akan dilakukan evaluasi untuk mengetahui kinerja terbaik dalam memprediksi hepatitis.

Setiap tahapan pemrosesan dalam penelitian ini akan dijelaskan secara lebih detail melalui masing-masing subbab.

- *Data Collection*

Data yang digunakan dalam penelitian ini diambil dari situs Halaman *website UCI Machine Learning Repository* terkait dengan Hepatitis Dataset [12]. Dataset Hepatitis terdiri dari 155 *record* data dengan sejumlah 20 atribut atau fitur data yang terdiri yakni: *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime, histology*, serta atribut *class* sebagai label datanya.

- *Pre-Processing Data*

Tahapan *pre-processing* data dimulai dengan mempersiapkan data mentah sebelum dilakukan proses selanjutnya melalui *missing value handling* dan *feature selection*.

- *Missing Value Handling*

Berdasarkan data yang tercantum di dalam dataset Hepatitis, terdapat 75 *record* data yang memiliki *missing value* sehingga untuk menanganinya penulis melakukan penghapusan pada *record* data yang memiliki nilai *missing value* tersebut.

- *Feature Selection*

Tahap ini bertujuan untuk memilih atribut (*feature*) data yang memiliki kontribusi besar dalam pembentukan model *machine learning*. Seleksi Fitur membantu dalam memahami data, mengurangi kebutuhan komputasi, mengurangi efek *curse* dimensi dan meningkatkan kinerja prediktor [5]. Teknik yang digunakan untuk melakukan filter pada *feature selection* yaitu *Pearson's Correlation*. Tujuan penggunaan metode ini dalam menyeleksi fitur didasarkan pada [7] agar mendapatkan fitur dengan nilai *Pearson's Correlation* tertinggi. Seleksi fitur menggunakan metode pemfilteran dengan koefisien korelasi *Pearson* bisa meningkatkan akurasi sehingga dapat efektif, akurat, dan cepat pada jumlah fitur yang banyak. Rumus *Pearson's Correlation* [13] dituliskan dalam persamaan (1).

$$R(i) = \frac{cov(x_i, y)}{\sqrt{var(x_i) * var(Y)}} \quad (1)$$

- Algoritma *Naïve Bayes*

Naïve Bayes merupakan sebuah metode klasifikasi probabilistik yang dapat digunakan untuk menghitung kombinasi frekuensi dan nilai dari kumpulan data yang telah disediakan. Metode ini [14], [15] menerapkan teorema *bayes* dalam menghitung probabilitas untuk semua atribut independen yang diberikan oleh nilai *class variable*. Berikut adalah rumus dasar dari teorema *bayes* disajikan pada persamaan (2).

$$P(C|X) = \frac{P(C).P(X|C)}{P(X)} \quad (2)$$

Dimana nilai X mewakili atribut, dan nilai C mewakili *class*. $P(C|X)$ digunakan untuk menghitung nilai peluang C terhadap X , sedangkan $P(X|C)$ digunakan untuk menghitung nilai peluang X terhadap C . Untuk $P(C)$ dan $P(X)$ sendiri mewakili peluang C dan peluang X . Oleh karena nilai X mewakili atribut, maka dapat dijabarkan pada persamaan (3).

$$X = (X_1, X_2, \dots, X_n) \quad (3)$$

Sehingga persamaan untuk teorema bayes disebutkan pada persamaan (4).

$$P(C|x_1, x_2, \dots, x_n) = \frac{P(C).P(x_1, x_2, \dots, x_n | C)}{P(x_1, x_2, \dots, x_n)} \quad (4)$$

Keuntungan menggunakan metode *naive bayes classifier* yaitu hanya membutuhkan sedikit data *training* agar bisa menentukan estimasi parameter yang dibutuhkan dalam proses klasifikasi [15]. Berdasarkan [16] terdapat 3 variasi algoritma *naive bayes* yaitu *multinomial*, *bernoulli*, dan *gaussian naive bayes*. Pada penelitian ini menerapkan *gaussian naive bayes* yang memiliki persamaan (5) [14]:

$$P = (X_i = x_i | C = c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right) \quad (5)$$

Dimana, nilai P adalah peluang dengan X_i (atribut ke- i) sama dengan x_i (nilai atribut ke- i). Sedangkan nilai C yang digunakan untuk menganalogikan *class* memiliki nilai yang sama dengan c_i (*sub class* ke- i). Kemudian terdapat μ yang mewakili nilai rata-rata untuk semua atribut, dan σ mewakili *standard deviation*.

- *Feature Scaling*
Feature Scaling [5] digunakan untuk mengurangi dampak variasi data yang berpengaruh pada kinerja perhitungan jarak atau *distance* untuk tiap *feature* data. Disini peneliti menerapkan fungsi *StandarScaler* untuk melakukan standarisasi skala sehingga dapat mengurangi ketidakstabilan pada numerik yang disebabkan karena angka yang diproses terlalu besar. Perhitungan standarisasi skala menggunakan *Standar Scaler* dituliskan ke dalam persamaan (6).

$$X_{new} = \frac{X_i - X_{mean}}{Standard\ Deviation} \quad (6)$$

- *K-Nearest Neighbor*
K-Nearest Neighbor adalah sebuah algoritma yang digunakan untuk melakukan klasifikasi pada objek berdasarkan pada data *learning* yang memiliki *distance* terdekat dengan objek tersebut [8]. Dengan demikian, *K-Nearest Neighbor* dapat dituliskan dengan persamaan (7).

$$Euclidean = \sqrt{\sum_{i=1}^k (X_i - Y_i)^2} \quad (7)$$

Terdapat 2 variasi *K-Nearest Neighbor*, yaitu Klasifikasi dan Regresi. Pada penelitian ini akan menerapkan *K-Nearest Neighbor Classification* (Klasifikasi). Sebelum menjalankan program klasifikasi, peneliti melakukan *splitting data* yang bertujuan untuk membagi data menjadi *data training* dan *data testing*. *Data training* digunakan untuk *training model*, sedangkan *data testing* digunakan untuk mengetahui kinerja algoritma yang sebelumnya yang sudah dilatih apabila menemukan data baru dengan format yang sesuai dengan *experience* sebelumnya [15].

C. Hasil dan Pembahasan

1. Pre-Processing Data

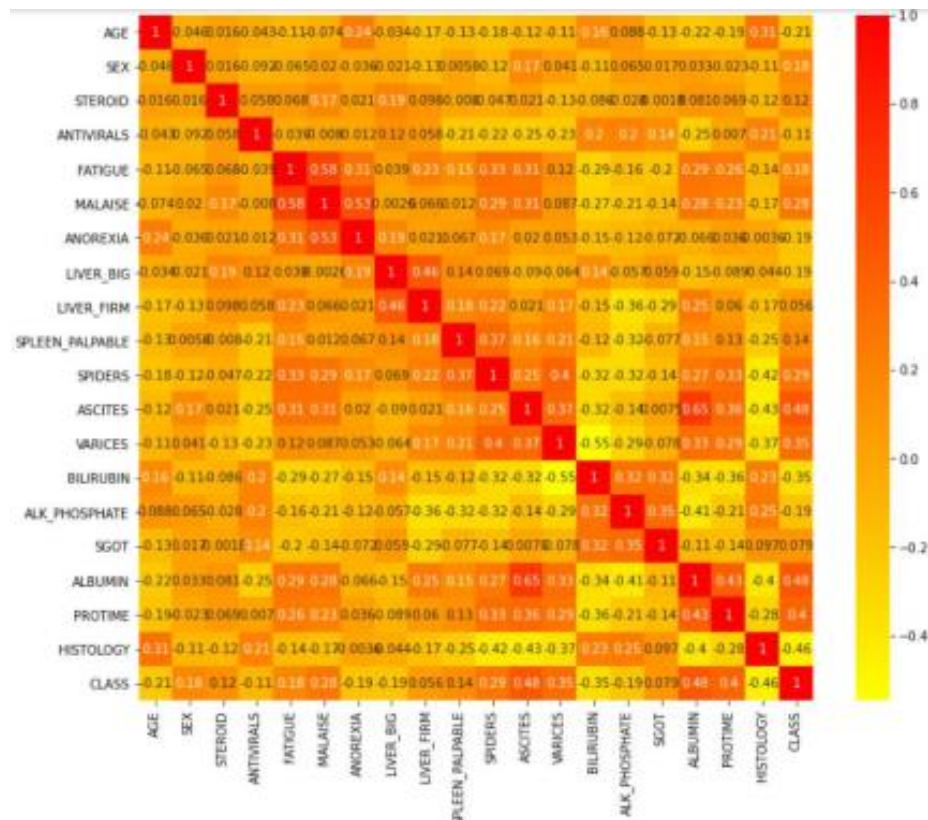
Dataset yang diperoleh kemudian diproses menggunakan platform *google collabs*. Dalam memproses dataset tersebut peneliti membutuhkan beberapa *library*, yaitu *pandas* yang digunakan untuk mengolah dan menganalisis data, *numpy* yang berguna dalam mempermudah operasi komputasi pada tipe data numerik, dan *seaborn* yang berperan dalam pembuatan grafik dan model statistik dari data yang diproses. Pertama, *import* dataset ke dalam *notebook* kemudian *library pandas* akan membaca isi dari dataset tersebut. Dari data sejumlah 155 baris (*record*), sebanyak 75 *record* diantaranya terdapat *missing value*. Untuk menanganinya, peneliti menghapus *record* yang mengandung *missing value* tersebut. Hasil dari data yang sudah bersih disajikan pada Tabel 1.

Tabel 1. Isi Dataset Hepatitis

	Age	Sex	Steroid	Antivirals	::	Prottime	Instology	Class
0	34	1	2	2	::	76	1	2
1	39	1	1	1	::	85	1	2
2	32	1	2	1	::	54	1	2
3	41	1	2	1	::	52	1	2
4	30	1	2	2	::	78	1	2
...	::
79	43	1	2	2	::	42	2	1

2. Feature Selection

Pada tahapan ini digunakan untuk memilih *attribute (feature)* data mana saja yang memiliki kontribusi besar dalam proses pembentukan model *machine learning* karena tidak semua fitur yang besar menghasilkan model yang bagus. Terkadang ada beberapa fitur yang tidak relevan justru bisa membuat kinerja model mengalami penurunan. Dalam tahap ini, peneliti menggunakan teknik *Pearson's Correlatif Coefisien*. Dalam pengimplementasiannya pada dataset, peneliti memanfaatkan *library seaborn* yang dikolaborasikan dengan *matplotlib (pyplot)* sehingga dapat menghasilkan visualisasi data seperti pada Gambar 2.



Gambar 2. Visualisasi Heatmap Pearson's Correlation

Kemudian peneliti melakukan proses seleksi fitur melakukan pemfilteran pada fitur yang memiliki nilai *relevant* lebih dari 0.4. Disini diperoleh 4 buah fitur yang terseleksi, yaitu *ASCITES*, *ALBUMIN*, *HISTOLOGY*, dan *CLASS* yang memiliki detail nilai seperti pada Tabel 2.

Tabel 2. Hasil *Feature Selection Pearson's Correlation*

Fitur	Nilai Relevan
<i>Ascites</i>	0.479211
<i>Albumin</i>	0.477404
<i>Histology</i>	0.456856
<i>Class</i>	1.00000

a. *Naïve Bayes*

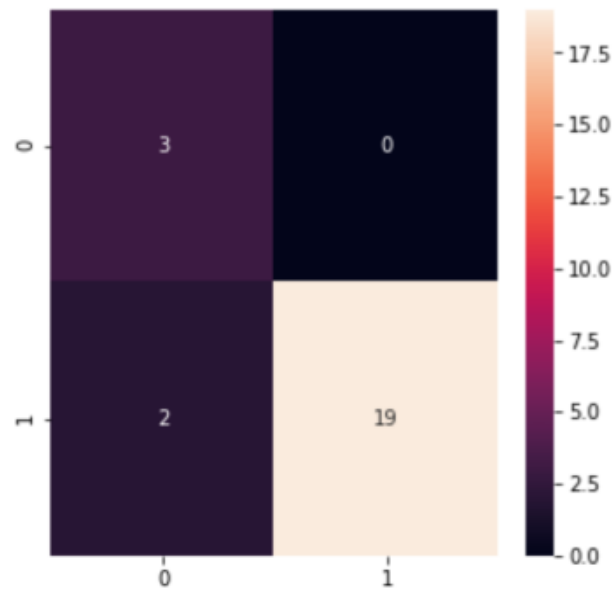
Berdasarkan hasil seleksi yang dilakukan pada tahap sebelumnya, maka fitur *ASCITES*, *ALBUMIN*, dan *HISTOLOGI* dimasukkan ke dalam input data, sedangkan fitur *class* dimasukkan ke dalam *output* data. Dalam proses selanjutnya, dataset kemudian dibagi dengan proporsi sebesar 70 % untuk *data training* dan 30% *data testing*. Kemudian diproses menggunakan metode *Gaussian Naive Bayes*.

Pengujian sistem (evaluasi) sistem klasifikasi pada metode *Naive Bayes* dilakukan menggunakan *confusion matrix*. Evaluasi ini berfungsi untuk mengetahui tingkat keakuratan dari model yang telah dibuat. Hasil akurasi didapatkan dengan cara melakukan perbandingan data pada tabel yang sudah diberi label "benar" atau "salah" pada model. Adapun detail pengujian menggunakan *confusion matrix* ditampilkan dalam Gambar 3.

		Actual Class	
		Class 1	Class 2
Predicted Class	Class 1	True Positive	False Negative
	Class 2	True Negative	False Positive

Gambar 3. *Confusion Matrix*

Dari penelitian yang dilakukan, diperoleh hasil *confusion matrix* seperti dibawah ini. Terlihat bahwa *true positif* (nilai benar yang diprediksi benar) menghasilkan nilai matrik sebesar 3 dan *true negative* (nilai benar yang diprediksi salah) menghasilkan nilai matrik sebesar 2. Sedangkan untuk *false negative* (nilai salah yang diprediksi salah) menghasilkan nilai matrik 0, dan *false positive* (nilai salah yang diprediksi benar) menghasilkan nilai matrik 19.

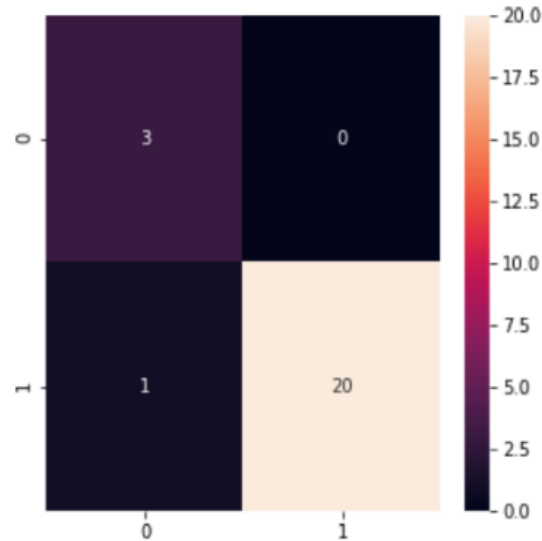


Gambar 4. *Confusion Matrix Naïve Bayes*

b. K-Nearest Neighbor

Sebelum melakukan klasifikasi menggunakan *K-Nearest Neighbor* (KNN), maka perlu dilakukan tahap *feature scaling* untuk membuat standarisasi skala pada data yang akan diproses. Sehingga dapat mengurangi ketimpangan pada data yang memiliki selisih yang besar dan akan memudahkan proses klasifikasi. *Feature scaling* ini memanfaatkan fungsi *library Standard Scaler*. Teknik *scaling* ini diterapkan untuk melakukan standarisasi skala pada *data training* dan *data testing* pada model. Setelah diperoleh hasil standarisasi, maka peneliti mulai menerapkan algoritma *K-Nearest Neighbor* ke dalam model tersebut. Pada tahapan ini ditentukan bahwa nilai n *neighbors* atau biasa disebut dengan k *neighbors*. Nilai ini digunakan untuk mengevaluasi hasil perhitungan jarak (*distance*) terdekat antara *data training* dengan tetangga terdekatnya. Dengan rasio *splitting* yang sama dengan *Naïve Bayes*, yaitu 70% *data training* dan 30% *data testing* maka algoritma *K-Nearest Neighbor* mulai menghitung nilai prediksi dari keluaran (*output*) yang akan dihasilkan.

Kemudian hasilnya ditampilkan ke dalam *confusion matrix* dengan detail metodenya sudah dijelaskan pada bagian metode *Naive Bayes*. Kemudian hasil *confusion matrix* ini ditampilkan visualisasinya agar memudahkan dalam proses analisis. Berikut adalah hasil visualisasi *confusion matrix* pada algoritma *K-Nearest Neighbor* sebagaimana Gambar 5.



Gambar 5. Confusion Matrix K-Nearest Neighbor

Berdasarkan hasil penelitian menggunakan metode *Gaussian Naive Bayes* dan *K-Nearest Neighbor*, maka peneliti mengukur tingkat *accuracy*, *precision*, *recall* dan *F-Measure* dari model yang digunakan menggunakan persamaan (8) hingga (11).

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (8)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (9)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (10)$$

$$F - Measure = 2 * \frac{(Recall*Precision)}{(Recall+Precision)} \quad (11)$$

Hasil pengukuran metode *Naive Bayes* dengan teknik *accuracy* sebesar 91,67%, *precision* sebesar 95%, *recall* sebesar 95% dan *F1-Score* sebesar 93% seperti pada pada Tabel 3.

Tabel 3. Hasil Klasifikasi Naive Bayes

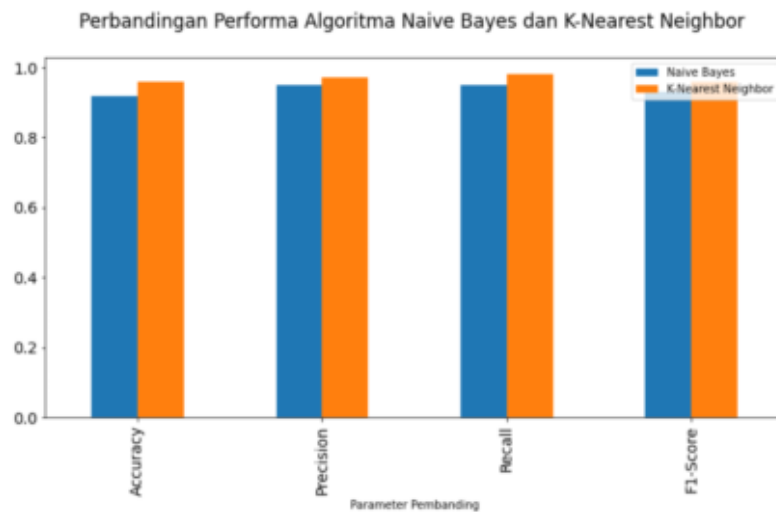
Parameter	Score
<i>Accuracy</i>	91,67%
<i>Precision</i>	95%
<i>Recall</i>	95%
<i>F1-Score</i>	93%

Hasil pengukuran metode *K-Nearest Neighbor* dengan teknik *accuracy* sebesar 95,83%, *precision* sebesar 97%, *recall* sebesar 98% dan *F1-Score* sebesar 96% disajikan pada pada Tabel 4.

Tabel 4. Hasil Klasifikasi K-Nearest Neighbor

Parameter	Score
<i>Accuracy</i>	95.83%
<i>Precision</i>	97%
<i>Recall</i>	98%
<i>F1-Score</i>	96%

Perbandingan dari kinerja dan performa dari metode *Naïve Bayes* dan *K-Nearest Neighbor* disajikan dalam bentuk grafik batang seperti pada Gambar 6 dapat disimpulkan bahwa *K-Nearest Neighbor* lebih unggul dibandingkan dengan metode *Naïve Bayes* dari semua teknik pengukuran.



Gambar 6. Grafik Perbandingan Kinerja

D. Simpulan

Penggunaan algoritma *supervised learning* dengan metode *Naive Bayes* dan *K-Nearest Neighbor* telah dilakukan untuk memprediksi adanya penyakit hepatitis berdasarkan Dataset Hepatitis yang diperoleh dari *UCI Machine Learning Repository*. Eksperimen dilakukan memanfaatkan platform *Google Collaboratory* dan menghasilkan kinerja sebesar 91.67% ketika menggunakan algoritma *Naive Bayes*. Sedangkan penggunaan algoritma *K-Nearest Neighbor* menunjukkan adanya peningkatan nilai akurasi hingga sebesar 95.83%. dari eksperimen yang dilakukan ini menunjukkan bahwa penggunaan *supervised learning model* dengan metode *Naive Bayes* maupun *K-Nearest Neighbor* menunjukkan adanya potensi untuk digunakan dalam pengembangan berbagai sistem yang berkaitan dengan prediksi penyakit hepatitis.

E. Referensi

- [1] D. Novianti, "Implementasi Algoritma Naïve Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner," *Paradig. - J. Komput. dan Inform.*, vol. 21, no. 1, pp. 49–54, 2019, doi: 10.31294/p.v21i1.4979.
- [2] A. Damayanti and G. Testiana, "Penerapan Data Mining Untuk Prediksi Penyakit Hepatitis C Menggunakan Algoritma Naïve Bayes," *J. Manaj. Inform. Jayakarta*, vol. 3, no. 2, pp. 177–186, 2023, [Online]. Available: <https://doi.org/10.52362/jmijayakarta.v3i2.1098>
- [3] S. Sulastri, K. Hadiono, and M. T. Anwar, "Analisis Perbandingan Klasifikasi Prediksi Penyakit Hepatitis Dengan Menggunakan Algoritma K-Nearest Neighbor, Naïve Bayes Dan Neural Network," *Dinamik*, vol. 24, no. 2, pp. 82–91, 2020, doi: 10.35315/dinamik.v24i2.7867.
- [4] N. Sharfina and N. G. Ramadhan, "Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 8, no. 1, p. 33, 2023, doi: 10.31328/jointecs.v8i1.4456.
- [5] E. Rajiah, H. Hasanuddin, and D. Wahyuni, "Prediksi Penyakit Hepatitis C dan Sirosis Hati dengan Penerapan SMOTE pada Metode Jaringan Saraf Tiruan Propagasi Balik," *Prism. Fis.*, vol. 11, no. 3, p. 71, 2023, doi: 10.26418/pf.v11i3.65229.
- [6] A. H. S. Jones and M. S. Makmun, "Implementasi Metode CART untuk Klasifikasi Diagnosis Penyakit Hepatitis Pada," *J. INISTA*, vol. Vol.3, no. No.2, pp. 61–70, 2021.
- [7] S. Utara, "Prediksi Penyakit Hepatitis Dengan Analisis Pembobotan Metode Swara dan Pendekatan K-Nearst," no. December, 2023.
- [8] M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 2007, pp. 1541–1546. doi: 10.1109/ICCIT.2007.148.
- [9] Y. Findawati, I. R. I. Astutik, A. S. Fitroni, I. Indrawati, and N. Yuniasih, "Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast," *J. Phys. Conf. Ser.*, vol. 1402, no. 6, 2019, doi: 10.1088/1742-6596/1402/6/066046.
- [10] P. Sainanda Cahyani Moonallika, K. Queena Fredlina, and I. Kresna Sudiatmika, "Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara)," *Progresif J. Ilm. Komput.*, vol. 6, no. 1, pp. 47–56, 2020.
- [11] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decis. Anal. J.*, 2022, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2772662222000261>
- [12] UCI Machine Learning, "UCI Machine Learning Repository." [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Hepatitis>
- [13] H. Husdi, "Pengenalan Ekspresi Wajah Pengguna Elearning Menggunakan Artificial Neural Network Dengan Fitur Ekstraksi Local Binary Pattern Dan Gray Level Co-Occurrence Matrix," *Ilk. J. Ilm.*, vol. 8, no. 3, pp. 212–219, 2016,

- doi: 10.33096/ilkom.v8i3.58.212-219.
- [14] S. Maheswari and R. Pitchai, "Heart Disease Prediction System Using Decision Tree and Naive Bayes Algorithm," *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.*, vol. 15, no. 8, pp. 712–717, 2018, doi: 10.2174/1573405614666180322141259.
- [15] V. T. Fauziastuti and L. A. Rakhman, "A Review of Students' Graduation Classification: A Comparison of Naive Bayes Classifier and K-Nearest Neighbour," *1st Int. Multidiscip. ...*, 2020, doi: 10.2991/assehr.k.200303.052.
- [16] K. Ganesh Reddy and P. Santhi Thilagam, "Naïve bayes classifier to mitigate the DDoS attacks severity in Ad-Hoc networks," *Int. J. Commun. Networks Inf. Secur.*, vol. 12, no. 2, pp. 221–226, 2020.