Journal of Information Technology and Computer Science (INTECOMS)
Volume 8 Nomor 4, Tahun 2025
e-ISSN : 2614-1574
p-ISSN : 2621-3249

KOMPARASI AKURASI ALGORITMA LOGISTIC REGRESSION DAN NAIVE BAYES PADA DATASET HEPATITIS

COMPARISON OF THE ACCURACY OF LOGISTIC REGRESSION AND NAIVE BAYES ALGORITHMS ON THE HEPATITIS DATASET

Muhammad Riansyah[1], Harry Pratama Fiqna[2], Syaiful Bahri[3]
[123]Program Studi Pendidikan Teknik Informatika, STKIP Al Maksum, Langkat, Indonesia
mohriansyah9@gmail.com

ABSTRACT

This study aims to compare the accuracy of two classification algorithms, namely Logistic Regression and Naive Bayes, in the process of diagnosing hepatitis. The research data was obtained from the Kaggle.com website and then divided into 80% training data and 20% test data. The evaluation process was carried out through a validation method using several measurement metrics, including accuracy, classification error, precision, recall, and f1-score. Based on the test results, the Logistic Regression algorithm successfully achieved an accuracy level of 90.32%, while the Naive Bayes algorithm only achieved an accuracy of 83.87%. These results indicate that Logistic Regression provides better performance than Naive Bayes. The findings of this study are expected to be a reference for selecting classification algorithms in machine learning-based medical data processing.

Keywords: *Comparison, Logistic Regression Algorithm, Naive Bayes, machine learning, Hepatitis*

ABSTRAK

Penelitian ini bertujuan untuk membandingkan tingkat akurasi dua algoritma klasifikasi, yaitu *Logistic Regression* dan *Naive Bayes*, dalam proses mendiagnosis penyakit hepatitis. Data penelitian diperoleh dari situs Kaggle.com yang kemudian dibagi menjadi 80% data latih dan 20% data uji. Proses evaluasi dilakukan melalui metode validasi dengan menggunakan beberapa metrik pengukuran, antara lain *accuracy*, *classification error*, *precision*, *recall*, dan *F1-score*. Berdasarkan hasil pengujian yang dilakukan, algoritma *Logistic Regression* berhasil mencapai tingkat akurasi sebesar 90,32%, sedangkan algoritma *Naive Bayes* hanya memperoleh akurasi sebesar 83,87%. Hasil ini menunjukkan bahwa *Logistic Regression* memberikan kinerja yang lebih baik dibandingkan *Naive Bayes*. Temuan penelitian ini diharapkan dapat menjadi referensi pemilihan algoritma klasifikasi dalam pengolahan data medis berbasis *machine learning*.

Kata Kunci: Komparasi, Algoritma *Logistic Regression*, *Naive Bayes, machine learning,* Hepatitis

PENDAHULUAN

Penyakit hepatitis merupakan salah satu gangguan serius pada organ hati yang dapat menyebabkan kerusakan jangka panjang, seperti sirosis dan kanker hati. Deteksi dini sangat diperlukan untuk meningkatkan efektivitas pengobatan serta menurunkan angka kematian. Dengan berkembangnya teknologi informasi dan ketersediaan data kesehatan, pemanfaatan algoritma machine learning menjadi alternatif yang menjanjikan dalam membantu proses diagnosis penyakit, termasuk hepatitis.

Salah satu permasalahan yang diidentifikasi dalam penelitian ini adalah belum adanya kajian komprehensif terkini yang secara khusus membandingkan performa algoritma *Logistic Regression* dan *Naive Bayes* dalam klasifikasi data hepatitis. Kaunang et al. (2022) membandingkan kedua algoritma bahwa

Logistic Regression memperoleh akurasi sekitar 97,9%, sedangkan *Naive Bayes* sekitar 96%. Meskipun demikian, studi tersebut belum mengevaluasi metrik performa lain seperti presisi, recall, dan F1-score secara menyeluruh, serta tidak dilakukan analisis statistik untuk menguji signifikansi perbedaan kinerja kedua algoritma.

Tinjauan pustaka terkini menunjukkan bahwa kedua algoritma memiliki kelebihan masing-masing. Li et al. (2022) menunjukkan bahwa *Logistic Regression* cukup kompetitif dibandingkan algoritma lain seperti *Logistic Regression* dalam mendeteksi infeksi hepatitis C. Di sisi lain, *Naive Bayes* masih banyak digunakan dalam pengolahan data medis karena kesederhanaan dan efisiensinya, meskipun asumsinya yang mengharuskan independensi antar fitur sering kali tidak terpenuhi pada data medis (Melinte-Popescu et al., 2023). Selain itu, penelitian oleh Menon et al. (2022) membuktikan bahwa *Logistic Regression* tetap efektif bahkan ketika data mengalami proses imputasi berulang, menandakan kestabilan performa algoritma tersebut pada kondisi data yang kompleks.

Penelitian ini bertujuan untuk membandingkan performa algoritma *Logistic Regression* dan *Naive Bayes* dalam mengklasifikasikan dataset hepatitis. Evaluasi dilakukan tidak hanya berdasarkan akurasi, tetapi juga mencakup metrik presisi, recall, dan F1-score untuk memberikan gambaran menyeluruh terhadap kinerja model. Selain itu, dilakukan uji statistik (seperti uji-t) guna menguji signifikansi perbedaan performa kedua algoritma.

Manfaat dari penelitian ini memberikan referensi empiris bagi akademisi dan praktisi di bidang data science dan kesehatan mengenai pemilihan algoritma klasifikasi yang paling sesuai dalam kasus diagnosis hepatitis. Dengan pendekatan yang komprehensif, hasil penelitian ini juga diharapkan dapat berkontribusi pada pengembangan sistem pendukung keputusan medis berbasis machine learning yang lebih akurat, efisien, dan dapat diandalkan dalam deteksi dini penyakit hepatitis.

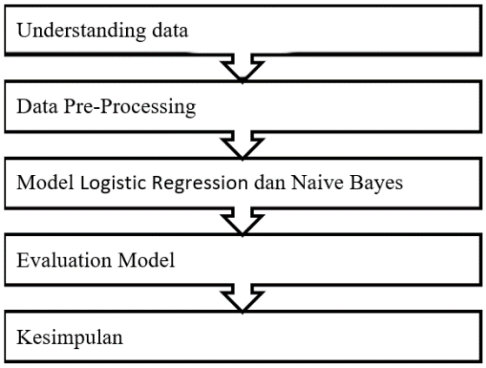
METODE

Penelitian ini menggunakan metode kuantitatif dengan desain eksperimental untuk membandingkan performa dua algoritma klasifikasi, yaitu *Logistic Regression* dan *Naive Bayes*, dalam mendiagnosis hepatitis berdasarkan data pasien. Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari sumber terbuka dan sudah tersedia secara publik. Dataset yang digunakan dalam penelitian ini berasal dari *Kaggle.com*, yaitu Hepatitis.

Dataset ini memuat berbagai fitur medis pasien, antara lain*: Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver_Big, Liver_Firm, Spleen_Palpable, Spiders, Ascites, Varices, Bilirubin, Alk_Phosphate, Sgot, Albumin, Protime, Histology, Class*). Dataset Hepatitis berisi 155 data dengan 20 atribut, dan analisis dilakukan menggunakan perangkat lunak RapidMiner.

Prosedur metode yang digunakan dalam penelitian ini dijelaskan secara ringkas melalui diagram alur yang dapat dilihat pada Gambar 1..

Gambar 1. Alur penelitian



Data Understanding

Tahap Data Understanding meliputi pengumpulan dan pemeriksaan data untuk memperoleh pemahaman yang mendalam tentang data yang akan digunakan. Selain itu, pada tahap ini juga dilakukan identifikasi masalah dengan memahami isi data serta menemukan aspek-aspek penting yang dapat dijadikan dasar dalam merumuskan hipotesis awal (Ordila et al., 2020).

Data Pre-Processing

Dalam penelitian ini, proses pra-pemrosesan data dilakukan dengan beberapa teknik. Tahap awal adalah pembersihan data (data cleansing) yang bertujuan untuk mengevaluasi kualitas data. Proses ini meliputi penggabungan data, pendeteksian nilai yang hilang (missing values), penghapusan data duplikat, serta pengisian data yang kosong (data imputation) (Darwis, 2021). Tahap kedua adalah Label Encoding, yaitu proses mengonversi label data kategorikal menjadi bilangan bulat unik sesuai dengan urutan alfabet, menggunakan teknik pengkodean khusus (Ardiansyah, 2020). Tahap ketiga adalah *Feature Selection*, yaitu proses memilih sejumlah fitur tertentu dari keseluruhan fitur yang ada dalam dataset dengan tujuan meningkatkan efisiensi dan akurasi model (Rahmansyah et al., 2018).

Algoritma

Dalam penelitian ini, digunakan dua algoritma klasifikasi, yaitu *Logistic Regression* dan *Naïve Bayes.*

Logistic Regression

Probabilitas adalah kemungkinan terjadinya suatu peristiwa dari keseluruhan peristiwa dalam sebuah percobaan, dengan nilai yang selalu berada antara 0 dan 1. Sementara itu, odds didefinisikan sebagai perbandingan antara probabilitas terjadinya peristiwa dengan probabilitas peristiwa tersebut tidak terjadi (Harlan, 2018).

$$0(Y) = \frac{P(Y)}{1-P(Y)} \qquad (1)$$

Regresi logistik merupakan salah satu algoritma pembelajaran mesin yang paling populer setelah regresi linier. Kedua metode ini memiliki banyak kesamaan, tetapi perbedaan utama terletak pada tujuan penggunaannya. *Regresi linier* digunakan untuk memprediksi atau memperkirakan nilai numerik, sedangkan regresi logistik digunakan untuk melakukan klasifikasi. (Gunawan et al., 2020).

Naive Bayes

Klasifikasi Bayes adalah metode statistik yang memperkirakan probabilitas keanggotaan suatu data dalam sebuah kelas tertentu, misalnya probabilitas bahwa sebuah data termasuk dalam kelas tertentu. Pengelompokan ini didasarkan pada teorema Bayes. Pada algoritma Naïve Bayes, asumsi yang digunakan adalah bahwa nilai setiap atribut dalam suatu kelas bersifat independen satu sama lain. Berikut adalah rumus Naïve Bayes: (Zamri, 2022).

$$P(Y \mid X) = P(Y) \,||\, P(X \mid Y) \qquad (2)$$

Dimana:
P(Y | Y) : Probalitas data dengan vektor X pada kelas Y
P(Y) : Probalitas awal kelas Y dan P(Xi | Y) adalah probalitas independen kelas Y pada semua fitur dalam vektor X

Evaluasi

Pada tahap ini, evaluasi kinerja model algoritma yang digunakan dalam metode pembelajaran klasifikasi dilakukan dengan menggunakan algoritma *Logistic Regression* dan *Naïve Bayes*. Penilaian kinerja model klasifikasi didasarkan pada jumlah prediksi yang benar dan salah dalam mengelompokkan objek. Dalam penelitian ini, evaluasi model dilakukan menggunakan confusion matrix, yang menampilkan perbandingan antara hasil

klasifikasi sebenarnya dan prediksi dari model (Naufal, 2021).

Tabel 1. *Confusion matrix*

Prediksi		
Aktual		*Positive*
Negative		
Positive	*True Positive*	*False*
Positive Negative	*False Positive*	
True Negative		

Pada Tabel 1, TP (True Positive) menunjukkan jumlah prediksi positif yang tepat, TN (True Negative) adalah jumlah prediksi negatif yang benar, FP (False Positive) merupakan jumlah prediksi positif yang salah, dan FN (False Negative) adalah jumlah prediksi negatif yang salah. Dalam penelitian ini, metrik evaluasi performa klasifikasi yang digunakan meliputi accuracy, classification error, precision, recall, dan F1-score. Persamaan (3), (4), (5), (6) dan (7) secara berurutan memperlihatkan rumus perhitungan untuk accuracy, precision, recall, dan classification error (Naufal, 2021).

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

$$\text{Classificasion Error} = \frac{\text{FP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} = \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score\ =\ 2\ x\ \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

HASIL DAN PEMBAHASAN

Dataset

Data yang digunakan dalam penelitian ini berasal dari dataset hepatitis yang diunduh dari https://www.kaggle.com, dengan jumlah atribut sebanyak 20 (*Age*, *Sex*, *Steroid*, *Antivirals*, *Fatigue*, *Malaise*, *Anorexia*, *Liver_Big*, *Liver_Firm*, *Spleen_Palpable*, *Spiders*, *Ascites*, *Varices*, *Bilirubin*, *Alk_Phosphate*, *Sgot*, *Albumin*, *Protime*, *Histology*, *Class*), Jumlah instance dalam dataset ini adalah 155 data. Salah satu atribut pada dataset tersebut adalah label kelas yang terdiri dari kategori "live" dan "die". Data hepatitis kemudian dibagi menjadi dua bagian, yaitu 80% untuk data pelatihan *(training)* dan 20% untuk data pengujian *(testing)*.

Tabel 2. Dataset hepatitis

No	Elemen	Keterangan
1	Dataset	Hepatitis
2	Atribut	20
3	Type	*Int, Float, Bool*
4	Kelas	2
5	Total Data	155

Hasil Perhitungsn *Confusion Matrix*

Hasil perhitungan confusion matrix menggambarkan kinerja model dalam mengklasifikasikan data dengan akurat, termasuk jumlah prediksi yang benar dan salah untuk setiap kelas. Matriks ini terdiri dari empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), yang masing-masing merefleksikan hasil prediksi dibandingkan dengan data sebenarnya. Berdasarkan nilai-nilai ini, berbagai metrik evaluasi seperti akurasi, classification error, presisi, recall, dan F1-score dapat dihitung untuk menilai kualitas model secara menyeluruh.

Tabel 3. Hasil *confusion matrix* algoritma *Logistic Regression*

	true live	true die	class precision
pred. live	25	3	89.29%
pred. die	0	3	100.00%
class recall	100.00%	50.00%	

Sumber: Rapidminer

a. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ = 90,32%

b. $Classifikasi\ Error = \frac{FP+FN}{TP+TN+FP+FN}$ = 9,68%

c. $Precision = \frac{TP}{TP+FP}$ = 94,64%

d. $Recall = \frac{TP}{TP+FN}$ = 75,00%

e. $F1\ Score\ =\ 2\ x\ \frac{Precision * Recall}{Precision + Recall}$ = 83,68%

Tabel 4. Hasil *confusion matrix* algoritma *Naïve Bayes*

	true live	true die	class precision
pred. live	22	2	91.67%
pred. die	3	4	57.14%
class recall	88.00%	66.67%	

Sumber: Rapidminer

a. *Accuracy* = $\frac{TP+TN}{TP+TN+FP+FN}$ = 83,87%
b. *Classifikasi Error* = $\frac{FP+FN}{TP+TN+FP+FN}$ = 16,13%
c. *Precision* = $\frac{TP}{TP+FP}$ = 74,40%
d. *Recall* = $\frac{TP}{TP+FN}$ = 77,33%
e. *F1 Score* = $2\ x\ \frac{Precision * Recall}{Precisioon + Recall}$ = 75,78%

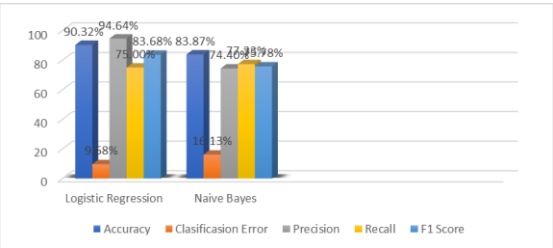
Hasil Akurasi

Algoritma *Logistic Regression* mencapai akurasi sebesar 90,32% dengan classification error 9,68%, precision 94,64%, recall 75,00% dan f1-score 83,68%. Di sisi lain, algoritma Naïve Bayes menunjukkan performa yang lebih baik dengan akurasi 83,87%, classification error 16,13%, precision 74,40%, recall 77,33% dan f1-score 75,78%. Hal ini menunjukkan bahwa *Logistic Regression* menunjukkan performa yang sedikit lebih unggul dalam membedakan antar kelas dalam datase,. namun, kelemahan *Naïve Bayes* terletak pada asumsi bahwa fitur-fitur saling independen, yang sering kali tidak sesuai dengan kondisi nyata, sehingga dapat menyebabkan penurunan akurasi dalam beberapa kasus.

Tabel 5. Hasil Perbandingan algoritma *Logistic Regression* dengan *Naïve Bayes*

	Dataset Hepatitis	
	Logistic Regression	**Naïve Bayes**
Accuracy	90,32%	83,87%
Clasificasion Error	9,68%	16,13%
Precision	94,64%	74,40%
Recall	75,00%	77,33%
F1-Score	83,68%	75,78%

Grafik perbandingan Klasifikasi *Naive Bayes* dan *Logistic Regression*:



Gambar 2. Grafik Tingkat Akurasi

KESIMPULAN

Penelitian ini berhasil melakukan evaluasi dan perbandingan kinerja algoritma *Logistic Regression* dan *Naïve Bayes* dalam mengklasifikasikan penyakit Hepatitis menggunakan dataset dari Kaggle.com. Hasil menunjukkan bahwa *Logistic Regression* memiliki performa yang lebih unggul dibandingkan *Naïve Bayes*, terutama dari segi akurasi dan tingkat kesalahan. Algoritma *Logistic Regression* mencapai akurasi sebesar 90,32% dengan tingkat klasifikasi error 9,68%, sementara *Naive Bayes* hanya memperoleh akurasi 83,87% dengan klasifikasi error sebesar 16,13%.

Hal ini mengindikasikan bahwa *Logistic Regression* memiliki performa yang sedikit lebih baik dalam membedakan kelas-kelas dalam dataset. Namun, *Naïve Bayes* memiliki kelemahan pada asumsi *independence* antar fitur, yang sering kali tidak sesuai dengan kondisi sebenarnya, sehingga dapat menurunkan *accuracy* dalam beberapa situasi.

Penelitian ini memberikan wawasan penting bagi praktisi medis dan peneliti dalam memilih metode klasifikasi yang tepat untuk diagnosis penyakit, terutama pada kondisi di mana akurasi sangat menentukan. Mengingat tingginya angka kejadian penyakit hepatitis dan dampaknya terhadap kesehatan masyarakat, penerapan algoritma machine learning seperti *Logistic Regression* dapat berperan signifikan dalam deteksi dini dan pengobatan yang lebih efektif.

Selain itu, penelitian ini menekankan pentingnya pemanfaatan teknik analisis data yang lebih maju dalam bidang kesehatan untuk meningkatkan hasil klinis. Sebagai saran untuk penelitian selanjutnya, disarankan menggunakan dataset yang

lebih besar dan beragam serta mengeksplorasi algoritma lain yang berpotensi meningkatkan akurasi dan keandalan dalam klasifikasi penyakit hepatitis.

DAFTAR PUSTAKA

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak, 15*(1), 131-145.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak, 15*(1), 131-145.

Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi prediksi penyakit diabetes mellitus menggunakan metode grid Seacrh Pada algoritma logistic regression. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *6*(3), 280-284.

Harlan, J. (2018). Analisa Regresi Logistik.

Kaunang, F. J. (2022). *A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms*. 8ISC Proceedings: Technology.

Li, T.-H.S., Chiu, H.-J., & Kuo, P.-H. (2022). *Hepatitis C Virus Detection Model by Using Random Forest, Logistic-Regression and ABC Algorithm*. IEEE Access.

Melinte-Popescu, A. S. et al. (2023). *Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity*. J Clin Med.

Menon, N. et al. (2022). *The Effect of Multiple Imputation of Routine Pathology Variables on Laboratory Diagnosis of Hepatitis C Infection*. arXiv preprint.

Naufal, M. F. (2021). Analisis Perbandingan Algoritma Svm, Knn, Dan Cnn untuk Klasifikasi Citra Cuaca. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *8*(2), 311-317.

Ordila, R., Wahyuni, R., Irawan, Y., & Sari, M. Y. (2020). Penerapan Data Mining Untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit Dengan Algoritma Clustering (Studi Kasus: Poli Klinik Pt. Inecda). *Jurnal Ilmu Komputer*, *9*(2), 148-153.

Rahmansyah, A., Dewi, O., Andini, P., Ningrum, T. H. P., & Suryana, M. E. (2018). Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*.

Zamri, D. (2022, September). Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan Algoritma Naïve Bayes dan KNN: Comparison of Data Mining Methods for Predition of Floods with Naïve Bayes and KNN Algorithm. In *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat* (pp. 40-48).