JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

PENGGUNAAN ANALYSIS OF VARIANCE PADA ALGORITMA NAÏVE BAYES DALAM PREDIKSI PENYAKIT HEPATITIS

Erras Lindiarda Mahentar*[1)], Yulison H. Chrisnanto [2)], Asep Id Hadiana [3)]
1. Teknik Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani, Indonesia
2. Teknik Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani, Indonesia
3. Teknik Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani, Indonesia

Article Info
Kata kunci: Anova, naïve bayes, Hepatitis

Keywords: Anova, naïve bayes, Hepatitis

Article history:
Received 24 April 2025
Revised 30 May 2025
Accepted 11 June 2025
Available online 1 March 2026

DOI :
https://doi.org/10.29100/jipi.v11i1.7717

* Corresponding author.
Corresponding Author
E-mail address:
erras.lindiarda@student.unjani.ac.id

ABSTRAK

Hepatitis merupakan penyakit serius yang memerlukan prediksi dini untuk mencegah komplikasi. Algoritma Naïve Bayes banyak digunakan dalam prediksi penyakit karena efisiensinya dalam menangani dataset yang besar dan sederhana. Namun, tantangan seperti ketidakseimbangan kelas, data yang tidak relevan, dan noise dalam dataset dapat mengurangi kinerja prediksi. Studi ini mengusulkan integrasi Analisis Varians (ANOVA) dengan Naïve Bayes untuk meningkatkan akurasi prediksi. ANOVA diterapkan sebagai metode pemilihan fitur berbasis filter untuk mengidentifikasi fitur yang paling signifikan, sehingga mengurangi efek noise dan fitur yang tidak relevan. Dataset yang digunakan dalam studi ini bersumber dari Repositori UCI, dengan validasi yang dilakukan menggunakan validasi silang 10 kali lipat. Hasilnya menunjukkan bahwa integrasi ANOVA meningkatkan akurasi prediksi Naïve Bayes dari 83,71% menjadi 85% sekaligus mengurangi jumlah fitur dari 19 menjadi lebih sedikit tanpa mengorbankan akurasi dan menghindari overfitting. Perangkat lunak yang dikembangkan dalam studi ini mendukung visualisasi dan prediksi data, yang selanjutnya memperkuat implementasi praktisnya. Penelitian ini berkontribusi pada optimalisasi prediksi penyakit hepatitis melalui pendekatan ANOVA. Penelitian selanjutnya disarankan untuk membandingkan metode ini dengan algoritma lain, seperti Decision Tree atau Support Vector Machine, dan mengujinya pada kumpulan data yang lebih besar dan lebih kompleks.

ABSTRACT

Hepatitis is a serious disease that requires early prediction to prevent complications. The Naïve Bayes algorithm is widely used in disease prediction due to its efficiency in handling large and simple datasets. However, challenges such as class imbalance, irrelevant data, and noise in the dataset can reduce prediction performance. This study proposes integrating Analysis of Variance (ANOVA) with Naïve Bayes to improve prediction accuracy. ANOVA is applied as a filter-based feature selection method to identify the most significant features, thereby reducing the effects of noise and irrelevant features. The dataset used in this study is sourced from the UCI Repository, with validation conducted using 10-fold cross-validation. The results indicate that the integration of ANOVA enhances the prediction accuracy of Naïve Bayes from 83.71% to 85% while reducing the number of features from 19 to fewer without sacrificing accuracy and avoiding overfitting. The software developed in this study supports data visualization and prediction, further strengthening its practical implementation. This research contributes to optimizing hepatitis disease prediction through the ANOVA approach. Future studies are suggested to compare this method with other algorithms, such as Decision Tree or Support Vector Machine, and to test it on larger and more complex datasets.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

I. PENDAHULUAN

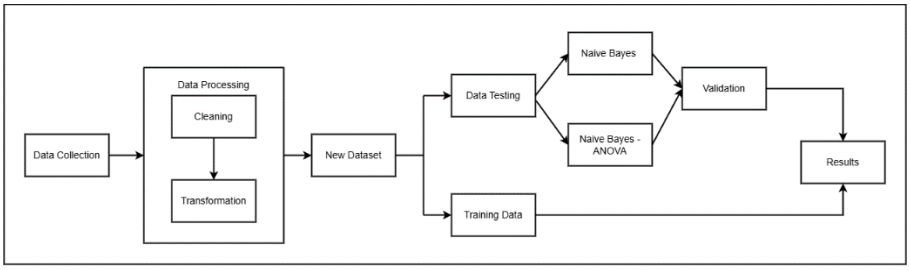
Hepatitis merupakan salah satu penyakit global yang paling menantang dengan konsekuensi serius bagi kesehatan masyarakat. Menurut WHO, sekitar 354 juta orang di dunia hidup dengan hepatitis B dan C, dengan lebih dari 1 juta kematian setiap tahun akibat komplikasi seperti sirosis dan kanker hati [1]. Hepatitis dapat berkembang menjadi kondisi kronis yang meningkatkan risiko komplikasi serius, sehingga prediksi dini dan akurat sangat penting [2]. Berbagai metode klasifikasi telah dikembangkan untuk meningkatkan akurasi prediksi, seperti decision tree, SVM, dan neural network [3]. Di antara metode tersebut, Naïve Bayes banyak digunakan dalam prediksi medis karena efisiensi komputasi, kemampuan menangani data berdimensi tinggi, serta asumsi sederhana namun efektif mengenai independensi fitur [4]. Naïve Bayes menunjukkan akurasi tinggi dalam klasifikasi penyakit liver, dengan keunggulan pada kecepatan pelatihan dan kemampuan generalisasi [5]. Kelebihan ini menjadikan Naïve Bayes relevan untuk prediksi hepatitis, yang datanya cenderung kompleks, bervariasi, dan tidak selalu lengkap [6]. Namun, prediksi hepatitis menghadapi tantangan statistik yang memengaruhi akurasi, salah satunya adalah ketidakseimbangan kelas, di mana jumlah kasus positif jauh lebih sedikit dibandingkan kasus negatif [7][8]. Ketidakseimbangan ini dapat menyebabkan model menjadi bias terhadap kelas mayoritas, sehingga mengurangi sensitivitas dalam mendeteksi kasus positif yang sangat krusial untuk pencegahan dan intervensi dini. Selain itu, keberadaan outlier dan noise dalam data dapat semakin menghambat kinerja model dengan memperkenalkan variabilitas yang tidak diinginkan. Jika tidak ditangani dengan tepat, hal ini dapat membahayakan integritas analisis secara keseluruhan [9]. Naïve Bayes, sebagai metode probabilistik yang banyak digunakan dalam klasifikasi biner dan multikelas, dikenal efektif menangani data berdimensi tinggi dengan komputasi efisien dan akurasi yang cukup baik [10]. Naïve Bayes dikenal efektif dalam aplikasi medis karena mampu menangani data berdimensi tinggi dengan akurasi cukup baik dan komputasi efisien. Namun, nilai hilang yang tidak ditangani, serta outlier dan noise dalam data medis, dapat menurunkan integritas, validitas, dan kestabilan kinerja model [11]. Naïve Bayes banyak digunakan dalam klasifikasi biner dan multikelas, termasuk prediksi penyakit kronis seperti hepatitis, namun tetap memerlukan pendekatan lebih mendalam untuk mengatasi variabilitas data dan ketidakseimbangan kelas guna meningkatkan kinerja model [12]. Untuk memaksimalkan potensi Naïve Bayes pada data yang tidak ideal, dibutuhkan pendekatan pemilihan fitur yang relevan, salah satunya melalui Analisis Varians (ANOVA) yang membantu mengidentifikasi fitur dengan pengaruh signifikan terhadap kelas target sebagai strategi feature selection yang terstruktur [13]. Penelitian [14] menunjukkan bahwa kombinasi ANOVA dan Naïve Bayes terbukti dapat meningkatkan akurasi klasifikasi hingga 6% pada diagnosis kanker payudara dibandingkan Naïve Bayes tanpa optimasi fitur. Namun, pendekatan serupa masih jarang diterapkan dalam prediksi hepatitis, padahal berpotensi besar dalam mengatasi ketidakseimbangan kelas dan mengurangi fitur tidak relevan. Penelitian ini bertujuan mengintegrasikan ANOVA ke dalam Naïve Bayes untuk mengoptimalkan pemilihan fitur dan meningkatkan akurasi prediksi hepatitis. Dengan mengeksplorasi penggunaan ANOVA [15] untuk menilai pengaruh fitur terhadap kelas target, penelitian ini berupaya meningkatkan kinerja Naïve Bayes dalam mendeteksi penyakit secara lebih akurat dan efisien. Dalam penelitian [16], dilakukan perbandingan antara efektivitas algoritma Naïve Bayes dan C4.5 dalam memprediksi hepatitis. Hasilnya menunjukkan bahwa algoritma Naïve Bayes mencapai akurasi yang lebih tinggi, yaitu 83,71% dibandingkan dengan C4.5 yang mencapai 77,29%. Akan tetapi, penelitian ini tidak membahas masalah seperti ketidakseimbangan kelas, nilai yang hilang, atau fitur yang tidak relevan dalam kumpulan data. Penelitian [17] menggabungkan metode Naïve Bayes Naïve Bayes dengan pemilihan fitur berbasis wrapper menunjukkan peningkatan akurasi hingga 88% pada data kesehatan. Namun, diperlukan evaluasi lanjutan untuk menguji stabilitas model terhadap fitur yang beragam dan penerapannya di dunia nyata. Tantangan seperti keterbatasan akses data pasien dan integrasi ke sistem rekam medis elektronik menuntut model prediksi yang tidak hanya akurat, tetapi juga ringan, mudah diintegrasikan, dan sesuai dengan praktik klinis yang ada. Penelitian [18] mengeksplorasi penggunaan Regresi Logistik dan Naïve Bayes untuk memprediksi tahap awal penyakit hati berlemak, dengan hasil menunjukkan kedua algoritma mencapai akurasi baik. Namun, Naïve Bayes unggul dalam kecepatan komputasi dan kemampuan menangani data tidak seimbang. Meskipun demikian [19], penelitian ini memiliki keterbatasan seperti kurangnya analisis mendalam dari variabel yang digunakan dan tidak adanya teknik tambahan seperti Analisis Varians (ANOVA) untuk meningkatkan kinerja model. Dalam penelitian ini, dengan menggabungkan Naïve Bayes dan pemilihan fitur ANOVA untuk meningkatkan kinerja prediksi hepatitis, dengan Naïve Bayes sebagai model klasifikasi dan ANOVA untuk memilih fitur yang paling signifikan secara statistik [20]. Dengan demikian, subset fitur yang dipilih diharapkan dapat mengoptimalkan kinerja model Naïve Bayes. Penggunaan Matriks Konfusi [21] sebagai metrik evaluasi diharapkan memberikan pemahaman menyeluruh terhadap performa model, terutama dalam menangani ketidakseimbangan kelas. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem pendukung keputusan medis (clinical decision support system) yang lebih andal dan dapat diterapkan di lingkungan kesehatan yang riil.

.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

II. METODE PENELITIAN

Pada bagian ini, Gambar 1 menyajikan metodologi penelitian, dimulai dari pengumpulan data. Data kemudian menjalani tahap pra-pemrosesan untuk membersihkan dan mengubahnya menjadi format yang sesuai untuk analisis. Selanjutnya, data dianalisis dan dikategorikan berdasarkan variabel tertentu untuk membangun model prediksi penyakit hepatitis.



Gambar 1. Metode Penelitian

.

A. Diskripsi Atribut Data

Langkah awal sebelum pembersihan data melibatkan eksplorasi data. Pada tahap ini, setiap atribut dalam kumpulan data menjalani analisis komprehensif untuk memahami karakteristik, distribusi, dan relevansinya dengan variabel target. Wawasan yang diperoleh dari analisis ini akan memandu pemilihan atribut yang akan disertakan dalam model prediktif.

TABLE I
DESKRIPSI ATRIBUT DATA

No	Attribute	Initialization	Explanation
1.	Class	A	Kategori atau kelas hasil diagnostik, seperti adanya penyakit tertentu atau kategori yang menunjukkan kondisi kesehatan pasien.
2.	Age	B	Usia pasien dalam tahun, yang berfungsi sebagai variabel demografi dalam analisis kesehatan.
3.	Sex	C	Jenis kelamin pasien, biasanya dikodekan sebagai 1 untuk pria dan 2 untuk wanita.
4.	Steroid	D	Penggunaan steroid oleh pasien. Nilai ini menunjukkan apakah pasien menjalani perawatan dengan steroid.
5.	Antiviral	E	Penggunaan obat antivirus. Nilai ini menunjukkan apakah pasien menerima pengobatan antivirus.
6.	Fatigue	F	Tingkat kelelahan yang dirasakan pasien.
7.	Malaise	G	Ketidaknyamanan dan kelemahan yang dialami pasien sering dikaitkan dengan gejala penyakit tertentu.
8.	Anorexia	H	Pasien mengalami kehilangan nafsu makan.
9.	Liver Big	I	Pembesaran hati (hepatomegali), menunjukkan apakah hati pasien lebih besar dari normal.
10.	Liver Firm	J	Kekakuan hati, menunjukkan apakah hati terasa lebih keras saat diperiksa secara fisik.
11.	Spleen Palpabel	K	Kemampuan untuk merasakan limpa melalui pemeriksaan fisik biasanya menunjukkan limpa yang membesar.
12.	Spiders	L	Kehadiran angioma laba-laba (pembuluh darah kecil yang terlihat pada kulit), sering dikaitkan dengan gangguan fungsi hati.
13.	Ascites	M	Adanya cairan di rongga perut, yang biasanya disebabkan oleh penyakit hati atau gagal jantung.
14.	Varices	N	Pembesaran pembuluh darah vena (varises) di area tertentu, biasanya di kerongkongan atau lambung, terkait dengan tekanan darah tinggi di vena porta akibat penyakit hati..
15.	Bilirubin	O	Kadar bilirubin dalam darah, zat yang terbentuk dari pemecahan sel darah merah. Kadar yang tinggi dapat mengindikasikan masalah hati.
16.	Alk Phosphate	P	Tingkat enzim alkali fosfatase dalam darah, yang mungkin meningkat jika ada kerusakan hati atau penyakit tulang.
17.	SGOT	Q	Enzim hati yang dilepaskan ke dalam darah ketika terjadi kerusakan hati atau jantung.
18.	Albumin	R	Kadar protein albumin dalam darah. Kadar albumin yang rendah dapat mengindikasikan masalah pada hati atau ginjal.
19.	Protime	S	Waktu yang dibutuhkan darah untuk membeku. Waktu yang lebih lama dapat mengindikasikan gangguan fungsi hati.
20.	Histology	T	Hasil pemeriksaan histologi atau biopsi jaringan hati untuk mengidentifikasi adanya penyakit atau perubahan patologis.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

B. *Dataset Penyakit Hepatitis*

Kumpulan data yang digunakan dalam penelitian ini diperoleh dari Kaggle Hepatitis C Prediction Dataset, yang terdiri dari 20 atribut. Atribut yang digunakan dalam penelitian ini meliputi A, B, C, D, E, ..., Q, R, S, dan T.

TABLE II
DATASET PENYAKIT HEPATITIS

A	*B*	*C*	*D*	*E*	…	*Q*	*R*	*S*	*T*
No	30	2	1	2	…	18	4	60	1
No	50	1	1	2	…	42	3.5	100	1
No	78	1	2	2	…	32	4	100	1
Yes	31	1	?	1	…	52	4	80	1
Yes	34	1	2	2	…	200	4	100	1
No	34	1	2	2	…	28	4	75	1
Yes	51	1	1	2	…	22	4	100	1
Yes	23	1	2	2	…	45	4.9	100	1
No	39	1	2	2	…	48	4.4	100	1
Yes	30	1	2	2	…	120	3.9	100	1
Yes	39	1	1	1	…	30	4.4	85	1
Yes	32	1	2	1	…	249	3.7	54	1
No	41	1	2	1	…	60	3.9	52	1
Yes	30	1	2	2	…	144	4.9	78	1
No	47	1	1	1	…	60	4	100	1

C. *Pra-Pemprosesan*

Prapemrosesan data merupakan langkah penting sebelum analisis lebih lanjut. Tujuan utamanya adalah untuk meningkatkan kualitas data dan memastikan kesesuaiannya dengan model yang diinginkan. Prapemrosesan melibatkan pembersihan data, seperti penanganan nilai yang hilang, outlier, dan inkonsistensi, serta transformasi data, termasuk normalisasi, standardisasi, dan konversi tipe data.

1) Pembersihan Data

Pembersihan data merupakan langkah penting dalam memastikan kualitas data untuk analisis. Dalam studi ini, pembersihan data melibatkan identifikasi nilai yang hilang dalam kumpulan data. Setelah memeriksa kumpulan data Kaggle Hepatitis, beberapa titik data yang hilang terdeteksi. Dengan mengidentifikasi dan menangani nilai yang hilang ini, kualitas data untuk analisis ditingkatkan, menghasilkan hasil yang lebih akurat dan andal.

2) Transformasi Data

Transformasi data dilakukan untuk menyiapkan kumpulan data untuk analisis yang lebih efisien. Proses ini melibatkan konversi variabel, termasuk 'Kelas', 'Usia', 'Jenis Kelamin', dan lainnya, ke dalam format numerik. Karena atribut 'Kelas' bersifat kategoris, yang mewakili ada atau tidaknya hepatitis, atribut tersebut dikodekan secara numerik. Nilai 2 ditetapkan untuk 'Ya' (menunjukkan hepatitis) dan 1 untuk 'Tidak'. Kumpulan data numerik yang dihasilkan selanjutnya digunakan untuk prosedur analisis lebih lanjut.

TABLE III
DATA TRANSFORMASI

A	*B*	*C*	*D*	*E*	…	*Q*	*R*	*S*	*T*
1	30	2	1	2	…	18	4	60	1
1	50	1	1	2	…	42	3.5	100	1
1	78	1	2	2	…	32	4	100	1
2	31	1	?	1	…	52	4	80	1
2	34	1	2	2	…	200	4	100	1
1	34	1	2	2	…	28	4	75	1
2	51	1	1	2	…	22	4	100	1
2	23	1	2	2	…	45	4.9	100	1
1	39	1	2	2	…	48	4.4	100	1
2	30	1	2	2	…	120	3.9	100	1
2	39	1	1	1	…	30	4.4	85	1
2	32	1	2	1	…	249	3.7	54	1
1	41	1	2	1	…	60	3.9	52	1
2	30	1	2	2	…	144	4.9	78	1
1	47	1	1	1	…	60	4	100	1

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

D. Proses Prediksi Penyakit Hepatitis

Data tersebut dianalisis dan dikelompokkan berdasarkan variabel yang terlibat. Setelah eksplorasi data, model yang sesuai dipilih dengan mempertimbangkan sifat data. Set data dibagi menjadi set pelatihan dan pengujian untuk memfasilitasi pengembangan model. Teknik validasi silang 10 kali lipat digunakan untuk pelatihan. Selanjutnya, kinerja model dinilai menggunakan pengklasifikasi Naive Bayes dan pengklasifikasi Naive Bayes yang dikombinasikan dengan ANOVA. Sebelum fase prediksi, set data yang telah diproses sebelumnya akan dibagi menjadi dua subset terpisah: set pelatihan dan set pengujian. Set pelatihan akan mencakup 80% data, dan 20% sisanya akan dialokasikan ke set pengujian.

1) Analysis of Variance

Pada tahap ini dilakukan pemilihan data terhadap atribut-atribut yang akan digunakan menggunakan Analysis of Variance (ANOVA) untuk mendapatkan atribut-atribut yang lebih informatif sehingga dapat meningkatkan akurasi dan efisiensi algoritma Naïve Bayes. Proses ini diawali dengan menghitung nilai varians dari seluruh atribut yang ada pada klasifikasi. Berikut ini adalah proses perhitungan manual Analysis of Variance untuk prediksi penyakit hepatitis yang berisi 20 atribut. Rumus perhitungan varians antar kelompok (Sum of Squares Between, SSB) dan varians dalam kelompok (Sum of Squares Within, SSW).

$$SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(x - \bar{x})^2 \quad (1)$$

$$SSB = \sum_{j=1}^{k}(\bar{x}_j - \bar{x})^2 \quad (2)$$

Untuk menghitung nilai varians setiap atribut dalam dataset hepatitis, nilai rata-rata untuk setiap kelas Ya dan Tidak dihitung terlebih dahulu, misalnya dalam atribut Q(SGOT). Di sini kita ingin menentukan apakah ada perbedaan yang signifikan dalam distribusi nilai rata-rata antara kedua kelompok.

Kelas rata-rata (Q) (Ya)

$$\bar{x}_1 = \frac{18+42+32+28+48+60+60}{7} = \frac{288}{7} = 41{,}14 \quad (3)$$

Kelas rata-rata (Q) (Tidak)

$$\bar{x}_2 = \frac{52+200+22+45+120+30+249+144}{8} = \frac{862}{8} = 107{,}75 \quad (4)$$

Rata-rata Total (Q)

$$\bar{x}_{total} = \frac{18+42+32+28+48+60+60+52+200+22+45+120+30+249+144}{15} = \frac{1.150}{15} = 76{,}66 \quad (5)$$

Setelah membagi data menjadi dua kelompok, kita akan menghitung Sum of Squares Between (SSB) yang mengukur variasi antar kelompok. Hal ini dilakukan dengan menghitung selisih antara rata-rata kelompok dan rata-rata keseluruhan untuk setiap fitur yang dianalisis, lalu menjumlahkannya. Di sisi lain, kita juga menghitung Sum of Squares Within (SSW), yang merupakan variabilitas dalam setiap kelompok, yang diperoleh dengan menjumlahkan selisih antara nilai individual dan rata-rata kelompoknya.

Hitung Inter-Class Sum of Squares (SSB)

$$SSB = 7\,(41{,}14 - 76{,}66) + 8\,(107{,}75 - 76{,}66) \quad (6)$$

$$SSB = 7\,(-35{,}52)^2 + 8(31{,}09)^2 \quad (7)$$

$$SSB = 8{,}831 + 7{,}732 = 16{,}563 \quad (8)$$

Hitung Jumlah Kuadrat Dalam Kelas (SSW)

For Class 1	: $(18\text{-}41.14)^2+(42\text{-}41.14)^2+(32\text{-}41.14)^2+\ldots+(60\text{-}41.14)^2$
Result	: 535.45+0.739+83.53+…+355.69
Total SSW	: Class 1 ≈ 2812.11
For class 2	: $(52\text{-}107.75)^2+(200\text{-}107.75)^2+(120\text{-}107.75)^2+\ldots+(144\text{-}107.75)^2$
Result	: 3.108+8.510+150.60+…+1.314
Total SSW	: Class 2 ≈29.992
Total SSW	: SSW = 2812.11+29.992 = 2.842
Calculate Degrees of Freedom (df)	: df between classes: k – 1 = 2-1 = 1 : df within classes: N – k = 15 – 2 = 13

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

Setelah SSB dan SSW dihitung, kita dapat menentukan nilai F dengan membandingkan between-group variation (SSB) dengan within-group variation (SSW) melalui Mean Square Between (MSB) dan Mean Square Within (MSW). Nilai F ini akan menunjukkan apakah ada perbedaan yang signifikan antara kedua kelompok Hepatitis Ya dan Hepatitis Tidak pada atribut yang diuji, seperti Jenis Kelamin, Usia, atau SGOT.

Mean Square Between

$$MSB = \frac{SSB}{df_{between}} = \frac{16.563}{1} = 16.563 \quad (9)$$

Mean Square Within

$$MSW = \frac{SSW}{df_{between}} = \frac{2.842}{13} = 218.61 \quad (10)$$

Hitung Nilai F

$$F = \frac{MSB}{MSW} = \frac{16.563}{218.61} = 75{,}76 \quad (11)$$

Nilai F sebesar 75,76 umumnya dianggap sangat tinggi dalam konteks ANOVA. Nilai F yang tinggi menunjukkan bahwa variabilitas antarkelompok jauh lebih besar daripada variabilitas dalam kelompok. Ini menyiratkan perbedaan yang signifikan antara kelompok yang dibandingkan. Untuk menentukan apakah nilai ini signifikan, kita perlu membandingkannya dengan nilai F kritis dari tabel distribusi F dengan derajat kebebasan antarkelompok (1) dan derajat kebebasan dalam kelompok (13) pada tingkat signifikansi tertentu (misalnya, 0,05). Jika nilai F yang dihitung kurang dari nilai kritis, kita tidak memiliki cukup bukti untuk menolak hipotesis nol, yang berarti tidak ada perbedaan yang signifikan dalam tekanan darah rata-rata (Q) di antara kelompok. Oleh karena itu, fitur 'Q' tidak signifikan untuk klasifikasi dalam kumpulan data ini. Dari proses perhitungan, hasil Varians diperoleh untuk semua atribut dalam sampel kumpulan data hepatitis.

TABLE IV
SAMPLE DATASET HEPATITIS

No.	Attribute	ANOVA
1.	Age	2.21
2.	Sex	0.22
3.	Steroid	1.19
4.	Antiviral	2.46
5.	Fatigue	0.40
6.	Malaise	1.40
7.	Anorexia	0.21
8.	Liver Big	0.12
9.	Liver Firm	23.71
10.	Spleen Palpabel	774.22
11.	Spiders	33.33
12.	Ascites	196.27
13.	Varices	17.03
14.	Bilirubin	52.03
15.	Alk Phosphate	59.06
16.	SGOT	75.76
17.	Albumin	0.12
18.	Protime	0.24
19.	Histology	0.00

2) Naïve Bayes

Setelah pemilihan fitur optimal melalui analisis varians (ANOVA), model Naive Bayes akan dibangun. Model ini selanjutnya akan digunakan untuk membuat prediksi pada data baru. Perhitungan Naive Bayes akan dilakukan hanya dengan mempertimbangkan fitur-fitur yang telah melewati tahap pemilihan ANOVA. Untuk memberi contoh perhitungan Naive Bayes manual, mari kita pertimbangkan contoh kumpulan data.

TABLE V
SAMPLE DATA

C	*D*	*F*	*G*	*H*	*A*
2	2	1	1	1	1
1	1	1	2	2	2
2	2	1	1	1	2
1	2	1	1	1	1
1	2	1	1	1	2
2	2	1	1	1	1
1	1	1	1	1	2
2	2	1	1	1	1
1	1	1	1	2	1
2	1	1	1	1	2

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

C	*D*	*F*	*G*	*H*	*A*
1	2	1	1	1	1
1	2	1	1	2	1
1	2	1	2	2	1
1	1	2	2	1	1
2	1	2	2	1	1

Perhitungan probabilitas dibagi menjadi 2 yaitu:

a. Probabilitas prior/kelas dan
b. Probabilitas kondisi.

TABLE VI
PERHITUNGAN PROBABILITAS

Classification	*Frekuensi*	*Probabilitas*
C(Yes)	6	3
C(No)	4	2

P(Classification = yes) = 0.67
P(Classification = no) = 0.33

Perhitungan manual Naïve Bayes untuk probabilitas kondisi pada setiap atribut, berikut ini adalah salah satu atribut yang dihitung yaitu pada htm, lakukan perhitungan pada semua atribut.

P(C = Yes|no) = 2/5 = 0.4
P(C = Yes|yes) = 4/10 = 0.4
P(C = No|no) = 3/5 = 0.6
P(= No|ya) = 6/10 = 0.6

Setelah mendapatkan nilai probabilitas sebelumnya dan kondisi yang dibutuhkan, langkah klasifikasi dalam Perhitungan Manual Naive Bayes adalah sebagai berikut. Misalnya, untuk record tertentu, dilakukan sebagai berikut htm = 1, dm = 1, cad = 2, appet = 1, pe = 2, Klasifikasi apa yang akan diklasifikasikan? YA atau TIDAK?

a. P(Classification=y|x) = ((P(C= Yes|yes) * (P(D= false|yes) * P(F = true|yes) * P(G = false|yes) * P(H = true|yes)) * P(Classification|yes)
b. P(Classification=y|x) = (0.4*0.3*0.2*0.7*0.3)*(0.67) = 0.011256
c. P(Classification=n|x) = ((P(C= Yes|no) * (P(D= false|no) * P(F = true|no) * P(G = false|no) * P(H = true|no)) * P(Classificationj|no)
d. P(Classification=y|x) = (0.4*0.6*0.8*0.2)*(0.33) = 0.012672
e. Dari hasil perhitungan dapat diketahui jika record C = 1, D = 2, F = 1, G = 1, H = 1, maka akan diklasifikasikan pada Klasifikasi No.

III. HASIL DAN PEMBAHASAN

Percobaan yang dilakukan adalah uji akurasi komparatif menggunakan Anova dan tidak menggunakan Anova pada data Training. Pengujian ini dilakukan dengan menggunakan 560 dataset dengan 3 rasio data training dan data test yang berbeda, yaitu rasio 90:10 (504 data training dan 56 data test), rasio 80:20 (448 data training dan 112 data test), 70:30 (392 data training dan 164 data test). Dengan penggunaan rasio-rasio ini memberikan variasi dalam ukuran data pelatihan dan pengujian, yang memungkinkan peneliti untuk mengevaluasi kinerja model dalam berbagai kondisi. Rasio tersebut dipilih untuk menyeimbangkan antara pelatihan yang cukup dan evaluasi yang efektif, sekaligus memastikan model tidak overfit atau underfit.

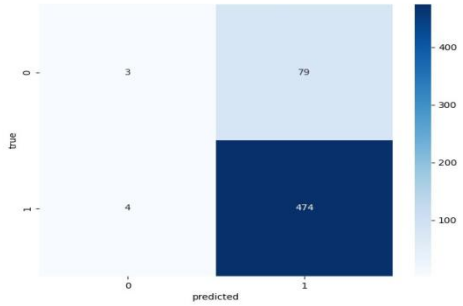
TABLE VII
HASIL PENGUJIAN

	Menggunakan Anova		
Jumlah Fitur	***Rasio 90:10 (360)***	***Rasio 80:20 (320)***	***Rasio 70:30 (280)***
	Akurasi (%)	***Akurasi (%)***	***Akurasi(%)***
10	83	84	85
	Tidak Menggunakan Anova		
jumlah Fitur	***Rasio 90:10 (360)***	***Rasio 80:20 (320)***	***Rasio 75:25 (280)***
	Akurasi (%)	***Akurasi (%)***	***Akurasi (%)***
19	83	83	83

Ketika ANOVA digunakan untuk pemilihan fitur, jumlah fitur berkurang dari 19 menjadi 10. Meskipun terjadi pengurangan fitur, akurasi model justru meningkat secara signifikan. Pada rasio split data 80:20, dan 70:30, akurasi model menggunakan ANOVA lebih tinggi, yakni 84% hingga 85%. Hal ini menunjukkan bahwa dengan

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.327-336

menggunakan ANOVA, fitur-fitur yang tidak relevan berhasil dihilangkan sehingga model menjadi lebih efisien dan akurat. Pengujian akurasi merupakan suatu proses untuk menguji tingkat akurasi hasil yang diperoleh dari suatu model yang digunakan. Pengujian ini melibatkan penggunaan metode chaotic matrix. Pengujian hasil prediksi yang dilakukan oleh model Naïve Bayes dan Anova pada data Test menggunakan metode chaotic matrix untuk menunjukkan bahwa metode ini lebih sensitif dalam mengukur ketidakpastian dan menangkap dinamika nonlinear dalam data. Dibandingkan dengan metode evaluasi tradisional, chaotic matrix memberikan pemahaman yang lebih mendalam, menjadikannya alat yang lebih akurat dan robust dalam analisis prediksi di bidang kesehatan.



Gambar 2 Confusion Matrix

Perhitungan menggunakan metode Naïve Bayes akan dilakukan dengan mempertimbangkan fitur-fitur yang telah lolos tahap seleksi ANOVA. Untuk mengilustrasikan perhitungan manual Naïve Bayes, mari kita ambil contoh data. Setelah itu, akurasi model dapat dievaluasi menggunakan rumus.

$$Akurasi = \frac{3+474}{3+474+79+4} \text{ x } 100\% = 85\% \quad (13)$$

Perhitungan berikutnya, yaitu presisi, akan dilakukan untuk mengukur seberapa akurat model dalam mengklasifikasikan sampel positif sebagai benar. Presisi dihitung dengan membagi jumlah positif benar dengan jumlah prediksi positif.

$$Precision = \frac{3}{3+79} \text{ x } 100\% = 36\% \quad (14)$$

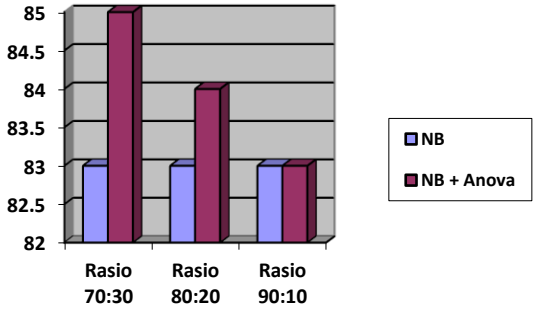
Perhitungan penarikan kembali dalam pengambilan informasi mengukur proporsi dokumen relevan yang berhasil diambil dari semua dokumen relevan dalam koleksi.

$$Recall = \frac{3}{3+7} \text{ x } 100\% = 42\% \quad (15)$$

Skor F1 adalah metrik evaluasi yang menggabungkan presisi dan perolehan kembali. Metrik ini memberikan nilai tunggal yang menunjukkan keseimbangan antara kemampuan model untuk mengidentifikasi contoh positif yang sebenarnya (presisi) dan kemampuan model untuk menemukan semua contoh positif yang sebenarnya (penarikan kembali). Rumus skor F1 adalah rata-rata harmonik tertimbang dari presisi dan perolehan kembali.

$$F1\ Score = \frac{2 \times 36 \times 42}{36+42} \text{ x } 100\% = 38\% \quad (16)$$

Berdasarkan penelitian dalam jurnal tersebut, akurasi prediksi menggunakan algoritma Naïve Bayes meningkat dari 83% menjadi 85% setelah seleksi fitur dengan ANOVA. Meskipun peningkatan ini terlihat kecil, dalam konteks medis, setiap peningkatan akurasi dapat berpengaruh pada pengambilan keputusan yang lebih tepat dan mengurangi risiko kesalahan diagnosis, terutama dalam kondisi kritis. Namun, peningkatan akurasi harus dilihat bersama dengan metrik lain seperti presisi, recall, dan F1 score. Untuk memastikan dampak peningkatan akurasi dalam pengambilan keputusan klinis, penting untuk memeriksa apakah metrik-metrik ini juga mengalami peningkatan signifikan. Oleh karena itu, evaluasi kinerja model harus mempertimbangkan semua metrik secara komprehensif.



Gambar 3 Hasil Akurasi Naive Bayes

A. DISKUSI

Dari hasil perbandingan pada tabel dan grafik terlihat bahwa masing-masing rasio pembagian data (90:10, 80:20, dan 70:30) menunjukkan peningkatan akurasi ketika diterapkan ANOVA, dari 83% menjadi 84–85%. Hal ini menunjukkan bahwa dengan dihilangkannya fitur-fitur yang kurang signifikan, model dapat fokus pada fitur-fitur yang relevan, sehingga Naïve Bayes dapat memperoleh hasil yang lebih optimal dalam klasifikasi penyakit hepatitis. Pemilihan fitur ANOVA juga membantu model untuk menghindari overfitting, yaitu kondisi ketika model terlalu banyak menyesuaikan diri dengan data latih hingga sulit bekerja dengan baik pada data uji. Meskipun ANOVA berhasil mengurangi jumlah fitur dari 19 menjadi 10, penting untuk memastikan bahwa fitur yang dipertahankan memiliki relevansi klinis yang tinggi. Fitur yang dipilih harus mencerminkan variabel-variabel yang benar-benar mempengaruhi prognosis atau prediksi penyakit, seperti hepatitis. Misalnya, faktor-faktor seperti usia, kadar enzim hati, atau riwayat medis dapat memiliki dampak langsung pada diagnosis, sementara fitur lain yang kurang relevan bisa saja hanya memperumit model tanpa memberikan kontribusi berarti. Oleh karena itu, pemilihan fitur yang tepat sangat penting agar model tetap memberikan hasil yang akurat dan berbasis bukti klinis. Dengan fitur yang lebih sedikit, Naïve Bayes bekerja lebih efektif dengan data yang lebih terfokus pada atribut-atribut yang relevan, menghindari overfitting, dan meningkatkan akurasi prediksi. Model ini memiliki potensi besar untuk diterapkan dalam lingkungan klinis nyata karena kesederhanaan dan efisiensinya, yang memungkinkan prediksi yang lebih cepat dan dapat diandalkan meskipun menggunakan data baru dengan variabilitas. Namun, penting untuk terus memantau kinerja model dengan data baru untuk memastikan keandalannya dalam kondisi klinis yang berbeda.

IV. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa integrasi ANOVA sebagai teknik seleksi fitur dapat meningkatkan kinerja algoritma Naïve Bayes dalam memprediksi hepatitis, dengan akurasi prediksi yang meningkat hingga 85%. Ini memperkuat temuan dari penelitian sebelumnya yang menunjukkan efektivitas Naïve Bayes dalam memprediksi penyakit hepatitis. Penelitian Anda lebih lanjut memperkenalkan inovasi dengan penggunaan ANOVA untuk mengurangi fitur yang tidak relevan, yang meningkatkan keakuratan model dan mengurangi kompleksitasnya. Selain itu, perangkat lunak yang diimplementasikan dalam penelitian ini memberikan kemudahan praktis dalam mengelola dan memprediksi data medis, yang berpotensi memberikan manfaat nyata bagi tenaga medis dalam melakukan deteksi dini hepatitis. Penelitian selanjutnya dapat mengembangkan model ini dengan menggunakan algoritma klasifikasi lain, seperti Decision Tree atau Support Vector Machine, untuk perbandingan lebih lanjut. Selain itu, penggunaan dataset yang lebih besar dan variasi parameter dalam metode ANOVA mungkin dapat menghasilkan peningkatan akurasi yang lebih signifikan. Integrasi model ini dengan data real-time juga disarankan untuk memberikan hasil prediksi.

V. DAFTAR PUSTAKA

[1] G. Health, S. Strategy, T. Ending, and V. Hepatitis, “Viral hepatitis 2016–2021,” no. June 2016, 2021.
[2] W. J. Gastroenterol, “Contents,” vol. 9327, no. 1, 2022.
[3] L. Zhang, J. Wang, R. Chang, and W. Wang, “Investigation of the effectiveness of a classification method based on improved DAE feature extraction for hepatitis C prediction,” *Sci. Rep.*, no. 1, pp. 1–12, 2024, doi: 10.1038/s41598-024-59785-y.
[4] V. Jansari and R. Pears, “A Positive-Confidence based approach to Classifying Imbalanced data : A Case Study on Hepatitis,” pp. 457–461, 2020, doi: 10.34297/AJBSR.2020.08.001319.Received.
[5] I. Kurniawan *et al.*, “Prediksi Kemungkinan Penyakit Liver menggunakan Algoritma Klasifikasi Naive Bayes,” vol. 15, no. 3, pp. 506–514, 2024.
[6] P. C. Austin, I. R. White, D. S. Lee, and S. Van Buuren, “Missing Data in Clinical Research : A Tutorial on Multiple Imputation,” *Can. J. Cardiol.*, vol. 37, no. 9, pp. 1322–1331, 2021, doi: 10.1016/j.cjca.2020.11.010.

[7] M. A. G. Hazber, E. M. Senan, and H. S. Alrashidi, "Feature Engineering Methods for Analyzing Blood Samples for Early Diagnosis of Hepatitis Using Machine Learning Approaches," 2025, doi: 10.32604/cmes.2025.062302.
[8] K. Gohari, A. Kazemnejad, M. Mohammadi, F. Eskandari, and S. Saberi, "A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients," *BMC Med. Res. Methodol.*, pp. 1–15, 2023, doi: 10.1186/s12874-023-02013-4.
[9] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets : review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.
[10] E. V. Study, "Clinical Validity of a Machine Learning Decision Support System for Early Detection of Hepatitis B Virus : A Binational External Validation Study," 2023.
[11] A. Statistics, M. Sciences, H. Education, and N. Cairo, "Handling Outlier Data as Missing Values by Imputation Methods : Application of Muhammad Metwally Seliem Email address :," pp. 273–286, 2020.
[12] M. Sayadi, V. Varadarajan, and E. Gozali, "Frontiers in Health Informatics Effective factors in diagnosing the degree of hepatitis C using machine learning," 2023.
[13] J. C. G, G. Suganya, M. Premalatha, A. Victy, Y. Phamila, and K. Karunamurthy, "Machine Learning and Deep Learning Methods for Building Intelligent Systems in," 2021.
[14] P. B. Cancer, "Predict Breast Cancer," 2023.
[15] R. Wibowo, "Top-k Feature Selection untuk Deteksi Penyakit Hepatitis Menggunakan Algoritme Naïve Bayes," pp. 1–9, 2020.
[16] W. D. Septiani, P. Studi, and M. Informatika, "DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS," vol. 13, no. 1, pp. 76–84, 2022.
[17] N. Singh and P. Singh, "Chemometrics and Intelligent Laboratory Systems A hybrid ensemble- fi lter wrapper feature selection approach for medical data classi fi cation," *Chemom. Intell. Lab. Syst.*, vol. 217, no. July, p. 104396, 2021, doi: 10.1016/j.chemolab.2021.104396.
[18] S. Fathmah, D. Kartini, F. Abadi, I. Budiman, and M. I. Mazdadi, "Implementation of PPCA Imputation , SMOTE-N Class Balancing in Hepatitis Classification Using Naïve Bayes," vol. 12, no. 2, pp. 169–176, 2024.
[19] L. Parisi and N. Ravichandran, "Knowledge-Based Systems Evolutionary feature transformation to improve prognostic prediction of hepatitis," *Knowledge-Based Syst.*, vol. 200, p. 106012, 2020, doi: 10.1016/j.knosys.2020.106012.
[20] M. Watanabe *et al.*, "Metabolomic markers and physiological adaptations for high phosphate utilization efficiency in rice," no. October 2019, pp. 2066–2079, 2020, doi: 10.1111/pce.13777.
[21] H. S. Sadiq and A. M. Abdulazeez, "Indonesian Journal of Computer Science," vol. 13, no. 1, pp. 4122–4140, 2024.