



# Analisis performa metode klasifikasi *Naïve Bayes Classifier* pada *Unbalanced Dataset*

Ericha Apriliyani<sup>a,1</sup>, Yulita Salim<sup>a,2</sup>

<sup>a</sup> Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.05, Makassar dan 90231, Indonesia

<sup>1</sup> 13020170113@umi.ac.id; <sup>2</sup> yulita.salim@umi.ac.id;

INFORMASI ARTIKEL	ABSTRAK
Diterima : 21 – 05 – 2022 Direvisi : 12 – 06 – 2022 Diterbitkan : 31 – 07 – 2022	Penelitian ini bertujuan menganalisis menganalisis performa metode <i>Naïve Bayes Classifier</i> pada berbagai variasi <i>unbalanced dataset</i> . Teknik klasifikasi yang digunakan yaitu dengan menggunakan pengukuran performa <i>accuracy</i> , <i>precision</i> , <i>recall</i> , dan <i>f-measure</i> dengan menggunakan metode <i>Naïve Bayes Classifier</i> . Dataset diambil dari <i>Kaggle</i> , dan dari tiap dataset memiliki data yang tidak seimbang atau <i>unbalance</i> , data inilah yang akan dihitung performanya menggunakan metode <i>Naïve Bayes Classifier</i> . Berdasarkan hasil pengujian menggunakan metode <i>Naïve Bayes Classifier</i> didapatkan pada dataset <i>Glass</i> yaitu <i>accuracy</i> sebesar 46%, <i>precision</i> sebesar 47%, <i>recall</i> sebesar 46% dan <i>f-measure</i> sebesar 43%, pada dataset <i>Heart Disease</i> yaitu <i>accuracy</i> sebesar 88%, <i>precision</i> sebesar 88%, <i>recall</i> sebesar 88% dan <i>f-measure</i> sebesar 88%, pada dataset <i>Kidney Disease</i> yaitu <i>accuracy</i> sebesar 100%, <i>precision</i> sebesar 100%, <i>recall</i> sebesar 100% dan <i>f-measure</i> sebesar 100%, pada dataset <i>Liver Disease</i> yaitu <i>accuracy</i> sebesar 78%, <i>precision</i> sebesar 82%, <i>recall</i> sebesar 78% dan <i>f-measure</i> sebesar 79%, pada dataset <i>Diabetes</i> yaitu <i>accuracy</i> sebesar 77%, <i>precision</i> sebesar 76%, <i>recall</i> sebesar 77% dan <i>f-measure</i> sebesar 76%, dan pada dataset <i>Breast Cancer</i> yaitu <i>accuracy</i> sebesar 94%, <i>precision</i> sebesar 94%, <i>recall</i> sebesar 94% dan <i>f-measure</i> sebesar 94%. Hasil penelitian ini menunjukkan metode <i>Naïve Bayes Classifier</i> pada <i>unbalanced dataset</i> memperoleh nilai performa yang tidak menentu.
<b>Kata Kunci:</b> <i>Naïve Bayes Classifier</i> <i>Unbalanced Dataset</i> <i>Analisis performa</i> <i>Data science</i> <i>Data mining</i>	 

## I. Pendahuluan

Data tidak seimbang atau *unbalanced dataset* adalah suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (*instance*) yang satu lebih sedikit atau lebih banyak dibanding menggunakan jumlah kelas data lainnya.[1] Dalam mengklasifikasikan sebuah dokumen dengan data yang tidak seimbang dapat menghasilkan tingkat *accuracy* yang berbeda.

*Machine learning* adalah bidang kecerdasan buatan atau *artificial intelligence* yang berkaitan dengan pengembangan teknik-teknik yang dapat diprogram dan dipelajari dari data masa lalu.[2] Pengenalan pola, *data mining*, dan *machine learning* sering digunakan untuk menggambarkan hal yang sama[3], [4], [13]–[17], [5]–[12]. Bidang ini bersinggungan dengan ilmu probabilitas dan statistik, terkadang dengan optimasi. *Machine learning* menjadi alat analisis dalam *data mining* [18].

*Data mining* adalah penemuan atau penemuan informasi baru dengan mencari pola atau aturan tertentu terhadap sejumlah data yang sangat besar [19]. *Data mining* juga dikenal sebagai rangkaian proses untuk menemukan nilai tambah berupa pengetahuan yang tidak diketahui secara manual dari suatu kumpulan data [20]. *Data mining*, juga dikenal sebagai *knowledge discovery in database* (KDD). KDD adalah aktivitas yang melibatkan pengumpulan dan penggunaan data historis untuk menemukan keteraturan, pola atau hubungan dalam *set data* yang berukuran besar [21].

Klasifikasi merupakan bentuk dasar dari analisis data yang mempunyai teknik untuk menentukan keanggotaan kelompok berdasarkan data-data yang sudah ada[22]–[27]. Konsep dasar dari klasifikasi adalah beberapa data yang memiliki struktur data yang mirip akan memiliki klasifikasi yang mirip pula [28].

*Naïve Bayes Classifier* (NBC) merupakan pengklasifikasi probabilitas sederhana untuk menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang

diberikan [29] Metode *Naïve Bayes* merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi [30]. Keuntungan menggunakan *Naïve Bayes* adalah metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi [31].

Terdapat sebuah penelitian yang membandingkan kinerja dari metode *Naïve Bayes* dan *K-Nearest Neighbor*, sehingga dapat diketahui bahwa metode *Naïve Bayes* lebih baik dalam melakukan klasifikasi dokumen. Hasil yang didapatkan menunjukkan metode *Naïve Bayes* memiliki kinerja yang lebih baik dengan tingkat *accuracy* 70%, sedangkan metode *K-Nearest Neighbor* memiliki tingkat *accuracy* yang cukup rendah yaitu 40% [32]. Sedangkan pada penelitian yang lain dilakukan penelitian *dataset* yang tidak seimbang menggunakan metode Algoritma klasifikasi C4.5, CART, *k-Nearest Neighbours* (k-NN) dan *Naïve Bayes*, yang dimana metode k-NN lebih unggul dari yang lainnya. Dengan menghitung skor nilai yang merupakan gabungan dari parameter kinerja “*accuracy*”, “*precision*”, “*recall*” dan “*AUC*” dimana nilai tertinggi dari masing-masing parameter adalah yang terbaik akan menghasilkan penilaian kinerja *klasifier* yang lebih representatif menunjukkan kualitas *klasifier*. [33] Dilakukan dua metode pengujian yaitu *10-fold Cross Validation* dan Pengujian Secara *Diskrit* untuk memastikan hasil penilaian kinerja yang representatif dari masing-masing *klasifier*. Penerapan pengujian terhadap empat algoritma klasifikasi dan analisis perbandingan kinerja menghasilkan urutan peringkat kualitas kinerja terbaik yaitu: 1. k-NN, 2. C4.5, 3. CART, 4. *Naïve Bayes* [34].

Berdasarkan pada penelitian yang telah dipaparkan sebelumnya. Dengan ini peneliti tertarik untuk menganalisis performa metode *Naïve Bayes Classifier* pada berbagai jenis *unbalanced dataset* [35]. Data tidak seimbang atau *unbalanced dataset* merupakan suatu keadaan dimana distribusi kelas data tidak seimbang, jumlah kelas data (*instance*) yang satu lebih sedikit atau lebih banyak dibanding dengan jumlah kelas data lainnya [36], *unbalanced dataset* yang digunakan pada penelitian ini ditunjukkan pada Tabel 1 berikut:

Tabel 1. Daftar *Unbalanced Dataset*

No.	Nama Data	Jumlah Data	Target Data	Klasifikasi	
1.	<i>Glass</i>	214	6	1	70
				2	76
				3	17
				5	13
				6	9
				7	29
				2.	<i>Heart Disease</i>
1	138				
3.	<i>Kidney Disease</i>	280	2	Ckd	174
				Notckd	106
4.	<i>Liver Disease</i>	583	2	1	416
				2	167
5.	<i>Diabetes</i>	2000	2	0	1316
				1	684
6.	<i>Breast Cancer</i>	569	2	0	212
				1	357

Berdasarkan pada Tabel 1. terdapat enam jenis *dataset* yang dimana tiap datanya memiliki jumlah data yang tidak seimbang, pada enam jenis data tersebut diperoleh dari *Kaggle*. Penelitian ini bertujuan untuk memperoleh nilai performa metode *Naïve Bayes Classifier* pada berbagai jenis *unbalanced dataset*, dimana performa yang diukur adalah *accuracy*, *precision*, *recall*, dan *f-measure*.

## II. Metode

### A. Tahapan Penelitian

#### 1) Load Dataset

*Dataset* yang digunakan pada penelitian ini berjumlah enam *dataset* dan akan menggunakan data yang tidak seimbang.

#### 2) Tahap Preprocessing Data

Tahap ini dilakukan untuk menghilangkan permasalahan- permasalahan yang dapat mengganggu hasil dari proses data. Tahap *preprocessing* data terdiri dari 2 tahapan, yaitu:

##### a) Feature Selection

Tahap ini dilakukan untuk menghapus fitur-fitur yg tidak digunakan dalam memproses data.

##### b) Final Dataset

Tahap ini adalah tahap akhir dari *dataset* yang telah di proses.

### 3) Implementasi Cross Validation

Pada tahap ini, nilai *k-fold cross validation* yang digunakan adalah  $k=5$ , dimana data akan menjadi 5 lipatan. Tujuannya adalah untuk menemukan kombinasi data yang terbaik.

### 4) Implementasi Metode Naïve Bayes Classifier

Pada tahap ini merupakan proses pengklasifikasian data pada beberapa *dataset*. Pada *dataset* pertama yaitu *dataset Glass* menjadi tujuh label yang dimana pada klasifikasi data yang keempat nilainya tidak ada dalam *database*, *dataset* kedua yaitu *dataset Heart Disease* terbagi menjadi dua label yaitu penyakit jantung (1) dan bukan penyakit jantung (0), *dataset* ketiga yaitu *dataset Kidney Disease* yang terbagi menjadi dua label yaitu penyakit ginjal kronis (Ckd) dan bukan penyakit ginjal kronis (Notckd), pada *dataset* keempat yaitu *dataset Liver Disease* terbagi menjadi dua label yaitu penyakit hati (1) dan bukan penyakit hati (0), pada *dataset* kelima yaitu *dataset Diabetes* terbagi menjadi dua label yaitu penyakit diabetes (1) dan bukan penyakit diabetes (0), dan *dataset* yang keenam atau yang terakhir adalah *dataset Breast Cancer* yang terbagi menjadi dua label yaitu penyakit kanker payudara (1) dan bukan penyakit kanker payudara (0).

### 5) Analisis Performa Metode Naïve Bayes Classifier

Dalam tahap ini, dilakukan perhitungan performa berupa *accuracy*, *precision*, *recall*, dan *f-measure* pada metode *Naïve Bayes Classifier* pada enam *dataset* yaitu *dataset Glass*, *dataset Heart Disease*, *dataset Kidney Disease*, *dataset Liver Disease*, *dataset Diabetes* dan *dataset Breast Cancer*

### 6) Pengambilan Keputusan

Kesimpulan diambil berdasarkan hasil performa (*accuracy*, *precision*, *recall*, *f-measure*) pada metode *Naïve Bayes Classifier* dalam mengklasifikasi enam *dataset* yaitu *dataset Glass*, *dataset Heart Disease*, *dataset Kidney Disease*, *dataset Liver Disease*, *dataset Diabetes* dan *dataset Breast Cancer*

## III. Hasil dan Pembahasan

### A. Implementasi

Pada tahap implementasi menggambarkan beberapa tahap yang terdiri dari implementasi kebutuhan perangkat lunak, perangkat keras, implementasi analisis *missing value*, implementasi *feature selection*, implementasi *cross validation*, dan implementasi metode *Naive Bayes Classifier* (NBC).

#### 1) Manual Instalasi

Dari enam jenis *dataset* yang ada, terdapat satu *dataset* yang memiliki *missing value* yaitu *dataset Kidney Disease* pada *dataset* ini terdapat 149 *record* data yang memiliki *missing value* dan 11 *field* yang termasuk.

#### 2) Implementasi Feature Selection

Pada tahap ini, penulis melakukan proses penghapusan data yang tidak digunakan seperti data yang *non numeric* dan data yang *null*.

#### 3) Implementasi Cross Validatio

Implementasi *cross validation* dilakukan pada setiap enam *dataset*, *dataset Glass*, *dataset Heart Disease*, *Kidney Disease*, *dataset Liver Disease*, *dataset Diabetes* dan *dataset Breast Cancer*. Setiap *dataset* akan dibagi menjadi 5 lipatan, dimana setiap lipatannya terdiri dari 80% data training dan 20% data testing, setiap lipatan akan bergiliran menjadi data testing.

#### 4) Implmentasi Metode Naïve Bayes Classifier

Implementasi metode *Naïve Bayes Classifier* pada tahap ini merupakan proses perhitungan untuk menentukan nilai tertinggi pada performa akurasi presisi, *recall*, dan *f-measure* dimana penulis menggunakan *dataset Heart Disease* dan *dataset Kidney Disease*.

Tabel 2. Sample Dataset Heart (Data Training)

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Target
57	1	0	150	276	0	0	112	1	0.6	1	1	1	0
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
68	1	2	180	274	1	0	150	1	1.6	1	0	3	0
57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
53	0	0	130	264	0	0	143	0	0.4	1	0	2	1

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Target
61	1	2	150	243	1	1	137	1	1	1	0	2	1
62	1	2	130	231	0	1	146	0	1.8	1	3	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
41	0	1	105	198	0	1	168	0	0	2	1	2	1

Tabel 3. Sample Dataset Heart (Data Testing)

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Target
69	0	3	140	239	0	3	151	0	1.8	2	2	2	?
45	0	1	130	234	0	0	175	0	0.6	1	0	2	?

Pada Tabel 2. menunjukkan 10 data yang digunakan sebagai data *training* dan pada Tabel 3. menunjukkan 2 data yang digunakan sebagai data *testing*. Tahap selanjutnya yaitu proses implementasi metode *Naïve Bayes Classifier* menggunakan persamaan 2, 3, dan 4.

Pada Tabel 4. menunjukkan hasil perhitungan nilai rata-rata (*mean*) setiap *class* pada setiap atribut, dimana data *training* di setiap atribut dijumlahkan lalu dibagi dengan jumlah data setiap *class* dengan menggunakan rumus *Naïve Bayes Classifier* pada persamaan 3.

Tabel 4. Hasil perhitungan nilai rata-rata (*mean*)

Class	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal
0	59.4	0.8	0.4	148.4	252	0.2	0.4	131	0.8	0.92	1.2	0.6	2.6
1	52.2	0.6	1.4	129	233.8	0.2	0.8	154.6	0.4	0.72	1.4	0.8	2.2

Pada Tabel 5. menunjukkan hasil perhitungan standar deviasi setiap *class* pada setiap atribut, dimana data *training* di setiap atribut dikurangkan dengan nilai rata-rata dan dipangkatkan dua, lalu dibagi dengan jumlah data setiap *class* dengan menggunakan rumus *Naïve Bayes Classifier* pada persamaan 4.

Tabel 5. Hasil perhitungan standar deviasi

Class	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal
0	5.41E+00	4.47E-01	8.94E-01	2.04E+01	3.07E+01	4.47E-01	5.48E-01	2.33E+01	4.47E-01	9.44E-01	4.47E-01	5.48E-01	8.94E-01
1	9.58E+00	5.48E-01	8.94E-01	1.60E+01	2.39E+01	4.47E-01	4.47E-01	1.80E+01	5.48E-01	7.01E-01	5.48E-01	1.30E+00	4.47E-01

Pada Tabel 6. menunjukkan 10 data yang digunakan sebagai data *training* dan pada Tabel 7. menunjukkan 2 data yang digunakan sebagai data *testing*.

Tabel 6. Sample dataset kidney (Data training)

Age	Bp	Sg	Al	Su	Bgr	Bu	Sc	Sod	Pot	Hemo	Pcv	Wc	Rc	Target
80	1.025	0	0	111	44	1.2	146	3.6	16.3	40	8000	6.4	4.5	0
42	80	1.02	0	0	132	24	0.7	140	4.1	14.4	50	5000	4.5	0
82	80	1.01	2	2	140	70	3.4	136	4.2	13	40	9800	4.2	0
24	70	1.025	0	0	140	23	0.6	140	4.7	16.3	48	5800	5.6	0
62	80	1.01	1	2	309	113	2.9	130	2.5	10.6	34	12800	4.9	0
52	90	1.015	4	3	224	166	5.6	133	47	8.1	23	5000	2.9	1
55	80	1.01	3	1	214	73	3.9	137	4.9	10.9	34	7400	3.7	1
40	70	1.015	3	4	253	150	11.9	132	5.6	10.9	31	8800	3.4	1
64	90	1.015	3	2	463	64	2.8	135	4.1	12.2	40	9800	4.6	1
63	70	1.01	3	0	380	60	2.7	131	4.2	10.8	32	4500	3.8	1

Tabel 7. Sample dataset kidney (Data Testing)

Age	Bp	Sg	Al	Su	Bgr	Bu	Sc	Sod	Pot	Hemo	Pcv	Wc	Rc	Target
73	100	1.01	3	2	295	90	5.6	140	2.9	9.2	30	7000	3.2	?
68	70	1.015	3	1	208	72	2.1	138	5.8	9.7	28	12200	3.4	?

Pada Tabel 6. menunjukkan 10 data yang digunakan sebagai data *training* dan pada Tabel 7. menunjukkan 2 data yang digunakan sebagai data *testing*. Tahap selanjutnya yaitu proses implementasi metode *Naive Bayes Classifier* menggunakan persamaan 2, 3, dan 4.

Pada Tabel 8. menunjukkan hasil perhitungan nilai rata-rata (*mean*) setiap *class* pada setiap atribut, dimana data *training* di setiap atribut dijumlahkan lalu dibagi dengan jumlah data setiap *class* dengan menggunakan rumus *Naive Bayes Classifier* pada persamaan 3.

Tabel 8. Hasil perhitungan Nilai Rata-rata (*mean*)

Class	Age	Bp	Sg	Al	Su	Bgr	Bu	Sc	Sod	Pot	Hemo	Pcv	Wc	Rc
0	58	62.205	0.813	0.6	23	153	46.24	30.72	109.92	6.36	18.86	1634.4	6681.28	4.74
1	54.8	80	1.013	3.2	2	306.8	102.6	5.38	133.6	13.16	10.58	32	7100	3.68

Pada Tabel 9. menunjukkan hasil perhitungan standar deviasi setiap *class* pada setiap atribut, dimana data *training* di setiap atribut dikurangkan dengan nilai rata-rata dan dipangkat dua, lalu dibagi dengan jumlah data setiap *class* dengan menggunakan rumus *Naive Bayes Classifier* pada persamaan 4.

Tabel 9. Hasil Perhitungan Standar Deviasi

Class	Age	Bp	Sg	Al	Su	Bgr	Bu	Sc	Sod	Pot	Hemo	Pcv	Wc	Rc
0	2.49E+01	3.45E+01	4.55E-01	8.94E-01	4.92E+01	9.62E+01	4.50E+01	6.45E+01	5.96E+01	5.62E+00	1.20E+01	3.56E+03	4.88E+03	5.41E-01
1	9.73E+00	1.00E+01	2.74E-03	4.47E-01	1.58E+00	1.10E+02	5.11E+01	3.83E+00	2.41E+00	1.89E+01	1.50E+00	6.12E+00	2.32E+03	6.22E-01

## B. Pembahasan

Pada penelitian ini menggunakan bahasa pemrograman *Python* dan *library scikit learn*, mulai dari tahap *load dataset*, analisis *missing value*, klasifikasi *feature selection* dan implementasi metode *Naive Bayes Classifier*, hingga proses perhitungan nilai performa (akurasi, presisi, *recall*, dan *f-measure*).

### 1) Pembahasan Dataset

*Dataset* yang digunakan sebagai objek penelitian ini sebanyak enam *dataset*, dimana datanya berupa *numeric*. Pada dataset pertama yaitu *dataset Glass* terbagi menjadi tujuh label, pada klasifikasi data yang keempat nilainya tidak ada dalam *database*, dataset kedua yaitu *dataset Heart Disease* terbagi menjadi dua label yaitu penyakit jantung (1) dan bukan penyakit jantung (0), dataset ketiga yaitu *dataset Kidney Disease* yang terbagi menjadi dua label yaitu penyakit ginjal kronis (Ckd) dan bukan penyakit ginjal kronis (Notckd), pada *dataset* keempat yaitu *dataset Liver Disease* terbagi menjadi dua label yaitu penyakit hati (1) dan bukan penyakit hati (0), pada *dataset* kelima yaitu *dataset Diabetes* terbagi menjadi dua label yaitu penyakit *diabetes* (1) dan bukan penyakit *diabetes* (0), dan *dataset* yang keenam atau yang terakhir adalah *dataset Breast Cancer* yang terbagi menjadi dua label yaitu penyakit kanker payudara (1) dan bukan penyakit kanker payudara (0). *Dataset* ini diperoleh dari *repository Kaggle*.

## C. Kesimpulan Pengujian

Berdasarkan hasil pengujian performa metode *Naive Bayes Classifier* (NBC) pada pengujian performa akurasi, presisi, *recall*, dan *f-measure* dari keenam *dataset* maka dapat disimpulkan pada Tabel 10.

Tabel 10. Hasil Pengujian Performa

Dataset	Accuracy	Precision	Recall	F-measure
<i>Glass</i>	46%	47%	46%	43%
<i>Heart Disease</i>	88%	88%	88%	88%
<i>Kidney Disease</i>	100%	100%	100%	100%
<i>Liver Disease</i>	58%	82%	58%	59%
<i>Diabetes</i>	77%	76%	77%	76%
<i>Breast Cancer</i>	88%	89%	88%	88%

Pada Tabel 8. menunjukkan nilai performa tertinggi yang mewakili dari setiap *dataset* dan nilainya tersebut merupakan nilai maksimum dari hasil perhitungan *cross validation*. Setiap *dataset* mempunyai hasil performa

yang berbeda-beda dari keenam *dataset*, *Kidney Disease* merupakan *dataset* dengan performa tertinggi dengan nilai akurasi 100%, presisi 100%, *recall* 100% dan *f-measure* 100%, sedangkan *dataset* dengan performa terendah adalah *dataset Glass* dengan nilai akurasi 46%, presisi 47%, *recall* 46% dan *f-measure* 43%. Hasil performa dari *Naïve Bayes Classifier* menunjukkan nilai yang tidak menentu atau nilai yang bisa berubah-ubah dari *dataset* yang sifatnya *unbalanced*.

#### IV. Kesimpulan

Berdasarkan penelitian ini dapat disimpulkan bahwa Penerapan metode *Naïve Bayes Classifier* pada *unbalanced Dataset* memperoleh nilai performa yang tidak menentu. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Glass* yaitu *accuracy* sebesar 46%, *precision* sebesar 47%, *recall* sebesar 46% dan *f-measure* sebesar 43%. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Heart Disease* yaitu *accuracy* sebesar 88%, *precision* sebesar 88%, *recall* sebesar 88% dan *f-measure* sebesar 88%. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Kidney Disease* yaitu *accuracy* sebesar 100%, *precision* sebesar 100%, *recall* sebesar 100% dan *f-measure* sebesar 100%. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Liver Disease* yaitu *accuracy* sebesar 78%, *precision* sebesar 82%, *recall* sebesar 78% dan *f-measure* sebesar 79%. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Diabetes* yaitu *accuracy* sebesar 77%, *precision* sebesar 76%, *recall* sebesar 77% dan *f-measure* sebesar 76%. Nilai performa metode *Naïve Bayes Classifier* pada *Dataset Breast Cancer* yaitu *accuracy* sebesar 94%, *precision* sebesar 94%, *recall* sebesar 94% dan *f-measure* sebesar 94%. Bagian ini adalah menerangkan opini, pendapat penulis berdasarkan hasil dan pembahasan penelitian yang dilakukan. Kesimpulan menjawab hipotesis, maksud, dan tujuan penelitian. Selain itu, kesimpulan menyajikan pernyataan singkat tentang pentingnya temua yang diperoleh dan implikasinya di masa depan.

#### Daftar Pustaka

- [1] A. Saifudin, U. Pamulang, R. S. Wahono, U. Dian, and N. Semarang, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 2, pp. 76–85, 2015.
- [2] I. Fauzan, "Artificial Intelligence (AI) Pada Proses Pengawasan dan Pengendalian Kepegawaian - Sebuah Eksplorasi Konsep Setelah Masa Pandemi Berakhir," *J. Civ. Serv.*, vol. 14, no. 1, pp. 31–42, 2020.
- [3] F. Tangguh and Y. Islami, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022.
- [4] L. B. C. Tanujayaa, B. Susanto, and A. Saragiha, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indones. J. data Sci.*, vol. 1, no. 3, pp. 68–78, 2020.
- [5] Wahyu ngestisari, "The Perbandingan Metode ARIMA dan Jaringan Syaraf Tiruan untuk Peramalan Harga Beras," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 96–107, 2020, doi: 10.33096/ijodas.v1i3.18.
- [6] A. A. D. Halim and S. Anraeni, "Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN)," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 01–12, 2021, doi: 10.33096/ijodas.v2i1.23.
- [7] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.
- [8] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.
- [9] Harlinda and Nasir, "Perancangan sistem pendukung keputusan dalam pengalokasian dana bantuan sosial di kabupaten pinrang dengan menggunakan metode AHP," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 44–51, 2020.
- [10] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.
- [11] H. Azis, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, pp. 1–4, 2018, doi: <https://doi.org/10.33096/ijodas.v1i1.3>.

- [12] M. D. U. Ulhaq and Irawati, "Implementasi Metode Visekriterijumsko Kompromisno Rangiranje (VIKOR) Pada Seleksi Program Keluarga Harapan Komponen Pendidikan Berbasis Web," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 38–49, 2021, doi: 10.33096/ijodas.v2i1.30.
- [13] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.
- [14] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.
- [15] I. P. Putri, "Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.
- [16] D. Susanti, "Analisis Modifikasi Metode Playfiar Cipher Dalam Pengamanan Data," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–80, 2020.
- [17] A. Prasetya Wibawa, W. Lestar, A. Bella Putra Utama, I. Tri Saputra, and Z. Nabila Izdihar, "Multilayer Perceptron untuk Prediksi Sessions pada Sebuah Website Journal Elektronik," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 57–67, 2020, doi: 10.33096/ijodas.v1i3.15.
- [18] Dwi Retnosari, "Sistem Aplikasi Data Mining Untuk Menampilkan," *J. Integr. Sist. Ind. UMJ*, vol. 1, no. 2, pp. 13–20, 2014.
- [19] F. Mulianingsih, K. Anwar, F. A. Shintasiwi, and A. J. Rahma, "Program Studi Tadris Ilmu Pengetahuan Sosial Institut Agama Islam Negeri Kudus Artificial Intellegence dengan Pembentukan Nilai dan Karakter di Bidang Pendidikan," *Ijtimaiya J. Soc. Sci. Teach.*, vol. 4, no. 2, pp. 148–154, 2020.
- [20] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.)*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.
- [21] R. D. Fitriani, H. Yasin, and T. Tarno, "PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," *J. Gaussian*, vol. 10, no. 1, pp. 11–20, 2021, doi: 10.14710/j.gauss.v10i1.30243.
- [22] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020.
- [23] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.
- [24] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.
- [25] A. A. Karim, H. Azis, and Y. Salim, "Kinerja Metode C4.5 dalam Penyaluran Bantuan Dana Bencana 1," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 84–87, 2018.
- [26] L. Nurhayati and H. Azis, "Perancangan Sistem Pendukung Keputusan Untuk Proses Kenaikan Jabatan Struktural Pada Biro Kepegawaian," *Semin. Nas. Teknol. Inf. dan Multimed.*, pp. 6–7, 2016.
- [27] F. Muharram, H. Azis, and A. R. Manga, "Analisis Algoritma pada Proses Enkripsi dan Dekripsi File Menggunakan Advanced Encryption Standard (AES)," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 112–115, 2018.
- [28] Y. I. Kurniawan, "Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 455, 2018, doi: 10.25126/jtiik.201854803.
- [29] A. Saleh, "Implementasi Metodel Klasifikasi Naive Bayes," *Citec J.*, vol. 2, no. 3, pp. 207–217, 2015, doi: 10.20895/inista.v1i2.73.
- [30] F. Liantoni and H. Nugroho, "Klasifikasi Daun Herbal Menggunakan Metode Naïve Bayes Classifier Dan Knearest Neighbor," *J. Simantec*, vol. 5, no. 1, pp. 9–16, 2015.
- [31] Herman *et al.*, "Comparison of Artificial Neural Network and Gaussian Naïve Bayes in Recognition of Hand-Writing Number," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2018, pp. 276–279, doi: 10.1109/EIConCIT.2018.8878651.

- 
- [32] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 427, 2018, doi: 10.25126/jtiik.201854773.
- [33] C. A. Sugianto, "Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk SMAN 1 Gibeber Untuk Siswa Baru Menggunakan Decision Tree," *J. TEDC*, vol. 9, no. 1, pp. 39–43, 2015, doi: 10.31227/osf.io/vedu7.
- [34] C. Anam and N. Rusdiana, "Analisis Pemeringkatan Kualitas Klasifier Pada Dataset Tidak Seimbang," *JIMP - J. Inform. Merdeka Pasuruan*, vol. 5, no. 1, pp. 38–44, 2020, doi: 10.37438/jimp.v5i1.248.
- [35] Maryamah, M. F. Asikin, D. Kurniawaty, S. K. Sari, and I. Cholissodin, "Implementasi Metode Naive Bayes Classifier untuk Seleksi Asisten Praktikum pada Simulasi Hadoop Multinode Cluster," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, pp. 273–278, 2016.
- [36] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR," *Isd*, vol. 3, no. 1, pp. 2528–5114, 2018.