Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Klasifikasi Kualitas Air Menggunakan K-Nearest Neighbors, Naïve Bayes, Dan Logistic Regression

Mandy Sandra Denny[a], Dyah Erny Herwindiati[b]
[a]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, mandy.535210034@stu.untar.ac.id
[b]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, dyahh@fti.untar.ac.id

Submitted: 26-09-2024, Reviewed: 04-10-2024, Accepted 27-10-2024
https://doi.org/10.47233/jteksis.v6i4.1649

Abstract
Water is an important natural resource for living creatures on earth, but water is very easily contaminated with bacteria and dangerous substances. Therefore, it is important to pay attention to the water quality on earth. This research aims to classify water quality as safe or unsafe and compare the methods used to obtain the highest accuracy value from the implementation of algorithms in machine learning. In this research, three methods are used, namely K-Nearest Neighbors (KNN), Naïve Bayes, and Logistic Regression, and then the classification results will be compared to choose the most suitable method. In this research, the dataset used is Water Quality from the Kaggle website, which contains 7999 samples with 20 features and 1 target class. The results obtained from the method used will be compared and take the highest accuracy value. The accuracy results of the KNN, Naive Bayes, and Logistic Regression methods are 89.62%, 78.69%, and 89.81%, respectively. The highest accuracy result is Logistic Regression, so this method is the best method for classifying water quality data.
Keywords: *KNN, Water Quality, Logistic Regression, Machine Learning, Naïve Bayes*

Abstrak
Air merupakan sumber daya alam yang penting bagi kehidupan makhluk hidup di bumi, namun air sangat mudah terkontaminasi bakteri dan zat berbahaya. Oleh karena itu, penting untuk memperhatikan kualitas air di bumi. Penelitian ini bertujuan untuk mengklasifikasikan kualitas air apakah aman atau tidak aman dan membandingkan metode yang digunakan untuk mendapatkan nilai akurasi tertinggi dari implementasi algoritma pada machine learning. Dalam penelitian ini, digunakan tiga metode, yaitu K-Nearest Neighbors (KNN), Naïve Bayes, dan Logistic Regression, kemudian hasil klasifikasi akan dibandingkan untuk memilih metode yang paling cocok. Dalam penelitian ini, dataset yang digunakan adalah Water Quality dari situs web Kaggle, yang berisi 7999 sampel dengan 20 fitur dan 1 kelas target. Hasil yang diperoleh dari metode yang digunakan akan dibandingkan dan mengambil nilai akurasi tertinggi. Hasil akurasi dari metode KNN, Naive Bayes, dan Logistic Regression masing-masing sebesar 89.62%, 78.69%, dan 89.81%. Hasil akurasi tertinggi merupakan Logistic Regression, sehingga metode ini merupakan metode yang paling baik untuk klasifikasi data kualitas air.
Keywords: KNN, Kualitas Air, Logistic Regression, Machine Learning, Naïve Bayes

This work is licensed under Creative Commons Attribution License 4.0 CC-BY International license

PENDAHULUAN

Air merupakan sumber daya alam dan senyawa kimia yang sangat bagi kehidupan makhluk hidup di bumi. Sumber air seperti air laut, air sungai, air tanah, dan lain-lain membentuk volume air 1,4 triliun kilometer kubik yang merupakan hampir 71% wilayah di Bumi. Air mengandung banyak zat dan mineral, tetapi beberapa ada yang tidak dapat diterima oleh tubuh manusia. Tercemarnya sumber air atau lingkungan sekitar membuat bakteri dan zat mineral berbahaya dapat masuk ke dalam air [1]. Karena itu memperhatikan kualitas air sangat diperlukan. Kualitas air digunakan untuk menggambarkan kondisi air, termasuk karakteristik kimiawi, fisik dan biologis sesuai dengan tujuan penggunaan airnya [2]. Kualitas air dapat dipengaruhi oleh limbah industri dan limbah rumah tangga dalam bentuk cair maupun padat. Perlu dilakukan pengecekan kualitas air untuk mengontrol penyakit dan bakteri yang ada pada air, agar dapat memenuhi standar kualitas air bersih dan air layak konsumsi [3].

Machine Learning adalah cabang kecerdasan buatan atau *Artificial Intelligence* (AI) yang fokus menggunakan model komputer dan algoritma untuk mempelajari pola data, melakukan prediksi, dan membuat keputusan [4]. Proses belajar mesin dilakukan dengan dua tahap, yaitu latihan (*training*) dan pengujian (*testing*). *Machine Learning* dibagi menjadi tiga kategori, yaitu *Supervised Learning, Unsupervised Learning, dan Reinforcement Learning*. *Supervised Learning* adalah klasifikasi data dengan label untuk mengklasifikasikan kelas, *Unsupervised Learning* sama dengan *clustering* yang tidak membutuhkan label pada data, sedangkan *Reinforcement Learning* berada di antara *Supervised Learning* dan *Unsupervised Learning* [5].

Penelitian terkait klasifikasi kualitas air sudah banyak dilakukan dengan menggunakan penerapan metode *Machine Learning*. Penelitian yang

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

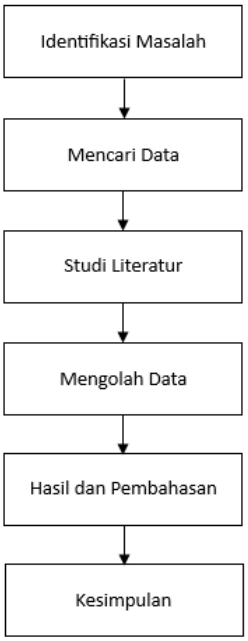
E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

dilakukan oleh Maulana dan kawan-kawan untuk klasifikasi kualitas air bersih dengan melakukan komparasi antara metode KNN dan Naïve Bayes, dari hasil penelitian tersebut didapatkan hasil akurasi metode KNN 82.42% dan Naïve Bayes 70.32%. Sedangkan penelitian yang dilakukan oleh Hartanti, D. dan Pradana, A. I dengan membandingkan metode *Decision Tree*, *Logistic Regression*, *Support Vector Machine* (SVM), dan *Artificial Neural Network* (ANN). Hasil akurasi metode penelitian tersebut adalah *Decision Tree* 60.19%, *Logistic Regression* 62.80%, SVM 68.59%, dan ANN 69.54% [6].

Penelitian sebelumnya menunjukkan bahwa tingkat akurasi dari beberapa teknik klasifikasi berbeda. Berdasarkan penelitian sebelumnya, klasifikasi dengan metode KNN dan *Naïve Bayes* memiliki hasil akurasi yang lebih baik dibandingkan penelitian menggunakan metode *Decision Tree*, *Logistic Regression*, SVM dan ANN. Oleh karena itu, dalam penelitian ini akan membandingkan tiga metode machine learning yang diambil dari dua penelitian sebelumnya, yaitu KNN, *Naïve Bayes*, dan *Logistic Regression* untuk menentukan metode yang lebih baik untuk membantu klasifikasi kualitas air. Dipilihnya tiga algoritma tersebut karena mudah dipahami dan diimplementasikan. Penelitian ini menggunakan *dataset Water Quality* dari Kaggle dengan tipe data csv. Tujuan dari penelitian ini adalah untuk mengklasifikasikan kualitas air berdasarkan dua kelas, yaitu aman atau tidak aman dengan parameter dirincikan pada Tabel 1. Selain itu juga untuk membandingkan performa dari setiap algoritma yang digunakan untuk mendapatkan algoritma terbaik dalam melakukan klasifikasi kualitas air.

METODE PENELITIAN

Tahapan atau alur penelitian yang digunakan untuk melakukan klasifikasi kualitas air menggunakan algoritma *K-Nearest Neighbors*, *Naïve Bayes*, dan *Logistic Regression* dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Penjelasan tahapan penelitian, yaitu

1. Tahap pertama dimulai dengan mengidentifikasi masalah untuk mencari topik yang akan digunakan, disini terkait dengan pentingnya kualitas air bagi kehidupan makhluk hidup.
2. Tahap kedua setelah menentukan topik peneliti mencari data yang berhubungan atau terkait dengan topik yang digunakan. Data yang digunakan dapat diambil dari sumber terpercaya, disini peneliti menggunakan data dari situs Kaggle.
3. Tahap ketiga studi literatur peneliti mengumpulkan informasi-informasi yang terkait dengan topik yang diambil. Informasi yang digunakan dalam penelitian ini berupa jurnal, buku, dan sumber-sumber yang terpercaya.
4. Tahap keempat setelah selesai mengumpulkan studi literatur, pengolahan data terdiri dari beberapa langkah. Pertama *preprocessing* data untuk menghilangkan nilai NULL pada data, memberi label, membagi data menjadi data latih dan data uji, melakukan proses model fit atau *classification*, dan *predict model*.
5. Tahap kelima setelah mengolah data akan mendapat hasil akurasi, maka hasil yang didapat kemudian akan dibahas. Untuk melihat akurasi digunakan metode evaluasi model dengan *Confusion Matrix* dan laporan klasifikasi yang terdiri dari *precision*, *recall*, f1-*score*, dan *support*.

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

6. Tahap keenam dari pembahasan hasil dibuat sebuah kesimpulan dari proses atau tahapan yang sudah dilalui [7].

Dataset Water Quality diambil dari situs Kaggle [8]. Dalam datasets berisi 7999 sampel, 20 fitur atau parameter, dan 1 kelas target untuk menentukan aman atau tidaknya kualitas air. Dataset dengan kelas target aman berjumlah 912 data, kelas target tidak aman berjumlah 7084 data, dan kelas target yang tidak diketahui berjumlah 3data. Datasets yang digunakan seperti pada Tabel 1 [9].

Tabel 1. Parameter Dataset Kualitas Air

No	Parameter	Deskripsi
1	Aluminum	berbahaya jika lebih besar dari 2.8
2	Ammonia	berbahaya jika lebih besar dari 32.5
3	Arsenic	berbahaya jika lebih besar dari 0.01
4	Barium	berbahaya jika lebih besar dari 2
5	Cadmium	berbahaya jika lebih besar dari 0.005
6	Chloramine	berbahaya jika lebih besar dari 4
7	Chromium	berbahaya jika lebih besar dari 0.1
8	Copper	berbahaya jika lebih besar dari 1.3
9	Fluoride	berbahaya jika lebih besar dari 1.5
10	Bacteria	berbahaya jika lebih besar dari 0
11	Viruses	berbahaya jika lebih besar dari 0
12	Lead	berbahaya jika lebih besar dari 0.015
13	Nitrates	berbahaya jika lebih besar dari 10
14	Nitrites	berbahaya jika lebih besar dari 1
15	Mercury	berbahaya jika lebih besar dari 0.002
16	Perchlorate	berbahaya jika lebih besar dari 56
17	Radium	berbahaya jika lebih besar dari 5
18	Selenium	berbahaya jika lebih besar dari 0.5
19	Silver	berbahaya jika lebih besar dari 0.1
20	Uranium	berbahaya jika lebih besar dari 0.3
21	is_safe	aman atau tidak aman

Sumber: www.kaggle.com/datasets/mssmartypants/water-quality

Dalam penelitian ini terdapat tahapan dalam mengelola data, yaitu preprocessing dan labeling data, pembagian data atau data splitting, dan implementasi algoritma.

2.1. *Preprocessing* dan *Labeling Data*

Data perlu dilakukan *preprocessing* untuk menghilangkan data yang missing atau NULL. Karena dataset yang digunakan hanya memiliki sedikit baris data yang NULL, maka data NULL akan dihilangkan. Keputusan yang berkualitas didasarkan pada data yang berkualitas, karena itu *preprocessing* langkah yang perlu dilakukan dalam proses penemuan pengetahuan [10].

Penggunaan algoritma dapat dikategorikan menjadi *supervised* atau *unsupervised*, dan implementasinya digunakan sebagai analisis pola (*unsupervised*) atau klasifikasi (*supervised*) [11]. Penelitian ini menggunakan *Supervised Learning*, metode klasifikasi dengan data yang diberi label untuk mengklasifikasikan kelas [12]. *Datasets* yang digunakan belum memiliki *label*, maka dari itu perlu dilakukan proses *labeling data*. Dari keterangan setiap fitur di atas dikonversi ke bilangan biner, yaitu 0 dan 1. Kolom untuk kelas target yang digunakan adalah 'is_safe' dengan dua label, yaitu 0 dan 1.

2.2 Pembagian Data atau *Data Splitting*

Peneliti membagi membagi kolom fitur dan target untuk menjadi data latih dan data uji. Data latih dan data uji dibagi menjadi perbandingan 8:2, sama dengan 80% untuk data latih dan 20% untuk data uji. Kemudian menggunakan random state sama dengan 42 agar pembagian data yang dilakukan sama ketika kode program dijalankan berulang, dan untuk mendapatkan hasil yang konsisten agar bisa membandingkan hasil akurasi antar algoritma.

2.3 Implementasi Algoritma

Penelitian ini menggunakan algoritma *K-Nearest Neighbors*, *Naïve Bayes*, dan *Logistic Regression*.

2.3.1 *K-Nearest Neighbors* (KNN)

Algoritma *KNN* adalah metode klasifikasi objek dengan data pembelajaran yang jaraknya paling dekat atau perbedaan nilai yang kecil dengan objek yang diuji, dimana jumlah tetangga terdekat ditentukan secara manual dan dinyatakan dengan k [13]. Hasil perhitungan kemudian dikelompokkan dengan data latih yang mempunyai tetangga terdekat terbanyak dalam nilai jangkauan tertentu. Salah satu metrik yang digunakan untuk menghitung jarak data latih dan data uji adalah Euclidean Distance, rumus *Euclidean Distance* dapat dilihat pada persamaan (1) [14].

$$D(x,y) = \sqrt{\sum_{i-1}^{n} (x_i - y_i)^2} \quad (1)$$

Keterangan:
D(x,y) = Jarak titik antara data latih dan data uji

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis
E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

x = Data latih (x1, x2, x3, …, xi)
y = Data uji (y1, y2, y3, …, yi)
n = Dimensi data
i = Variabel data

Penentuan nilai k terbaik tergantung pada data yang digunakan. Nilai k tinggi secara umum akan mengurangi efek noise, tetapi batasan antar klasifikasi akan lebih kabur [15].

2.3.2 *Naïve Bayes*

Algoritma *Naïve Bayes* adalah metode klasifikasi yang menggunakan metode probabilitas dan statistik, metode ini efisien untuk mengelompokkan data dengan memaksimalkan akurasi probabilistik [16]. *Naïve Bayes* atau Teorema Bayes dikemukakan oleh ilmuwan Inggris Thomas Bayes, dan digunakan untuk memprediksi peluang masa depan berdasarkan pengalaman sebelumnya [17]. Rumus *Naïve Bayes*:

$$P(H|X) = \frac{P(H)\,P(X|H)}{P(X)} \quad (2)$$

Keterangan:
X = Data kelas yang belum diketahui
H = Hipotesis data X
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X
P(H) = Probabilitas hipotesis H (*prior probability*)
P(X|H) = Probabilitas X berdasarkan kondisi hipotesis H
P(X) = Probabilitas X

2.3.3 *Logistic Regression*

Algoritma *Logistic Regression* adalah model linear untuk klasifikasi, menggambarkan hasil percobaan tunggal merupakan model dengan fungsi logistik. Regresi logistik digunakan untuk mendeskripsikan data dan hubungan antara satu variabel biner dependen dengan satu atau lebih variabel independen dengan tipe kategori [18]. Persamaan *Logistic Regression* seperti pada Persamaan (5), dan nilai peluang (Y = 1) dapat menggunakan persamaan (6) [19].

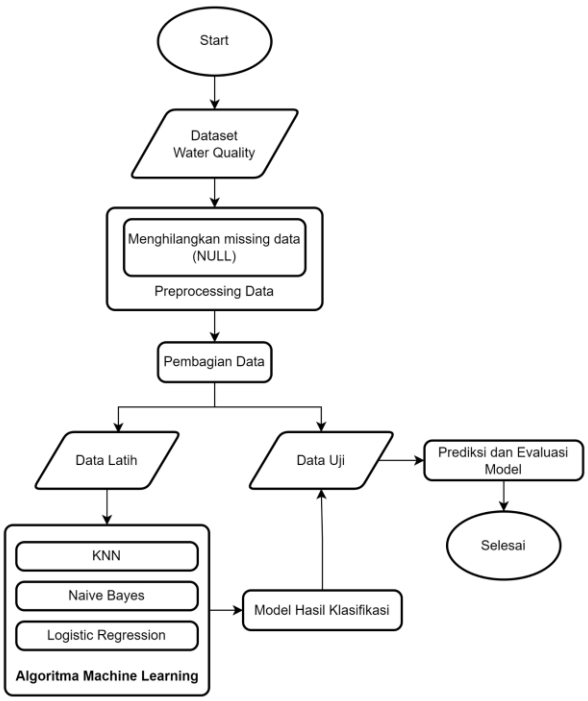
$$Ln\left(\frac{p}{1-p}\right) = B_0 + B_1 X \quad (3)$$

$$p = \frac{e^{(B_0+B_1X)}}{(1 + e^{(B_0+B_1X)})} \quad (4)$$

Keterangan:
p = Peluang
B_0 = Konstanta
B_1 = Koefisien masing-masing variabel

2.4 Rancangan Eksperimen

Flowchart diagram (diagram alir) tentang rancangan eksperimen yang meliputi input - proses - output dari penelitian ini dibuat dapat dilihat pada Gambar 2 [20].



Gambar 2. Rancangan Eksperimen

2.5 Metode Evaluasi

Melakukan evaluasi dari model yang telah dibangun dengan proses *training* adalah hal yang penting. Tujuan dari evaluasi adalah mengukur performa atau kinerja model apakah mampu melakukan klasifikasi dengan baik atau tidak. Metode evaluasi yang digunakan dalam penelitian ini adalah *Confusion Matrix* dan laporan klasifikasi yang terdiri dari *precision*, *recall*, f1-*score*, dan *support*.

2.5.1 *Confusion Matrix*

Confusion Matrix adalah tabel yang memberi informasi perbandingan hasil klasifikasi yang diprediksi dengan hasil klasifikasi yang sebenarnya [21]. *Confusion* matriks yang digunakan dalam penelitian ini memiliki dimensi 2 × 2, contohnya dapat dilihat pada Gambar 3

		Nilai Aktual	
		Positif	Negatif
Nilai Prediksi	Positif	TP	FN
	Negatif	FP	TN

Gambar 3. Confusion Matrix 2 × 2

Ada 4 nilai hasil klasifikasi yang akan digunakan untuk menghitung nilai akurasi dan laporan klasifikasi yang terdiri dari presisi, *recall*, dan f1-score, yaitu

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

1. *True Positive* (TP) adalah total sampel data yang diklasifikasikan positif
2. *True Negative* (TN) adalah total sampel data diklasifikasikan negatif
3. *False Positive* (FP) adalah total sampel data diklasifikasikan positif, tetapi sebenarnya negatif
4. *False Negative* (FN) adalah total sampel data diklasifikasikan negatif tetapi sebenarnya positif

2.5.2 Laporan Klasifikasi

Berdasarkan nilai yang didapat dari Confusion Matriks dapat dihasilkan laporan klasifikasi sebagai berikut.

1. Akurasi

   Akurasi adalah metrik untuk evaluasi model klasifikasi, yang merupakan perhitungan pecahan jumlah prediksi yang benar dengan total prediksi yang ada. Persamaan akurasi dapat dilihat pada persamaan 5.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

2. Presisi

   Presisi adalah informasi data yang diambil salah atau tidak tepat dan menunjukan tingkat ketepatan dalam klasifikasi [20]. Persamaan presisi dapat dilihat pada persamaan 6.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

3. *Recall*

   Recall adalah pengukuran proporsi positif yang benar diidentifikasi dan meenampilkan berapa banyak data positif diklasifikasikan dengan benar oleh sistem. Persamaan *recall* dapat dilihat pada persamaan 7.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

4. F1-*score* atau F-*measure*

   F1-*score* adalah gabungan dari metrik presisi dan *recall* yang merupakan "harmonisasi" dari kedua metrik tersebut karena mengambil rata-rata dari kedua metrik. Persamaan *recall* dapat dilihat pada persamaan 8.

$$F\text{—}measure = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

HASIL DAN PEMBAHASAN

Hasil dari pengolahan data yang telah dilakukan pada tahap *preprocessing* adalah dengan menghilangkan missing value dan mengubah tipe data menjadi float. Hasil dari preprocessing yang telah dilakukan pada dataset Water Quality Classification yang berisi 7999 sampel, setelah dilakukan preprocessing data jumlah sampel menjadi 7996 sampel. Info data setelah preprocessing dapat dilihat pada Gambar 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7996 entries, 1 to 7996
Data columns (total 21 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   0       7996 non-null   float64
 1   1       7996 non-null   float64
 2   2       7996 non-null   float64
 3   3       7996 non-null   float64
 4   4       7996 non-null   float64
 5   5       7996 non-null   float64
 6   6       7996 non-null   float64
 7   7       7996 non-null   float64
 8   8       7996 non-null   float64
 9   9       7996 non-null   float64
 10  10      7996 non-null   float64
 11  11      7996 non-null   float64
 12  12      7996 non-null   float64
 13  13      7996 non-null   float64
 14  14      7996 non-null   float64
 15  15      7996 non-null   float64
 16  16      7996 non-null   float64
 17  17      7996 non-null   float64
 18  18      7996 non-null   float64
 19  19      7996 non-null   float64
 20  20      7996 non-null   float64
dtypes: float64(21)
memory usage: 1.3 MB
```

Gambar 4. Info Data Setelah Dilakukan Preprocessing

Setelah menghilangkan NULL value dan juga mengubah data menjadi *float* dilakukan labeling atau memberi label pada data. Hasil setelah memberi label pada data dapat dilihat pada Gambar 5. Label data berisi 0 untuk berbahaya dan 1 untuk aman.

	0	1	2	3	4	5	6	7	8	9	...	11	12	13	14	15	16	17	18	19	20
1	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0
2	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	...	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0
3	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
4	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	...	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0
5	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	...	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7992	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
7993	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7994	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7995	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7996	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	...	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

7996 rows × 21 columns

Gambar 5. Hasil Data Setelah Diberi Label

Dari data hasil preprocessing, pembagian data diperlukan menjadi data latih dan data uji. Pada penelitian ini pembagian data untuk data latih dan data uji dibagi dengan perbandingan 8:2. Setelah pembagian data, selanjutnya data dapat digunakan untuk algoritma yang digunakan. Hasil metode

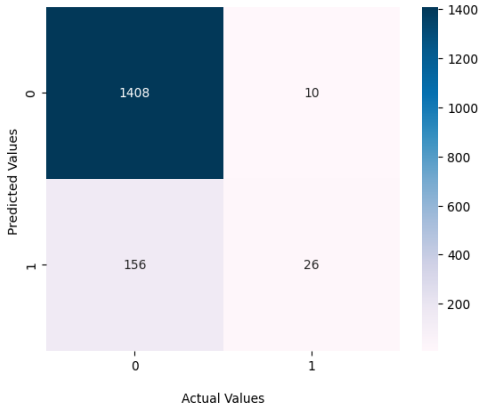
Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

evaluasi dari implementasi algoritma adalah sebagai berikut.

1. *K-Nearest Neighbor*

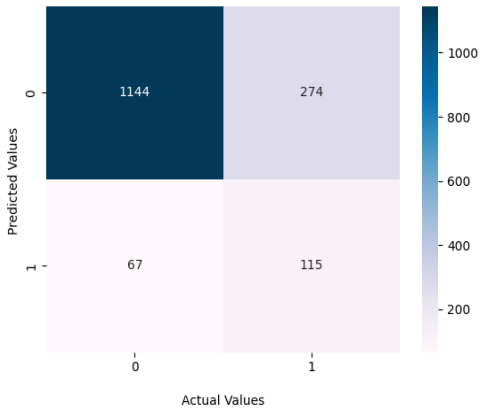
   Dalam eksperimen ini digunakan beberapa nilai k untuk mencari nilai k terbaik yang dapat digunakan. Nilai k yang digunakan adalah 3, 5, 7, 9, 11, dan 13. Pada perbandingan 8 : 2 hasil akurasi terbaik adalah 89.62% pada k = 11. *Confusion Matrix* dapat dilihat pada gambar 6.



Gambar 6. Confusion Matrix Algoritma KNN

2. *Naïve Bayes*

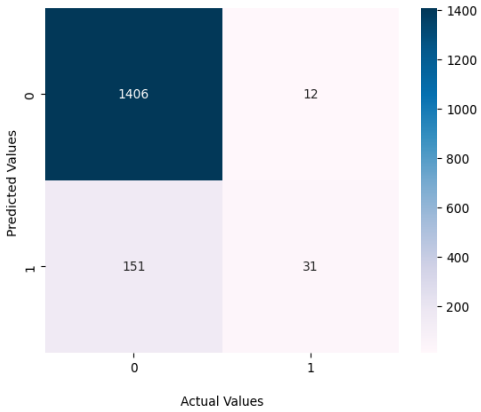
   Dalam eksperimen ini didapat hasil akurasi sebesar 78.69%. *Confusion Matrix* dapat dilihat pada gambar 7.



Gambar 7. Confusion Matrix Algoritma Naïve Bayes

3. *Logistic Regression*

   Dalam eksperimen ini didapat hasil akurasi sebesar 89.81%. *Confusion Matrix* dapat dilihat pada gambar 8.



Gambar 8. Confusion Matrix Algoritma Logistic Regression

Dari hasil *confusion matrix*, dapat dihasilkan perhitungan laporan klasifikasi presisi, *recall*, dan f1-*score* yang disajikan dalam bentuk tabel. Hasil dapat dilihat pada Tabel 2.

Tabel 2. Hasil Implementasi Algoritma

Algoritma	Akurasi		Presisi	Recall	F1-score
	Data Latih	Data Uji			
KNN	90.1%	89.6%	0.143	0.722	0.238
Naïve Bayes	75.2%	78.7%	0.632	0.296	0.403
Logistic Regression	89.37%	89.8%	0.17	0.721	0.275

Dari hasil pengujian nilai tertinggi untuk akurasi data latih adalah metode KNN dengan selisih 0.73% dibanding dengan metode Logistic Regression. Sedangkan untuk nilai akurasi data uji tertinggi adalah metode *Logistic Regression* dengan hasil 89.81%, tetapi hasil yang didapat tidak berbeda jauh dengan KNN dan tidak disarankan untuk menggunakan Naïve Bayes. Hasil presisi, *recall*, dan F1-*score* diperoleh dari memasukan nilai pada *confusion matrix* ke dalam rumus:

1. KNN

$$Accuracy = \frac{1434}{1600} = 0.8962 = 89.62\%$$

$$Precision = \frac{26}{182} = 0.1428 = 14.28\%$$

$$Recall = \frac{26}{36} = 0.7222 = 72.22\%$$

$$F1\text{-}score = \frac{0.2062}{0.865} = 0.2384 = 23.84\%$$

2. Naïve Bayes

$$Accuracy = \frac{1259}{1600} = 0.78687 = 78.69\%$$

$$Precision = \frac{115}{182} = 0.6318 = 63.19\%$$

$$Recall = \frac{115}{389} = 0.2956 = 29.56\%$$

$$F1\text{-}score = \frac{0.3735}{0.9275} = 0.4026 = 40.26\%$$

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

3. Logistic Regression

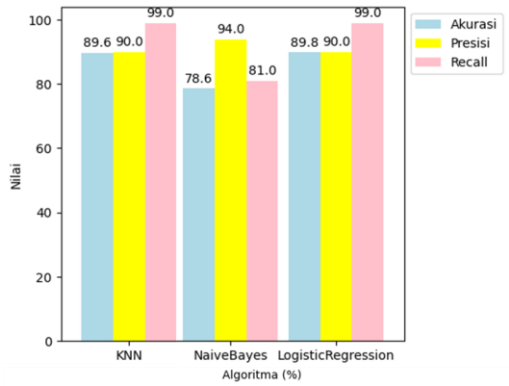
$$Accuracy = \frac{1437}{1600} = 0.78687 = 78.69\%$$

$$Precision = \frac{31}{182} = 0.1703 = 17.03\%$$

$$Recall = \frac{31}{43} = 0.7209 = 72.09\%$$

$$F1\text{-}score = \frac{0.2455}{0.8912} = 0.2754 = 27.54\%$$

Jika hasil yang didapat dibuat grafik maka visualisasi data seperti pada Gambar 9. Tetapi angka yang didapat berbeda dengan yang dijabarkan di rumus untuk presisi dan *recall*, karena dalam grafik menggunakan perhitungan untuk kelas *false*.



Gambar 9. Grafik Perbandingan Akurasi, Presisi, dan Recall

SIMPULAN

Dari hasil penelitian yang telah diperoleh dapat disimpulkan metode yang memberikan hasil akurasi data uji terbaik adalah *Logistic Regression* sebesar 89.81%, tetapi tidak berbeda jauh dengan metode klasifikasi KNN yang memberikan hasil sebesar 89.62%. Dalam penelitian menggunakan metode *Naive Bayes* memberikan hasil yang cukup karena hanya menghasilkan akurasi sebesar 78.69%, sehingga tidak disarankan menggunakan *Naive Bayes*. Untuk mendapatkan hasil akurasi yang lebih maksimal penelitian selanjutnya bisa menggunakan metode *machine learning* yang lain, seperti *Random Forest*, *Support Vector Machine* (SVM), atau *Neural Networks*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang pertama kepada Tuhan YME, orang tua, dosen pembimbing, dan kepada semua yang mendukung penelitian ini. Terima kasih atas saran dan bantuan yang diberikan dalam penelitian ini.

DAFTAR PUSTAKA

[1] Rahman, M. A., Hidayat, N., & Supianto, A. A. 2018. Komparasi Metode Data Mining K-Nearest Neighbor Dengan Na{\"\i}ve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2(12): 6346-6353.

[2] Rahman, F. Y., Purnomo, I. I. and Hijriana, N. 2018. Penerapan Algoritma Data Mining Untuk Klasifikasi Kualitas Air. Technologia: Jurnal Ilmiah. 13(3): 228-232.

[3] Aini, Q. 2018. Analisa Zat Organik Bod dan Cod Pada Air Sungai disekitar Pabrik Tahu di Daerah Junok Bangkalan. *Thesis*. Universitas Muhammadiyah Surabaya.

[4] Faiza, I. M., & Andriani, W. 2022. Tinjauan Pustaka Sistematis: Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir. Jurnal Minfo Polgan. 11(2): 59-63.

[5] Roihan, A., Sunarya, P. A., & Rafika, A. S. 2020. Pemanfaatan Machine Learning dalam Berbagai Bidang. Jurnal Khatulistiwa Informatika. 5(1): 490845.

[6] Hartanti, D. & Pradana, A. I. 2023. Komparasi Metode Decission Tree, Logistic Regression, SVM, dan ANN Dalam Klasifikasi Kualitas Air. SMARTICS Journal. 9(1): 1-6.

[7] Tangkelayuk, A. 2022. The Klasifikasi Kualitas Air Menggunakan Metode KNN, Na{\"\i}ve Bayes, dan Decision Tree. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi). 9(2): 1109-1119.

[8] MsSmartyPants. 2021. Water Quality. Available: https://www.kaggle.com/datasets/mssmartypants/water-quality.

[9] Pritalia, G. L. 2022. Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum. KONSTELASI: Konvergensi Teknologi dan Sistem Informasi. 2(1).

[10] Setyadji, A. E. S., & Wibowo, A. P. 2021. Analisis Klaster Data Pasien Diabetes untuk Identifikasi Pola dan Karakteristik Pasien. Jurnal Teknologi Dan Sistem Informasi Bisnis. 5(3): 172-182.

[11] Diponegoro, M. H., Kusumawardani, S. S., & Hidayah, I. 2021. Tinjauan pustaka sistematis: implementasi metode deep learning pada prediksi kinerja murid. Jurnal Nasional Teknik Elektro Dan Teknologi Informasi. 10(2): 131-138.

[12] Aziz, A., Zy, A. T., & Sunge, A. S. 2024. Prediksi Penjualan Obat Dan Alat Kesehatan Terlaris Menggunakan Algoritma K-Nearest Neighbor. Jurnal Teknologi Dan Sistem Informasi Bisnis. 6(1): 117-124.

[13] Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan kinerja metode naive bayes dan k-nearest neighbor untuk klasifikasi artikel berbahasa indonesia. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 5(4): 427-434.

[14] Baharuddin, M. M., Azis, H., & Hasanuddin, T.. 2019 Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca. ILKOM Jurnal Ilmiah. 11(3): 269-274.

[15] Arinda, P. H., & Angga, L. 2024. Sistem Prediksi Penjualan Produk APD Terlaris di PT A3 Karunia Sidoarjo menggunakan Metode Naive Bayes. Jurnal Teknologi Dan Sistem Informasi Bisnis. 6(2): 431-437.

[16] Sari, Y. S. 2021. Penerapan Metode Na{\"\i}ve Bayes Untuk Mengetahui Kualitas Air Di Jakarta. Jurnal Ilmiah FIFO. 13(2): 222-228.

[17] Muharemi, F., Logof{\u{a}}tu, D., & Leon, F. 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication. 3(3): 294-307.

[18] F. D. Pramakrisna, F. D. Adhinata and N. A. F. Tanjung. 2022. Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression. Teknika. 11(2): 90-97.

[19] A. S. Dharma, O. Y. Silitonga and H. J. Manurung. 2019. PERBANDINGAN ALGORITMA Naïve BAYES, ID3,

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 6 No. 4 Oktober 2024 Hal. 851-858
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

DAN TAN PADA KLASIFIKASI SMS SPAM. Journal of Maritime and Education (JME). 1(2).

[20] M. I. Fikri, T. S. Sabrila and Y. Azhar. 2020. Perbandingan metode na{\"\i}ve bayes dan support vector machine pada analisis sentimen twitter. SMATIKA Jurnal: STIKI Informatika Jurnal. 10(02): 71-76.

[21] K. Nugroho and S. Murdowo. 2022. Klasifikasi Jenis Hewan Pada Kebun Binatang Dengan Menggunakan Metode Deep Neural Network. Jurnal Ilmiah Infokam. 18(1): 46-51