Klasifikasi Kualitas Air Menggunakan Metode *KNN, Naïve Bayes* Dan *Decision Tree*

Aldi Tangkelayuk[*1], Evangs Mailoa[2]
[1,2]Jl. Diponegoro 52-60, Salatiga 50711, Indonesia
[3]Jurusan Teknik Informatika, FTI UKSW, Salatiga
e-mail: *[1]67201022@student.uksw.edu, [2]evangs.mailoa@uksw.edu

Abstrak

Air merupakan kebutuhan yang sangat penting bagi kehidupan manusia, namun tidak semua air aman untuk dikomsumsi, sehingga diperlukan adanya identifikasi kualitas air yang baik untuk di komsumsi. Klasifikasi merupakan salah satu teknik data mining, dimana dapat membangun model dari sampel data menjadi satu kelompok yang sama. Klasifikasi memiliki banyak algoritma yang sering di gunakan seperti algoritma *Naive Baiyes, Desicion Tree,* dan *K-nearest Neighbors.* Penelitian ini menggunakan dataset kualitas air dengan tiga algoritma yaitu *Naive Baiyes, Desicion Tree,* dan *K-nearest Neighbors*. Ketiga algoritma ini akan dilakukan perbandingan pada proses klasifikasi data untuk mengetahui metode mana yang paling akurat, dilihat dari tingkat akurasi yang paling tinggi. Hasil penelitian ini menunjukkan metode *K-nearest Neighbors* memiliki tingkat akurasi paling tinggi sebesar 86.88% dibandingkan dengan *Decision Tree* sebesar 80.84% dan *Naïve Bayes* sebesar 63.60%, sehingga metode *K-nearest Neighbors* merupakan metode yang paling baik untuk klasifikasi data.

Kata kunci: Kualitas air**,** Klasifikasi, *K-nearest neighbors, Naïve Bayes, Decision Tree.*

Abstract

Water is a very important need for human life, but not all water is safe for consumption, so it is necessary to identify good water quality for consumption. Classification is one of the data mining techniques, which can build models from sample data into the same group. Classification has many algorithms that are often used such as the Naive Baiyes algorithm, Desicion Tree, and K-nearest Neighbors. This study uses a water quality dataset with three algorithms, namely Naive Baiyes, Desicion Tree, and K-nearest Neighbors. These three algorithms will be compared in the data classification process to find out which method is the most accurate, seen from the highest level of accuracy. The results of this study indicate that the K-nearest Neighbors method has the highest accuracy rate of 86.88% compared to Decision Tree of 80.84% and Naïve Bayes of 63.60%, so the K-nearest Neighbors method is the best method for data classification.

Keywords: *Water quality, Classification, K-nearest neighbors, Naïve Bayes, Decision Tree.*

1. PENDAHULUAN

Data mining merupakan salah satu metode untuk menentukan pola tertentu dari sekumpulan data yang berjumlah besar. Data mining memiliki banyak teknik salah satunya teknik klasifikasi. Klasifikasi merupakan teknik pembelajaran data untuk menghasilkan prediksi nilai dari serangkaian atribut [1]. Klasifikasi banyak digunakan untuk memprediksi kelas pada label tertentu, yaitu dengan mengklasifikasi data (membangun model) berdasarkan *training set*

dan nilai-nilai (label kelas) dalam mengklasifikasikan atribut tertentu. Klasifikasi dibagi menjadi lima kategori berdasarkan perbedaan konsep matematika, yaitu berbasis statistik, berbasis jarak, berbasis pohon keputusan, berbasis jaringan syaraf, dan berbasis *rule*. Klasifikasi memiliki banyak algoritma namun pada penelitian ini menggunakan algoritma *Decision Tree, KNN,* dan *Naïve Bayes.* Ketiga algoritma ini, *Decision Tree* merupakan salah satu metode yang paling sering di gunakan khususnya dalam klasifikasi data. Pada studi kasus analisis sentiment pengguna layanan BPJS yang menggunakan metode *KNN, Naïve Bayes,* dan *Decision Tree,* membuktikan bahwa metode *Decision Tree* memiliki tingkat akurasi tinggi dalam klasifikasi data[2]. Pada studi kasus Komparasi Metode Data Mining *K-Nearest Neighbor* dengan *Naïve Bayes* untuk klasifikasi kualitas air bersih, terbukti bahwa metode *KNN* memiliki tingkat akurasi yang tinggi dibandingkan dengan *Naïve Bayes*[3]. Jika dibandingkan dengan metode *Naïve Bayes,* metode ini jarang mendapatkan tingkat akurasi yang tinggi, sehingga pada penelitian ini akan membandingkan ketiga algoritma tersebut dengan melihat tingkat akurasinya, metode mana yang paling baik untuk klasifikasi.

Penelitian ini menggunakan data kualitas air menggunakan data *Water Potability* dengan tipe data csv, berisi data kualitas air dengan *value* yang berbeda-beda pada sepuluh *attributte*, *attributte potability* merupakan *attributte* yang diberi label sebagai hasil *conclusion* apakah air tersebut aman atau tidak aman untuk diminum dengan presentasi dari masing-masing *value attributte* yang ada pada data tersebut.

Berdasarkan permasalahan yang ada yaitu ingin membandingkan ketiga metode *Decision Tree, KNN,* dan *Naïve Bayes* maka dilakukan penelitian dengan judul "Klasifikasi Kualitas Air Menggunakan Metode *KNN, Naïve Bayes, dan Decision Tree"* menggunakan *software rapid miner* untuk mengetahui nilai akurasi yang paling besar dari ketiga metode yang akan diimplementasikan kedalam klasifikasi data yakni Analisis Perbandingan Akurasi *Water Quality* menggunakan Klasifikasi Data *KNN, Naïve Bayes, dan Decision Tree*. Tujuan penelitian ini untuk membandingkan diantara ketiga metode yang paling baik digunakan dalam klasifikasi kualitas air dengan hasil akurasi yang paling maksimal.

2. LANDASAN TEORI

Penelitian yang membahas metode *Naïve Bayes, KNN,* dan *Decision Tree* terhadap Analisis Sentimen Transportasi KRL *Commuter* dengan permasalahan kondisi lalu lintas di kota Jakarta yang begitu padat dan kemacetan terus meningkat menyebabkan warga yang hendak ingin bekerja memerlukan transportasi yang lebih nyaman. Penelitian ini menggunakan media *social twitter* untuk mendapatkan data bersifat random sebanyak 127 data. Menggunakan metode *Naive Bayes Classifier, KNN, dan Desicion Tree* dengan beberapa tahapan yaitu *Convert Emoticon, Cleansing, Case Folding, Tokenizing,* dan *Stemming*[4]. Hasil yang didapatkan metode *Decision Tree* memiliki akurasi yang terbesar dibandingkan dengan *KNN* dan *Naïve Bayes* dimana *Decision Tree* memiliki akurasi 100%, *precision* 100%, *sensitivity* 100%, dan *specificity* 100%. Metode KNN memiliki akurasi 80%, *precision* 100%, *sensitivity* 50%, *specificity* 100%, dan metode *Naïve Bayes* memiliki akurasi 80%, *precision* 66,67%, *sensitivity* 100%, dan *specificity* 66,67%.

Penelitian yang membahas Klasifikasi Kualitas Air Bersih pada Studi Kasus PDAM Tirta Kencana Kabupaten Jombang menggunakan Komparasi Metode Data Mining *K-Nearest Neighbor* dengan *Naïve Bayes*. Dibutuhkan pengawasan dan pengolahan lingkungan sekitarnya termasuk sumber air agar dapat menghasilkan kualitas air yang bersih sesuai dengan standar kualitas air bersih dan layak dikonsumsi oleh manusia[3]. Hasil akurasi sebesar 82,42% pada metode *K-Nearest Neighbor* dan *Naïve Bayes* sebesar 70,32%, dapat disimpukan metode *K-Nearest Neighbor* adalah metode yang paling baik dalam menentukan kualitas air yang bersih dan aman untuk diminum.

Pada penelitian Analisis Sentimen Pengguna Layanan *BPJS* menggunakan metode *KNN, Decision Tree,* dan *Naïve Bayes* yang membahas tentang manusia menggunakan layanan *BPJS* yang sering kali terjadi *pro* dan *kontra*, hal tersebut merupakan alasan mengapa dilakukan penelitian analisis *sentiment* data mining terhadap pengguna *BPJS* pada media *social twitter* dengan 1.000 data yang di*filter* menjadi 903 dikarenakan adanya data yang terduplikat. Implementasikan metode KNN, *Decision Tree*, dan Naïve Bayes guna untuk membandingkan tingkat akurasi dari tiga metode yang digunakan[2]. Penelitian ini menggunakan *software rapid miner* versi 9.9 dimana hasil penelitian diperoleh metode *KNN* tingkat akurasinya sebesar 95.58%, *Decision Tree* sebesar 96,13%, dan metode *Naïve Bayes* sebesar 89.14%, maka dapat disimpulkan metode yang paling baik digunakan adalah *Decision Tree.*

Data Mining adalah proses mengekstrak informasi untuk memperoleh informasi yang baru[5]. Penelitian yang dilakukan kali ini menggunakan teknik data mining yang mengimplementasikan metode *K-nearest Neighbors, Naïve Bayes,* dan *Decision Tree* untuk membandingkan hasil akurasi yang paling maksimal dari ketiga metode yang digunakan. Data mining adalah operasi *resourcing* serta penggunaan data yang digunakan untuk mencari hubungan atau pola dari suatu kumpulan data yang besar untuk memperoleh informasi yang baru[6].

Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya [6]. Algoritma klasifikasi *KNN* merupakan metode untuk mengklasifikasikan objek yang didasarkan pada data latih yang memiliki jarak terdekat[4]. Prinsip kerja algoritma dari *KNN* yaitu menentukan dan mencari jarak terdekat dengan nilai k *neighbor* terdekat dalam data *training* dengan data yang akan diuji. Nilai k yang terbaik untuk algoritma ini tergantung berdasarkan nilai sebuah data, dimana biasanya nilai k yang tinggi mengurangi efek kesalahan atau *noise* pada proses klasifikasi, akan tetapi membuat sebuah batasan antar klasifikasi menjadi tidak maksimal. Penelitian ini akan dilakukan proses perhitungan untuk menghasilkan hasil akurasi data data yang ada dengan metode *KNN*. Rumus pencarian jarak dengan menggunakan rumus *euclidian*:

$$\mathbf{d_i} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{p}} (\mathbf{x_{2i}} - \mathbf{x_{1i}})^2} \quad (1)$$

dimana x1 adalah sampel data ; d adalah jarak ; x2 adalah data uji ; p adalah dimensi data;i adalah variabel data.

Naive Bayes Classifier adalah metode dalam data mining untuk mengklasifikasikan data. Cara kerja metode *Naive Bayes Classifier* menggunakan perhitungan *probabilitas.* Naive Bayes merupakan algoritma yang terdapat pada teknik klasifikasi [7]. Konsep dasar *Naive Bayes* mengunakan *teorema Bayes*, yaitu *teorema* yang digunakan dalam statistika yang digunakan menghitung suatu peluang, *Naive Bayes Classifier* menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada dan menentukan kelas yang paling optimal[8]. *Naive bayes classifier* berfungsi menghitung dan mencari nilai *probabilitas* paling tinggi untuk mengklasfikasikan sebuah data uji dengan kategori yang tepat. Teknik prediksi *probabilitas* yang sederhana didasarkan pada penerapan *teorema bayes* atau aturan *bayes* merupakan suatu teknik yang diimplementasikan pada algoritma *naïve bayes*. Rumus *naïve bayes* :

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

dimana X adalah data dengan kelas yang belum diketahui ; H adalah *Hipotesis* data X merupakan kelas spesifik ; P(H|X) adalah *Probabilitas hipotesis* H berdasar kondisi X ; P(H)

adalah *Probabilitas hipotesis* H *(prior probability)* ; P(X|H) adalah *Probabilitas* X berdasar kondisi pada *hipotesis* H ; P(X) adalah *Probabilitas* dari X.

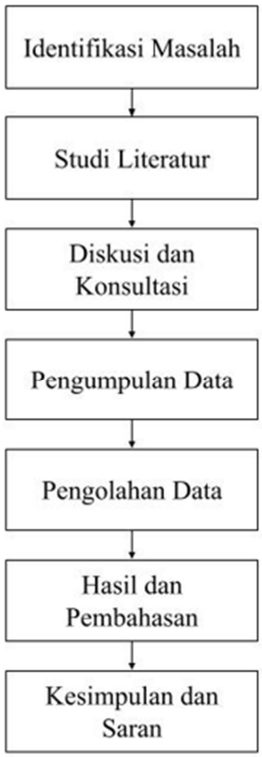
Proses klasifikasi data dapat menggunakan beberapa metode salah satunya adalah *Decision Tree* atau pohon keputusan. *Decision Tree* adalah algoritma yang umum digunakan untuk mengambil sebuah keputusan [9]. Decision Tree merupakan algoritma yang baik digunakan untuk klasifikasi atau prediksi [10].Pohon keputusan adalah metode klasifikasi yang melibatkan konstruksi pohon keputusan yang terdiri dari node keputusan yang di hubungkan dengan cabang-cabang dari simpul akar sampai ke node daun (akhir)[8]**..**Model *Decision Tree* adalah berbentuk sebuah pohon yang terdiri dari beberapa bagian yaitu *root node, internal node,* dan *terminal node*. *Data query* menelusuri *root node* dan *internal node* mencapai terminal *node* merupakan proses melakukan klasifikasi pada metode *Decision Tree* ini. Konsep *entropi* yang akan digunakan untuk menentukan atribut mana pada pohon keputusan akan terbagi, semakin tinggi *entory sampel* maka semakin tidak murni sampel tersebut. Rumus untuk menghitung *entropy sampel* adalah:

$$\textit{Entropy}(S) = - P_1 \log_2 P1 - P2 \log_2 P_2 \quad (3)$$

dimana p1, p2, p3….., pn masing-masing menyatalam *proposi* kelas 1, kelas 2,….. kelas n pada *output*.

3. METODE PENELITIAN

Dalam penelitian ini digunakan beberapa tahapan yang disajikan dalam bentuk Gambar 1 Tahapan Penelitian.



Gambar 1**.** Tahapan Penelitian

Tahap pertama dalam penelitian ini dimulai dengan mengidentifikasi masalah terkait pentingnya air bagi kehidupan manusia dengan masalah kualitas air yang layak diminum dan tidak layak diminum.

Tahap kedua merupakan studi literatur sebagai pengumpulan informasi yang terkait dengan penyusunan tugas akhir. Mengumpulkan informasi yang menunjang penelitian ini berupa jurnal, buku, referensi, dan sumber-sumber terpercaya yang lainnya.

Diskusi dan konsultasi tak luput dari metode penelitian dalam penyusunan tugas akhir ini, diskusi dan konsultasi dengan dosen pembimbing serta berbagai pihak yang ahli dalam bidang ini.

Tahap pengumpulan data dilakukan dengan teknik pengumpulan data yang menggunakan data dari *kaggle* tentang *Water Quality*. *Kaggle* merupakan *situs/platform* yang resmi untuk mengadakan perlombaan-perlombaan di bidang *Data Science*, dimana situs ini merupakan sumber pembelajaran *data science* (secara praktek).

Proses pengolahan data pada *software rapid miner* yang mencakup beberapa langkah, dimulai dari dataset, *preprocessing,* membagi kedua data ke dalam data *training* dan data *testing*, proses model *fit/classification, predict/apply model,* dan *result*. Proses pengolahan data yang dilakukan akan menghasilkan sebuah *result* atau hasil yang akan dibahas dan menghasilkan sebuah kesimpulan dalam proses penelitian yang dilakukan.

4. HASIL DAN PEMBAHASAN

Dataset

Dataset *Water Quality* diperoleh dari situs kaggle (www.kaggle.com/adityakadiwal/water-potability) [11]. Pada penelitian ini dataset yang digunakan adalah kualitas air dengan tipe data *csv* untuk proses klasifikasi dalam membandingkan hasil akurasi dari ketiga metode yang digunakaan yaitu *Naive Bayes, Decision Tree,* dan *KNN*. Hasil data yang diperoleh pada Tabel 1 Data Kualitas Air.

Tabel 1. Data Kualitas Air

Ph	Kekerasan	Padatan	Kloramin	Sulfat	Konduktivitas	Karbon_Organik	Tribalomethanes	Kekeruhan	Potabilitas
7.081	204.890	20791.319	7.300	368.516	564.309	10.380	86.991	2.963	0
3.716	129.423	18630.058	6.635	333.776	592.885	15.180	56.329	4.501	0
8.099	224.236	19909.542	9.276	333.776	418.606	16.869	66.420	3.056	0
8.317	214.373	22018.417	8.059	356.886	363.267	18.437	100.342	4.629	0
9.092	181.102	17978.986	6.547	310.136	398.411	11.558	31.998	4.075	0
5.582	188.313	28748.688	7.545	326.678	280.468	8.400	54.918	2.560	0
10.224	248.072	28749.717	7.513	393.663	283.652	13.790	84.604	2.673	0
8.636	203.362	13672.092	4.563	303.310	474.608	12.364	62.798	4.401	0
7.081	118.989	14285.584	7.804	268.647	389.376	12.706	53.929	3.595	0
11.180	227.231	25484.508	9.077	404.042	563.885	17.928	71.977	4.371	0
9.445	145.805	13168.529	9.444	310.583	592.659	8.606	77.577	3.875	1
9.025	128.097	19859.676	8.016	300.150	451.143	14.771	73.778	3.985	1
7.081	169.975	23403.637	8.520	333.776	475.574	12.924	50.862	2.747	1
6.800	242.008	39143.403	9.502	187.171	376.457	11.432	73.777	3.855	1
7.174	203.409	20401.102	7.682	287.086	315.550	14.534	74.406	3.940	1
7.658	236.961	14245.789	6.289	373.165	416.624	10.464	85.853	2.437	1
8.323	207.252	28049.646	8.827	297.813	358.726	18.709	60.911	4.052	1
5.934	223.858	23249.654	4.603	333.776	277.385	11.367	66.624	5.218	1
9.803	98.772	27357.457	9.218	323.199	512.429	14.169	59.454	2.765	1
6.102	215.268	15976.926	8.857	308.483	417.844	13.147	62.506	3.536	1

(www.kaggle.com/adityakadiwal/water-potability)

Besar jumlah data yang digunakan 2.081 baris data dengan sepuluh atribut diantaranya:

1. ph: pH 1 air (0 sampai 14)
2. *Hardness* (Kekerasan): kapasitas air untuk mengendapkan sabun dalam *mg/L*
3. *Solids* (Padatan): Total padatan terlarut dalam satuan *ppm*
4. *Chloramines* (Kloramin): Jumlah Kloramin dalam satuan *ppm*
5. *Sulfate* (Sulfat): Jumlah Sulfat yang terlarut dalam satuan *ppm*

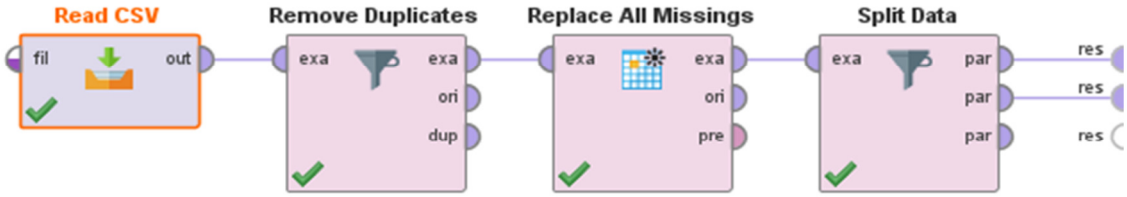
6. *Conductivity* (Konduktivitas): Konduktivitas listrik air dalam satuan *µS/cm*
7. *Organic Carbon* (Karbon Organik): Jumlah karbon organik dalam satuan *ppm*
8. *Trihalomethanes* (*Trihalomethanes*): Jumlah *Trihalomethanes* dalam satuan *µg/L*
9. *Turbidity* (Kekeruhan): Ukuran properti pemancar cahaya air dalam satuan *NTU*
10. *Potability* (sifat dapat diminum): menunjukkan apakah air aman untuk dikonsumsi manusia, dapat diminum 1 dan tidak dapat diminum 0. Atribut ph sampai dengan kekeruhan merupakan tipe data *real* dan atribut *potabilitas* merupakan tipe data *integer*.

Pre-processing dan *Labeling*

Data yang diperoleh pada penelitian ini perlu dilakukan proses *pre-processing*. Mengetahui sifat dari *text* data yang dikumpulkan sebelumnya maka dilakukan proses *labeling* data. *Attribute* yang diberi *label* dalam penelitian ini adalah *potability*, merupakan *attribute* yang menunjukkan apakah air aman untuk dikonsumsi manusia atau tidak. Air yang layak diminum 1 dan tidak dapat diminum 0. Proses *labeling* dapat dilakukan pengaturan warna pada label agar mempermudah proses penelitian. Beberapa metode *preprocessing* yang digunakan yaitu *Data Validation* untuk mendapatkan data yang baik dengan akurasi yang akurat**,** dilakukan peninjauan kembali tipe data yang diperoleh dan mengidentifikasi data agar tingkat akurasi yang diperoleh maksimal. Membuat data yang tidak konsisten menjadi konsisten dengan operator *replace all missing*. *Data Validation* mengidentifikasi sekaligus menghapus data yang tidak digunakan serta data yang non konsisten, dan data yang *missing*, dimana dari kondisi data yang mentah menjadi data yang siap diolah dan dapat dianalisis karena proses *cleansing* data dan *filtering* data pada proses *data validation.* Penelitian ini menggunakan metode *Data Integration dan Transformation* untuk meningkatkan hasil *accuracy* dari ketiga metode yang digunakan. Metode *Data Size Redution and Dicretization* digunakan untuk menghapus data yang terduplikat dengan menggunakan operator *remove duplicates.* Kondisi data awal sebanyak 3.276 menjadi 2.081 data yang clear karena adanya proses *Data Validation, Data Integration and Transformation,* dan *Data Size Redution and Dicretization* sehingga data dapat dianalisis untuk menghasilkan informasi data yang baru.

Data *Training* dan Data *Testing*

Proses pembagian data berupa data *training* dan data *testing* menggunakan *apply model*.



Gambar 2 *Split* Data

Operator *split* data berguna untuk melakukan pembagian antara data *training* dan data *testing*, rationya dibagi menjadi data *training* sebesar 0.7 (70%) dan data *testing* 0.3 (30%), menghasilkan jumlah data sebesar 1.456 untuk data *training* dan jumlah data sebesar 625 untuk data *testing*. Berdasarkan penelitian sebelumnya yang membahas perbandingan nilai akurasi dengan beberapa perbandingan nilai data training dan data testing yang berbeda, menghasilkan tingkat akurasi yang paling maksimal menggunakan data training 70% dan data testing 30% dengan tingkat akurasi sebesar 90.33% [12]. Data *training* dilakukan sebagai pembentuk model/pola/*knowledge* dan data *testing* dilakukan untuk pengujian model.

Pengukuran Akurasi dengan *Confusion Matrix*

Confusion Matrix merupakan metode klasifikasi berdasarkan hasil klasifikasi yang telah dilakukan, dimana akurasi klasifikasi mempengaruhi kinerja klasifikasi. *Confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. Pentingnya *confusion matrix* akan memberikan informasi seberapa baik model yang telah dibuat sebelumnya melalui pengukuran akurasi yang ada untuk mengetahui seberapa akurat model yang telah dibuat. *Confusion matrix* menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. *Confusion Matrix* digunakan untuk menghitung *accuracy*. *Confusion Matrix* ditampilkan pada Tabel 2 *Confusion Matrix* [13].

Tabel 2. *Confusion Matrix*

Kelas	Terklarifikasi Positif	Terklarifikasi Negatif
Positif	TP (*True Positive*)	FN (*False Negative*)
Negatif	FP (*False Positive*)	TN (*True Negative*)

Kinerja *Confusion Matrix* dapat diukur menggunakan dengan nilai TP, FP, FN, dan TN. *True Positive* merupakan data positif yang diprediksi benar. *True Negative* adalah data negatif yang diprediksi benar. *False Positive* adalah data negatif namun diprediksi sebagai data positif. *False Negative* adalah data positif namun diprediksi sebagai data negatif.
Mengitung *accuracy* menggunakan persamaan:

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

Hasil *Accuracy* Algoritma *Naive Bayes*

Klasifikasi data dengan algoritma *Naive Bayes* menghasilkan data yang disajikan dalam Tabel 3 *Confusion Matrix Naïve Bayes.*

Tabel 3. *Confusion Matrix Naïve Bayes*

***accuracy*: 63.60%**

	true 0	*true* 1	*class precision*
pred. 0	788	430	64.70%
pred. 1	100	138	57.98%
class recall	88.74%	24.30%	

Hasil akurasi sebesar 63.60%, dengan *class precision* untuk *pred.* nol *(pred. negative)* adalah 64.70% dan *pred* satu *(pred.positive)* adalah 57.98%. Hasil *accuracy* didapatkan menggunakan persamaan 4, dimana nilai *true positive* sebanyak 788, *true negative* sebanyak 138, *false negative* sebanyak 430, dan *false positive* 100. Hasil akurasi dapat dibuktikan dengan:

$$accuracy = \frac{788 + 138}{788 + 138 + 430 + 100} = 63.60\%$$

Performance Vektor:

Tabel 4. *Performance Vektor Naïve Bayes*

PerformanceVector:		
accuracy: 63.60%		
ConfusionMatrix:		
True	0	1

0	788	430
1	100	138

Performance Vector sendiri adalah bentuk deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan. Nilai *True Positive* sebanyak 788, merupakan nilai data positif yang artinya air aman untuk diminum dan diprediksi memiliki nilai yang benar. Nilai *False Positive* sebanyak 100, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif. Nilai *False Negative* sebanyak 430, data positif namun diprediksi sebagai data negatif. Nilai *True Negative* sebanyak 138, merupakan data negatif yang diprediksi benar.

Hasil *Accuracy* Algoritma *Decision Tree*

Klasifikasi data dengan algoritma *decision tree* menghasilkan data yang disajikan dalam Tabel 5 *Confusion Matrix Decision Tree*.

Tabel 5. *Confusion Matrix Decision Tree*

***accuracy*: 80.84%**

	true 0	*true* 1	*class precision*
pred. 0	817	208	79.71%
pred. 1	71	360	83.53%
class recall	92.00%	63.38%	

Hasil akurasi sebesar 80.84%, dengan *class precision* untuk *pred.* nol *(pred. negative)* adalah 79.71% dan *pred* satu *(pred.positive)* adalah 83.53%. Hasil *accuracy* yang didapatkan menggunakan persamaan 4, dimana nilai *true positive* sebanyak 817, *true negative* sebanyak 360, *false negative* sebanyak 208, dan *false positive* 71. Hasil akurasi dapat dibuktikan dengan:

$$accuracy = \frac{817 + 360}{817 + 360 + 208 + 71} = 80.84\%$$

Performance Vector

Tabel 6. Hasil *Performance Vektor Decision Tree*

PerformanceVector:		
accuracy: 80.84%		
ConfusionMatrix:		
True	0	1
0	817	208
1	71	360

Performance Vector sendiri adalah bentuk deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan. Nilai *True Positive* sebanyak 817, merupakan nilai data positif yang artinya air aman untuk diminum dan diprediksi memiliki nilai yang benar. Nilai *False Positive* sebanyak 71, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif. Nilai *False Negative* sebanyak 208, data positif namun diprediksi sebagai data negative. Nilai *True Negative* sebanyak 360, merupakan data negatif yang diprediksi benar.

Hasil *Accuracy* Algoritma *K-nearest neighbors*

Klasifikasi data dengan algoritma *K-nearest neighbors* menghasilkan data yang disajikan dalam Tabel 7 *Confusion Matrix KNN.*

Tabel 7. *Confusion Matrix KNN*

accuracy: 86.88%

	true 0	*true* 1	*class precision*
pred. 0	836	139	85.74%
pred. 1	52	429	89.19%
class recall	94.14%	75.53%	

Hasil akurasi didapatkan sebesar 86.88%, dimana *class precision* untuk *pred.* nol *(pred. negative)* adalah 85.74% dan *pred* satu *(pred.positive)* adalah 89.19%. Hasil *accuracy* yang didapatkan menggunakan persamaan 4, dimana nilai *true positive* sebanyak 836, *true negative* sebanyak 429, *false negative* sebanyak 139, dan *false positive* 52. Hasil akurasi dapat dibuktikan dengan:

$$accuracy = \frac{836 + 429}{836 + 429 + 139 + 52} = 86.88\%$$

Performance Vector:

Tabel 8. Hasil *Performance Vektor KNN*

PerformanceVector:		
accuracy: 86.88%		
ConfusionMatrix:		
True	0	1
0	836	139
1	52	429

Performance Vector merupakan bentuk deskripsi dari tabel hasil analisis yang diperoleh dalam penelitian yang dilakukan. Nilai *True Positive* (TP) memiliki nilai sebanyak 836, merupakan nilai data positif yang artinya air aman untuk diminum dan diprediksi memiliki nilai yang benar. Nilai *False Positive* sebanyak 52, dimana data negatif (air tidak dapat diminum) namun diprediksi sebagai data positif. Nilai *False Negative* sebanyak 139, data positif namun diprediksi sebagai data negatif. Nilai *True Negative* sebanyak 429, merupakan data negatif yang diprediksi benar.

Proses klasifikasi data menggunakan menggunakan beberapa operator untuk menjalankan metode *classification*, diantaranya terdapat *Read CSV, Split data, Apply Model,* dan *Performance. Metode Classification* seperti *KNN, Naïve Bayes, dan Decision Tree*. Operator tersebut memiliki fungsinya masing-masing, *Read CSV* berfungsi untuk *import* data *CSV* yang sudah didapatkan, didalam *Read CSV* dilakukan metode *pre-processing* dimana *pre-processing* berfungsi untuk melihat dataset yang di *import*, apakah terjadi data yang tidak konsisten atau *missing value*. Operator *Split* data berfungsi untuk mengambil *Example Set* sebagai inputnya dan mengirimkan *subset* dari *ExampleSet* tersebut melalui *port outputnya. Apply Model* berfungsi pengaplikasian metode *classification. Performance* berfungsi untuk melihat akurasi dari semua jenis metode *classification.*

Hasil Akurasi

Tabel 8. Perbandingan Hasil Akurasi

Naïve-Bayes	*K-nearest neighbors*	*Decision tree*
63.60%	86.88%	80.84%

Analisis perbandingan akurasi *Water Quality* menggunakan data hasil klasifikasi dengan *K-nearest neighbors, Naïve Bayes,* dan *Decision Tree* menunjukkan bahwa *K-nearest neighbors* merupakan metode yang menghasilkan tingkat akurasi paling tinggi yaitu 86.88% untuk klasifikasi data kualitas air yang digunakan dalam penelitian ini, sedangkan *Naïve-Bayes sebesar* 63.60% dan *Decision tree* sebesar 80.84%.

5. KESIMPULAN

Tujuan dilakukan penelitian ini untuk mengetahui hasil perbandingan tingkat keakuratan dari metode penelitian yang digunakan yaitu *K-nearest neighbors, Naïve Bayes,* dan *Decision Tree*. Dilihat dari *Class Recall* dan *Class Precision* metode yang menghasilkan tingkat keakuratan yang paling tinggi adalah *Decision Tree* yaitu sebesar 86.88%. Metode klasifikasi *Decision Tree* dan *KNN* pada penelitian ini cukup baik digunakan karena menghasilkan tingkat akurasi diatas 80%, namun untuk mendapatkan hasil akurasi yang lebih maksimal untuk penelitian selanjutnya bisa menggunakan metode yang lain

6. SARAN

Pada penelitian ini digunakan tiga metode yakni *K-nearest neighbors, Naïve Bayes*, dan *Decision Tree* dan menggunakan data sebanyak 2.081 data, oleh karena itu saran untuk penelitian selanjutnya bisa menggunakan lebih banyak metode lagi dan menggunakan data yang lebih banyak lagi untuk menghasilkan tingkat akurasi yang lebih tinggi.

UCAPAN TERIMA KASIH

Pada penelitian ini penulis mengucapkan terima kasih kepada Bapak Evangs Mailoa, S.Kom., M.Cs. selaku dosen pembimbing saya yang telah mengarahkan saya sehingga penelitian ini dapat selesaikan, penulis juga mengucapkan terima kasih kepada orang tua atas bantuan dalam melakukan penelitian ini.

DAFTAR PUSTAKA

[1] S. Wahyuningsih and D. R. Utari, *"Perbandingan Metode K-Nearest Neighbor , Naive Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit," Konf. Nas. Sist. Inf. 2018 STMIK Atma Luhur Pangkalpinang, 8 – 9 Maret 2018*, pp. 619–623, 2018.

[2] R. Puspita and A. Widodo, *"Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," J. Inform. Univ. Pamulang*, Vol. 5, No. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[3] M. A. Rahman, N. Hidayat, and A. Afif Supianto, *"Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Vol. 2, No. 12, Desember 2018, hlm. 6346-6353 e-ISSN*, Vol. 2, No. 12, pp. 925–928, 2018.

[4] N. T. Romadloni, I. Santoso, and S. Budilaksono, *"Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl," J. IKRA-ITH Inform.*, Vol. 3, No. 2, pp. 1–9, 2019.

[5] P. N. Harahap and S. Sulindawaty, *"Implementasi Data Mining Dalam Memprediksi Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus PT.Arma Anugerah Abadi Cabang Sei Rampah)," Matics*, Vol. 11, No. 2, p. 46, 2020, doi: 10.18860/mat.v11i2.7821.

[6] D. Cahyanti, A. Rahmayani, and S. A. Husniar, *"Analisis Performa Metode Knn pada Dataset Pasien Pengidap Kanker Payudara," Indones. J. Data Sci.*, Vol. 1, No. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[7] Z. Zulfauzi and M. N. Alamsyah, "Penerapan Algoritma Naive Bayes Untuk Prediksi Penerimaan Mahasiswa Baru Studi Kasus Universitas Bina Insan Fakultas Komputer," *J. Teknol. Inf. Mura*, Vol. 12, No. 02, pp. 156–165, 2020, doi: 10.32767/jti.v12i02.1096.

[8] U. I. Lestari, *"Penerapan Metode K-Nearest Neighbor Untuk Sistem Pendukung Keputusan Identifikasi Penyakit Diabetes Melitus," JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, Vol. 8, No. 4, pp. 2071–2082, 2021, doi: 10.35957/jatisi.v8i4.1235.

[9] F. Yulian Pamuji, V. Puspaning Ramadhan, and R. Artikel, *"Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy Info Artikel Abstrak," J. Teknol. dan Manaj. Inform.*, Vol. 7, No. 1, pp. 46–50, 2021, [Online]. Available: http://http//jurnal.unmer.ac.id/index.php/jtmi.

[10] E. Muningsih, *"Kombinasi Metode K-Means dan Decision Tree Dengan Perbandingan Kriteria dan Split Data," J. Teknoinfo*, Vol. 16, No. 1, p. 113, 2022, doi: 10.33365/jti.v16i1.1561.

[11] *"Water Quality,"* [Online]. Available: https://www.kaggle.com/adityakadiwal/water-potability.

[12] M. Fatchan and H. Sugeng, *"Anlisa Terpilihnya Tri Rismaharini Sebagi Menteri Sosial Dengan Pendekatan Algorithma Naïve Bayes,"* Vol. 1, No. 2, pp. 50–57, 2022.

[13] Sofyan Vivi Dwiyanu, *"Klasifikasi Kelulusan Mahasiswa Program Studi Teknik Informatika Menggunakan Algoritma K-Nearest Neighbor,"* 2019.