Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 12, Desember 2018, hlm. 6346-6353 http://j-ptiik.ub.ac.id

Komparasi Metode Data Mining *K-Nearest Neighbor* Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)

Maulana Aditya Rahman[1], Nurul Hidayat[2], Ahmad Afif Supianto[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]maulanaadityar@gmail.com,[2]ntayadih@ub.ac.id, [3]afif.supianto@ub.ac.id

Abstrak

Air adalah merupakan senyawa kimia yang sangat dibutuhkan bagi kelangsungan hidup makhluk hidup yang ada di bumi. Wilayah terluas di planet bumi merupakan air yang menutupi hampir 71% wilayah yang ada di bumi. Air juga merupakan zat yang sangat penting yang ada di bumi yang sangat dibutuhkan oleh semua makhluk hidup mulai dari tumbuhan, hewan dan manusia. Dibutuhkan pengawasan dan pengolahan lingkungan sekitar sumber air sehingga dapat menghasilkan kualitas air yang bersih sesuai dengan standar kualitas air bersih dan memenuhi standar air yang layak dikonsumsi oleh manusia. Untuk menentukan klasifikasi kualitas air bersih terdapat banyak metode yang dapat digunakan. Untuk memilih metode klasifikasi yang paling cocok, dapat dilakukan komparasi antara beberapa metode. Penelitian ini melakukan komparasi antara metode *K-Nearest Neighbor* dan *Naïve Bayes*. Berdasarkan dari beberapa penelitian, metode *K-Nearest Neighbor* dan *Naïve Bayes* merupakan metode yang cukup baik dan menghasilkan tingkat akurasi yang cukup tinggi. Berdasarkan hasil pengujian, diperoleh rata-rata nilai akurasi metode *K-Nearest Neighbor* sebesar 82.42% dan rata-rata nilai akurasi metode *Naïve Bayes* sebesar 70.32%. Dapat disimpulkan bahwa metode yang paling baik untuk klasifikasi kualitas air bersih adalah metode *K-Nearest Neighbor*.

Kata kunci: *kualitas air bersih, data mining, klasifikasi, komparasi metode, k-nearest neighbor, naïve bayes*

Abstract

Water is a chemical compound that is needed for the survival of living things on earth. The widest area on planet earth is water that covers almost 71% of the region on earth. Water is also a very important substance on earth that is needed by all living things from plants, animals and humans. It takes the supervision and processing of the environment around the water source so as to produce clean water quality in accordance with the standard of clean water quality and meet the standard of water that is suitable for human consumption. To determine the classification of clean water quality there are many methods that can be used. To choose the best classification method, it can be comparated between several methods. This study comparing the K-Nearest Neighbor and Naïve Bayes methods. Based on several studies, the K-Nearest Neighbor and Naïve Bayes methods are quite good and yield a high degree of accuracy. Based on the test result, the average accuracy value of K-Nearest Neighbor method is 82.42% and the average accuracy of Naïve Bayes method is 70.32%. It can be concluded that the best method for water quality classification is K-Nearest Neighbor method.

Keywords: *water quality, data mining, classification, comparison method, k-nearest neighbor, naïve bayes.*

1. PENDAHULUAN

Air adalah merupakan senyawa kimia yang sangat dibutuhkan bagi kelangsungan hidup makhluk hidup yang ada di bumi. Wilayah terluas di planet bumi merupakan air yang menutupi hampir 71% wilayah yang ada di bumi. Di bumi ketersediaan volume air sebesar 1,4 triliun kilometer kubik. Terdapat banyak sumber air yang ada di bumi seperti air laut, air sungai, air permukaan, air atmosfir, air rawa dan air tanah.

Air juga merupakan zat yang sangat penting yang ada di bumi yang sangat dibutuhkan oleh semua makhluk hidup mulai dari tumbuhan, hewan dan manusia. Tumbuhan memerlukan air

sebagai salah satu senyawa dalam proses fotosintesis. Hewan membutuhkan air untuk proses pencernaan makanan dan sebagai tempat tinggal. Air juga dibutuhkan manusia untuk keperluan sehari-hari seperti memasak, mencuci, mandi dan lain-lainnya. Oleh karena itu air sering disebut sebagai sumber kehidupan yang mana disitu ada air maka disitu pula terdapat kehidupan.

Air merupakan senyawa kompleks karena mengandung zat-zat dan mineral-mineral di dalamnya. Namun tidak semua zat dan mineral yang terkandung di air dapat dicerna dan diterima dengan baik oleh tubuh manusia. Air juga rentan terkontaminasi oleh bakteri-bakteri dan zat mineral yang berbahaya bagi tubuh manusia. Hal tersebut bisa terjadi dikarenakan tercemarnya sumber air atau tercemarnya lingkungan di sekitar sumber air. Begitu pentingnya peranan air dalam kehidupan sehingga dapat dinyatakan bahwa kualitas air dapat digunakan sebagai indikator tingkat kesehatan manusia(Situmorang, 2017). Dibutuhkan pengawasan dan pengolahan lingkungan sekitar sumber air sehingga dapat menghasilkan kualitas air yang bersih sesuai dengan standar kualitas air bersih dan memenuhi standar air yang layak dikonsumsi oleh manusia.

Untuk mengetahui bahwa air tersebut memeliki kualitas yang sesuai syarat kesehatan dapat diketahui dari zat-zat atau mineral yang terkandung didalamnya. Namun dalam penentuan kualitas air masih menggunakan perhitungan manual seperti *Water Quality Index (WQI)* dan *STORET*. Pada metode *STORET* masih menggunakan metode penghitungan secara manual dengan menghitung satu-persatu data parameter. Kelemahan dari metode ini membutuhkan waktu yang cukup lama yaitu 1 sampai 30 hari sesuai dengan parameter apa yang diukur dan diteliti dan biaya yang digunakan cukup mahal. Sehingga, dalam mengatasi permasalahan klasifikasi terhadap kualitas air, peneliti mengusulkan penggunaan metode klasifikasi data dan dapat memberikan solusi dalam membantu proses penetuan terhadap klasifikasi kualitas air yang lebih efektif dan efisien.

Untuk menentukan klasifikasi kualitas air bersih terdapat banyak metode yang dapat digunakan seperti *K-Nearest Neighbor*, *Naïve Bayes*, *K-Means*, *Decision Tree* dan lain sebagainya. Namun untuk memilih metode yang paling cocok, dapat dilakukan komparasi antara beberapa metode. Dalam penelitian sebelumnya oleh Rifwan Hamidi dengan menggunakan metode *Learning Vector Quantization* untuk klasifikasi kualitas air sungai menggunakan parameter sejumlah 7 masukan dengan menghasilkan akurasi sebesar 81.13%. Penelitian lain yang dilakukan oleh Mila Listiana pada tahun 2015 dalam kasus identifikasi tumbuh kembang anak balita dengan perbandingan algoritme *Decision Tree(C4.5)* dan *Naïve Bayes*, diperoleh hasil pengujian rata-rata nilai akurasi metode *Naïve Bayes* sebesar 96,89% dan algoritme *Decision Tree(C4.5)* sebesar 89,78%.

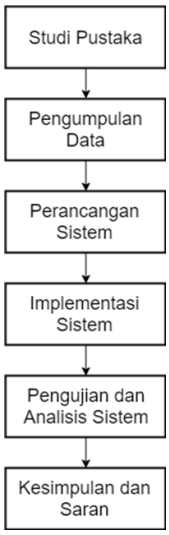
Penelitian lain yang dilakukan oleh Khafiizh Hastuti pada tahun 2012 pada prediksi data mahasiswa nonaktif dengan melakukan perbandingan metode klasifikasi *Logistic Regression, Decision Tree, Neural Network* dan *Naïve Bayes*. Penelitian tersebut menghasilkan tingkat akurasi metode *Logistic Regression* sebesar 81,64%, metode *Decision Tree* sebesar 95,29%, metode *Neural Network* sebesar 94,56% dan metode *Naïve Bayes* sebesar 93,47%. Selain itu terdapat penelitian lain yang dilakukan oleh Santoso pada tahun 2016 dengan melakukan komparasi metode *K-Nearest Neighbor* dan *Learning Vector Quantization (LVQ)* dengan studi kasus peramalan klasifikasi tingkat kemiskinan menghasilkan akurasi metode *K-Nearest Neighbor* sebesar 93.52% dan *Learning Vector Quantization (LVQ)* sebesar 75.93%.

Pada penelitian sebelumnya terdapat perbedaan tingkat akurasi beberapa metode untuk klasifikasi. Berdasarkan dari beberapa penelitian sebelumnya, metode *K-Nearest Neighbor* dan *Naïve Bayes* merupakan metode yang memiliki akurasi cukup tinggi. Untuk itu penelitian ini, peneliti akan melakukan komparasi antara metode *K-Nearest Neighbor* dan metode *Naïve Bayes* untuk mengetahui metode mana yang lebih baik dalam membantu melakukan klasifikasi terhadap kualitas air bersih. Berdasarkan uraian latar belakang, maka judul yang diambil dalam skripsi ini adalah "Komparasi Metode Data Mining *K-Nearest Neighbor* dan *Naïve Bayes* untuk Klasifikasi Kualitas Air Bersih".

2. METODOLOGI

Metodologi penelitian akan dibahas secara sistematik melalui langkah-langkah yang

spesifik untuk digunakan dalam menyelesaikan masalah penelitian. Tahap-tahap penelitian ini disajikan pada Gambar 1.



Gambar 1. Diagram Metode Tahapan Penelitian

2.1. Pengumpulan Data

Data yang digunakan dalam penilitian merupakan data kualitas air bersih yang diperoleh dari Kantor PDAM Tirta Kencana Kabupaten Jombang dalam jangka waktu tahun 2016 hingga tahun 2017.

Contoh data klasifikasi kualitas air bersih akan ditunjukan pada Tabel 1.

Tabel 1. Contoh Data Klasifikasi Air Bersih

No.	*Coliform*	*E. Coli*	Mangan	TDS	Khlorida	Kelas
1	0	0	0	152	12,29	Sesuai Syarat
2	0	0	0	72	12,9	Sesuai Syarat
3	23	0	0,5	201	9,4	Tidak Sesuai
4	23	0	0	235	15,8	Tidak Sesuai
5	0	0	2	225	17,3	Sesuai Syarat

2.2. Algoritme *Naïve Bayes*

Naïve Bayes merupakan metode pengklasifikasian suatu probabilitas dan statistik yang diperoleh Thomas Bayes seorang ilmuwan Inggris dengan cara melakukan prediksi peluang di masa depan berdasarkan pengalaman pada masa sebelumnya(Bustami, 2013). *Naïve Bayes* merupakan sebuah teknik prediksi probabilitas sederhana yang berdasarkan pada penerapan teorema *bayes* (aturan *bayes*) yang memiliki ketidakterkaitan antara suatu fitur dengan fitur lain dalam suatu data(Prasetyo, 2012).

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E)} \quad (1)$$

Keterangan:

- H adalah hipotesis data X yang merupakan suatu kelas spesifik
- E adalah data dengan kelas yang belum diketahui
- P(H|E) adalah probabilitas hipotesis H berdasar *evidence*/bukti E (*posterior probability*).
- P(H) adalah probabilitas hipotesis H (*prior probability*)
- P(E|H) adalah probabilitas *evidence* E berdasar kondisi hipotesis H.
- P(E) adalah probabilitas dari *evidence* E

2.3. Algoritme *K-Nearest Neighbor*

Prinsip kerja algoritme *K-Nearest Neighbor* (KNN) adalah mencari jarak terdekat dengan *k* tetangga (*neighbor*) terdekat dalam data *training* dengan data yang akan diuji. Teknik mengelompokkan data baru dengan cara menghitung jarak data baru ke beberapa data/tetangga (*neighbor*) terdekat. Algoritme *K-Nearest Neighbor* merupakan *instead-based learning*, dimana data *training* disimpan sehingga klasifikasi untuk *record* baru yang belum diklasifikasi dapat ditemukan dengan membandingkan kemiripan yang paling banyak dalam data *training* (Kustiyahningsih, 2013).

Untuk menghitung jarak dalam *K-Nearest Neighbor* digunakan fungsi *Euclidean Distance*

$$euc = \sqrt{\sum_{i=1}^{n}(x_{2i} - x_{1i})^{\,2}} \quad (2)$$

Keterangan:

- X_2 = data latih
- X_1 = data uji
- i = variabel data
- n = dimensi data

2.4. Akurasi

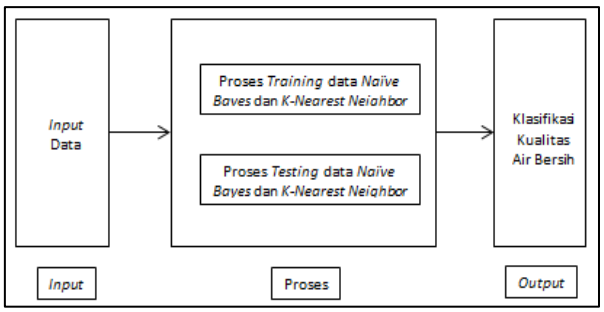
Pengujian akurasi adalah suatu ukuran seberapa dekat hasil pengukuran terhadap angka sebenarnya. Akurasi dapat diperoleh dari persentase kebenaran, yaitu perbandingan antara jumlah data benar dengan keseluruhan data. Akurasi dinyatakan dengan rumus

$$akurasi = \frac{total\ data\ benar}{total\ data}\ x\ 100\% \quad (2)$$

3. PERANCANGAN SISTEM

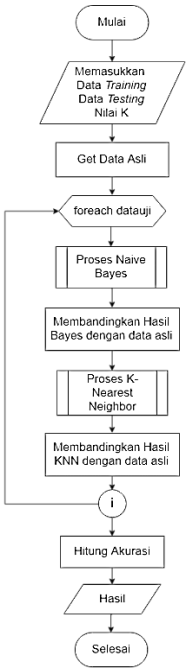
Langkah - langkah penelitian yang dilakukan secara terstruktur mulai dari memasukkan data hingga memperoleh hasil. Langkah-langkah tersebut terdiri dari tiga proses, yaitu proses input, proses perhitungan

dan proses output. Gambaran umum tiga proses tersebut akan dijelaskan pada Gambar 2.



Gambar 2. Diagram Perancangan Sistem

Proses perhitungan semua data dimulai ketika pengguna memasukkan jumlah data *training*, jumlah data *testing*, dan nilai *k*. Data training dan data testing tersebut dihitung menggunakan *K-Nearest Neighbor* dan *Naïve Bayes*. Selanjutnya dihitung akurasi yang dihasilkan dari kedua metode tersebut. Hasil yang ditampilkan berupa kesesuaian antara kedua metode tersebut dengan data asli serta akurasi sistem. Semua proses tersebut akan dijelaskan pada Gambar 3.



Gambar 3. Diagram Proses Perhitungan Semua Data

4. PENGUJIAN DAN ANALISIS

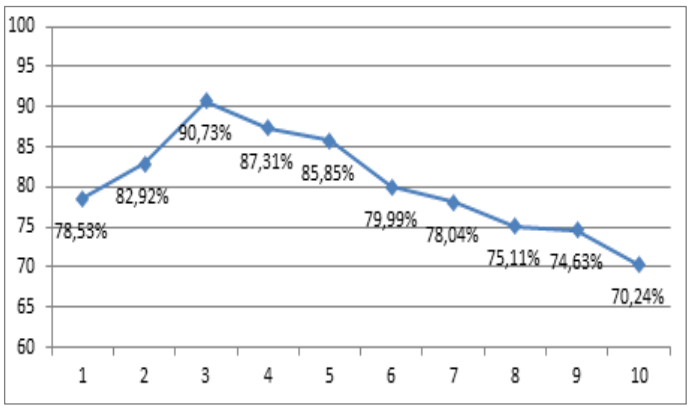
5.1. Pengujian Berdasarkan Nilai *K*

Pengujian ini dilakukan untuk mengetahui pengaruh nilai atribut *K* terhadap metode klasifikasi *K-Nearest Neighbor*. Pada pengujian ini dilakukan sebanyak 10 kali mengubah nilai *K* dalam setiap uji coba dimana setiap skenario pengujian *K* dilakukan sebanyak 5 kali. Hasil dari pengujian dengan nilai *K* terbaik akan digunakan dalam pengujian komparasi metode *K-Nearest Neighbor* dan *Naïve Bayes*. Nilai *K* yang digunakan pada pengujian komparasi nanti hanya menggunakan angka terkecil dengan akurasi terbaik dikarenakan untuk mempermudah penghitungan komaparasi selanjutnya. Data *training* yang yang digunakan pada pengujian sejumlah 127 data dan data *testing* sejumlah 40 data. Hasil pengujian berdasarkan nilai atribut *K* dapat dilihat pada Tabel 2. dan grafik pada Gambar 4.

Tabel 2. Hasil Pengujian Berdasarkan Nilai *K*

K	Uji Coba Ke-	Akurasi (%)
1	1	78.05%
	2	78.05%
	3	80.48%
	4	78.05%
	5	78.05%
	Rata-Rata	**78.53%**
2	1	85.36%
	2	85.36%
	3	85.36%
	4	80.48%
	5	78.04%
	Rata-Rata	**82.92%**
3	1	90.24%
	2	92.68%
	3	90.24%
	4	90.24%
	5	90.24%
	Rata-Rata	**90.73%**
4	1	87.8%
	2	87.8%
	3	87.8%
	4	90.24%
	5	82.92%
	Rata-Rata	**87.31%**
5	1	87.8%
	2	85.36%
	3	85.36%
	4	85.36%
	5	85.36%
	Rata-Rata	**85.85%**
6	1	80.48%
	2	82.92%
	3	78.04%
	4	80.48%
	5	78.04%
	Rata-Rata	**79.99%**
7	1	78.04%
	2	78.04%
	3	80.48%
	4	75.6%
	5	78.04%
	Rata-Rata	**78.04%**
8	1	73.17%
	2	75.6%
	3	75.6%
	4	75.6%
	5	756%
	Rata-Rata	**75.11%**
9	1	73.17%

	2	75.6%
	3	75.6%
	4	73.17%
	5	75.6%
	Rata-Rata	**74.63%**
	1	70.73%
	2	68.29%
10	3	70.73%
	4	70.73%
	5	70.73%
	Rata-Rata	**70.24%**



Gambar 4. Grafik Hasil Pengujian Berdasarkan Nilai *K*

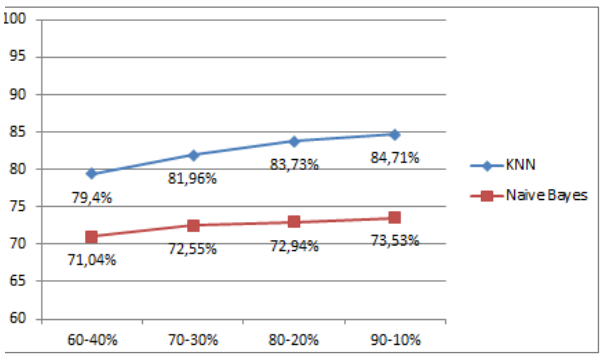
Pada pengujian berdasarkan nilai atribut *K* data yang digunakan sejumlah 167 data yang terbagi menjadi 127 data *training* dan 40 data *testing*. Hasil pengujian terendah terjadi ketika nilai atribut *K* bernilai 10 yaitu 70,24%. Hasil pengujian paling tinggi dan terbaik ketika nilai atribut *K* bernilai 3, dimana pada pengujian nilai tertinggi sebesar 90.73%. Pada Gambar 5.1 dapat ditarik kesimpulan bahwa akurasi *K-Nearest Neighbor* akan dipengaruhi terhadap jumlah nilai *K*. Semin banyak nilai *K*, maka semakin rendah tingkat akurasi, hal ini disebabkan oleh atribut yang digunakan memiliki kemiripan yang banyak sehingga semakin banyak tetangga atau nilai *K* yang diambil semakin banyak data dari kelas yang lain ikut dijadikan pertimbangan keputusan. Pada pengujian yang dilakukan peneliti akurasi yang terbaik didapatkan ketika *K* bernilai 3 yang nantinya akan digunakan dalam pengujian komparasi metode *K-Nearest Neighbor* dan *Naïve Bayes.*

5.2. Pengujian Berdasarkan Rasio Perbandingan atau *Percentage Split*

Pengujian berdasarkan rasio perbandingan atau *percentage split* menggunakan data sejumlah 100% dari keseluruhan data atau sejumlah 167 data yang akan dibagi berdasarkan rasio perbandingan yang ditentukan yaitu menggunakan 90% data *training* dan 10% data *testing*, 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, 60% data *training* dan 40% data. Dengan menggunakain nilai atribut *K* = 3. Pengujian ini dilakukan untuk mengetahui pengaruh rasio atau persentase tertentu dari jumlah data *training* dan data *testing* terhadap tingkat akurasi metode *K-Nearest Neighbor* dengan metode *Naïve Bayes*. Hasil pengujian dapat ditunjukan pada Tabel 3 dan grafik pada Gambar 5

Tabel 3. Hasil Pengujian Berdasarkan *Percentage Split*

Training	*Testing*	Uji Coba Ke-	*KNN (%)*	*NB (%)*
		1	85.07%	70.15%
		2	79.1%	68.66%
60%	40%	3	76.12%	73.13%
		4	76.12%	74.62%
		5	80.6%	68.66%
	Rata-Rata		**79.40%**	**71.04%**
		1	82.35%	76.47%
		2	80.4%	72.55%
70%	30%	3	78.43%	68.63%
		4	80.4%	80.4%
		5	88.24%	64.71%
	Rata-Rata		**81.96%**	**72.55%**
		1	91.18%	79.41%
		2	82.35%	79.41%
80%	20%	3	76.47%	64.71%
		4	80.4%	72.55%
		5	88.24%	68.63%
	Rata-Rata		**83.73%**	**72.94%**
		1	82.35%	76.47%
		2	88.24%	76.47%
90%	10%	3	82.35%	64.71%
		4	82.35%	79.41%
		5	88.24%	70.59%
	Rata-Rata		**84.71%**	**73,53%**
	Rata-Rata Total		82.45%	72.52%



Gambar 5. Grafik Hasil Pengujian Berdasarkan *Percentage Split*

Hasil pengujian berdasarkan rasio perbandingan atau *percentage split* metode *K-Nearest Neighbor*_memiliki rata-rata tingkat

akurasi paling tinggi pada rasio data *training* dan data *testing* 90%-10% yaitu sebesar 84.71% dan rata-rata akurasi terendah pada rasio data *training* dan data *testing* 60%-40% sebesar 79.40%. Sedangkan untuk metode *Naïve Bayes* rata-rata tingkat akurasi paling tinggi sebesar 73.53% juga pada rasio data *training* dan data *testing* 90%-10% dan rata-rata terendah pada rasio 60%-40% sebesar 71.04%.

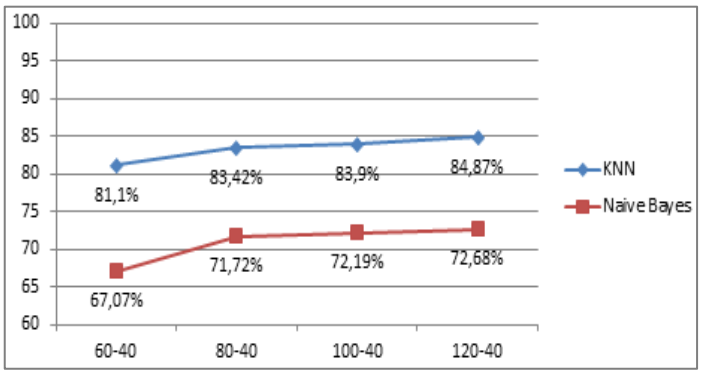
Pada grafik pada Gambar 5 diperlihatkan bahwa semakin besar selisih persentase atau rasio antara data *training* dan data *testing* maka semakin tinggi pula akurasi yang didapatkan. Hal tersebut dikarenakan jumlah data *training* yang semakin lebih banyak daripada jumlah data *testing* yang semakin lebih sedikit maka model *classifier* yang dibangun berdasarkan fakta dari data *training* akan lebih baik dan lebih lengkap untuk melakukan prediksi terhadap klasifikasi data baru atau data *testing*. Sehingga akurasi yang didapatkan akan jauh lebih baik juga.

5.3. Pengujian Berdasarkan Jumlah Data *Training*

Pengujian berdasarkan variasi jumlah data *training* ini berbeda dengan pengujian sebelumnya yaitu *percentage split* dimana pada pengujian sebelumnya data yang digunakan 100% dari keseluruhan data. Pada pengujian ini tidak menggunakan data seluruhnya, hanya menggunakan beberapa data yang nantinya dibagi menjadi beberapa data *training* dengan jumlah data *training* yaitu 60 data, 80 data, 100 data, dan 120 data. Dan menggunakan jumlah data testing yang sama yaitu 40 data, hal ini disebabkan pengujian ini hanya difokuskan terhadap jumlah data *training*. Dengan menggunakan nilai *K = 3*.Pengujian ini bertujuan untuk mengetahui pengaruh jumlah data *training* terhadap tingkat akurasi yang dihasilkan tanpa melihat data testing pada metode *K-Nearest Neighbor* dan metode *Naïve Bayes*. Hasil pengujian berdasarkan variasi jumlah data training dapat dilihat seperti pada Tabel 4 dan Gambar 6.

Tabel 4. Hasil Pengujian Berdasarkan Data *Training*

Training	*Testing*	Uji Coba Ke-	*KNN (%)*	*NB (%)*
		1	85.37%	70,73%
		2	85.37%	70.73%
60	40	3	75.6%	68.29%
		4	78.05%	63.41%
		5	85.37%	65.85%
	Rata-Rata		81,1%	67,07%
		1	80.49%	73.17%
80	40	2	87.8%	78.05%
		3	80.49%	63.41%
		4	82.93%	80.49%
		5	85.37%	63.41%
	Rata-Rata		83,42%	71,72%
		1	87.8%	65.85%
		2	87.8%	75.6%
100	40	3	82.93%	80.49%
		4	80.49%	73.17%
		5	80.49%	65.85%
	Rata-Rata		83,9%	72,19%
		1	90.24%	78.05%
		2	85.37%	70.73%
120	40	3	82.92%	65.85%
		4	82.92%	70.73%
		5	82.92%	78.05%
	Rata-Rata		84,87%	72,67%
	Rata-Rata Total		83,32%	70,91%



Gambar 6. Grafik Hasil Pengujian Berdasarkan Data *Training*

Hasil pengujian berdasarkan jumlah data *training* metode *K-Nearest Neighbor* memperoleh rata-rata nilai akurasi tertinggi sebesar 84.87% dengan data training sebanyak 120 data dan rata-rata nilai akurasi terendah sebesar 81.1% dengan jumlah data 60 data *training*. Rata-rata akurasi dari metode *K-Nearest Neighbor* adalah 83.32%. Sedangkan hasil pengujian metode *Naïve Bayes* diperoleh rata-rata nilai akurasi tertinggi sebesar 72.68% dengan data *training* sebanyak 120 data dan rata-rata nilai akurasi terendah sebesar 67.07% dengan jumlah data 60 data *training*. Rata-rata akurasi dari metode *Naïve bayes* adalah 70.91%. Pada grafik Gambar 6 diketahui bahwa semakin banyak jumah data *training* semakin tinggi pula tingkat akurasi metode.

Sama halnya dengan pengujian berdasarkan rasio perbandngan atau *percentage split*, semakin banyak jumlah data *training* maka semakin tinggi pula hasil akurasi yang didapatkan. Hal tersebut dikarenakan jika semakin banyak data *training* yang digunakan maka semakin baik dan semakin lengkap juga model *classifier* yang dibentuk berdasarkan fakta dari data *training* tersebut. Sehingga pada saat melakukan prediksi pada klasifikasi data

testing atau data baru, akurasi yang didapatkan akan semakin baik atau tinggi.

6. KESIMPULAN

Berdasarkan hasil dari pengujian dan analisis dapat ditarik kesimpulan sebagi berikut:

1. Implementasi algoritme K-Nearest Neighbor dan Naive Bayes pada klasifikasi kualitas air bersih, atribu-atribut yang digunakan dalam mebangun sistem berupa atribut komposisi pada air bersih meliputi Coliform, Escherichia Coli, Mangan, TDS dan Khlorida. Langkah-langkah komparasi metode untuk klasifikasi kualitas air bersih sebagai berikut :
   - Memasukkan jumlah data *training*, data *testing* dan nilai *K*
   - Mengambil data asli dari *database*
   - Melakukan proses perhitungan menggunakan metode *Naive Bayes*
   - Membandingkan hasil dari proses *Naive Bayes* dengan data asli
   - Melakukan proses perhitungan menggunakan metode *K-Nearest Neighbor*
   - Membandingkan hasil dari proses *K-Nearest Neighbor* dengan data asli
   - Melakukan proses perulangan untuk melakukan proses perhitungan lagi
   - Menghitung akurasi yang didapatkan dari kedua metode.
2. Pada pengujian berdasarkan nilai atribut *K* didapatkan rata-rata nilai akurasi tertinggi pada *K* bernilai 3 yaitu 90.73%. Pada pengujian berdasarkan *percentage split* nilai rata-rata keseluruhan metode *K-Nearest Neighbor* adalah 82.45% dan metode *Naive Bayes* sebesar 72.52%. Sedangkan pada pengujian berdasarkan jumlah data *training* metode *K-Nearest Neighbor* memiliki rata-rata akurasi keseluruhan sebesar 83.32% dan metode *Naive Bayes* sebesar 70.91%. Berdasarkan hasil akurasi yang didapatkan dari seluruh pengujian metode *K-Nearest Neighbor* merupakan metode yang terbaik dengan total rata-rata akurasi 82.89%.
3. Pengaruh nilai *K* terhadap akurasi pada algoritme *K-Nearest Neighbor* yaitu apabila nilai *K* semakin banyak maka akurasi yang didapat semakin kecil, hal tersebut dikarenakan oleh semakin banyak tetangga atau nilai *K* yang diambil semakin banyak data ikut dijadikan pertimbangan keputusan kelas.

   Rasio perbandingan jumlah data *training* dan data *testing* juga memiliki pengaruh terhadap akurasi algortime semakin besar selisih persentase atau rasio antara data *training* dan data *testing* maka semakin tinggi pula akurasi yang didapatkan. Hal tersebut dikarenakan jumlah data *training* yang semakin lebih banyak daripada jumlah data *testing* yang semakin lebih sedikit maka model *classifier* yang dibangun berdasarkan fakta dari data *training* akan lebih baik dan lebih lengkap untuk melakukan prediksi terhadap klasifikasi data baru atau data - *testing*. Sehingga akurasi yang didapatkan akan jauh lebih baik juga.

   Perbedaan jumlah data *training* juga memiliki pengaruh terhadap akurasi algoritme. Sama halnya dengan pengujian berdasarkan rasio perbandingan atau *percentage split*, semakin banyak jumlah data *training* maka semakin tinggi pula hasil akurasi yang didapatkan. Hal tersebut dikarenakan jika semakin banyak data *training* yang digunakan maka semakin baik dan semakin lengkap juga model *classifier* yang dibentuk berdasarkan fakta dari data *training* tersebut. Sehingga pada saat melakukan prediksi pada klasifikasi data - *testing* atau data baru, akurasi yang didapatkan akan semakin baik atau tinggi.

7. DAFTAR PUSTAKA

Bustami. 2013. "*Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*". TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146

Donna Prayoga, Novianto. 2018. "*Sistem Diagnosis Penyakit Hati Menggunakan Metode Naive Bayes".* Universitas Brawijaya. Malang.

Hamidi, Rifwan,dkk. 2017. "*Implementasi Learning Vector Quantization (LVQ) untuk Klasifikasi Kualitas Air Sungai*". Universitas Brawijaya, Malang.

Hastuti, Khafiizh. 2012. "*Analisis Komparasi Algoritme Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif*". Universitas Dian Nuswantoro, Semarang.

Iskandar, Derick dan Suprapto, Yoyon K. 2013. "*Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan antara Algoritme*

C4.5 dan Naive Bayes Classifier". Institut Teknologi Sepuluh November, Surabaya.

J. Kodoatie, Robert & Roestam Sjarief. 2010. "*Tata Ruang Air*". Yogyakarta.

Kusnawi. 2007. "*Pengantar Solusi Data Mining*". STMIK AMIKOM, Yogyakarta.

Kustiyahningsih Yeni, dkk. 2013. "*Sistem Pendukung Keputusan untuk Menentukan Jurusan pada Siswa SMA Menggunakan Metode KNN dan SMART*". Universitas Trunojoyo, Madura.

Kusumadewi, Sri. 2009. "*Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification*". Universitas Islam Indonesia, Yogyakarta.

Lestiana, Mila. 2015. "*Perbandingan Algoritma Decision Tree (C4.5) dan Naive Bayes pada Data Mining untuk Identifikasi Tumbuh Kembang Anak Balita*". Universitas Muhammadiyah, Surakarta.

Peraturan Menteri Republik Indonesia nomor 492 tahun 2010 tentang Persyaratan Kualitas Air minum. Jakarta : Kementrian Kesehatan Republik Indonesia.

Peraturan Pemerintah Republik Indonesia nomor 82 tahun 2001 tentang Pengelolaan Kualitas Air dan Pengendalian Pencemaran Air. Jakarta.

Santoso. 2016. "*Perbandingan Metode K-Nearest Neighbor(K-NN) dan Learning Vector Quantization (LVQ) untuk Permasalahan Klasifikasi Tingkat Kemiskinan".* Institut Teknologi Sepuluh November, Surabaya.

Situmorang, Manihar. 2017. "*Kimia Lingkungan*". Depok: PT RajaGrafindo

Tri Vulandari, Retno. 2017. "*Data Mining Teori dan Aplikasi Rapidminer*". Surakarta.