

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Machine learning (ML) adalah alat canggih dan fleksibel yang mengembangkan program komputer yang belajar dan berkembang dari pengalaman tanpa diprogram secara eksplisit untuk melakukan tugas tertentu [1]. ML memungkinkan komputer belajar dari data dan mengeksplorasi pola dan hubungan tersembunyi untuk memprediksi data baru [2]. ML menggunakan algoritma dan informasi *historis* untuk membantu proses prediksi dan klasifikasi [3]. Kemampuan ML dalam mempelajari sebuah data untuk mendapatkan sebuah pola atau informasi yang berharga sangat penting dalam banyak bidang di kehidupan kita, dengan data tersebut, komputer dapat membuat sebuah otomatisasi keputusan dan klasifikasi.

Klasifikasi merupakan teknik pembelajaran data untuk menghasilkan prediksi nilai dari serangkaian atribut [4]. Mengklasifikasikan dan mengekstrak informasi yang berguna berdasarkan *class* merupakan tugas klasifikasi, sehingga pengklasifikasian data dapat dijalankan dengan cepat dan efektif [5]. Klasifikasi banyak sekali penerapannya pada kehidupan sehari-hari seperti Diagnostik medis, Pengenalan wajah, ramalan cuaca dll [6], sehingga penerapan klasifikasi menjadi hal penting dan sangat bermanfaat dalam kehidupan kita.

Pada saat ini penggunaan algoritma untuk pemodelan data dan pembelajaran mesin banyak sekali dipakai untuk memecahkan masalah, ada beberapa algoritma

pembelajaran mesin dalam klasifikasi seperti *Naïve bayes*, *Support vector machine* *Decision tree* dan *Logistic regression* [7]. *K-Nearest Neighbor* (KNN) merupakan salah satu algoritma klasifikasi data mining yang baik dalam mengolah data berbentuk numerik dan banyak digunakan dalam penelitian [8], Algoritma *K-Nearest Neighbor* (KNN) menggunakan konsep mencari tetangga terdekat dan mengelompokkannya berdasarkan kemiripannya [9]. Standar kualitas air dapat dianalisis berdasarkan pengukuran konsentrasi kandungan unsur yang tercantum di dalam baku mutu kualitas air [10]. Berdasarkan penjelasan diatas, algoritma *K-Nearest neighbor* (KNN) peneliti gunakan dalam klasifikasi kualitas air karna efektif dalam menangani data yang bertipe numerik dan konsep yang sederhana, dengan melihat jumlah tetangga yang terdekat sehingga data air dapat diklasifikasikan dengan melihat data yang memiliki *value* yang serupa.

Air merupakan aspek penting dalam kehidupan manusia dan ekosistem, Sehingga air sangat dibutuhkan untuk kelangsungan hidup dan keseimbangan lingkungan [11]. Kualitas air merupakan salah satu faktor yang sangat berpengaruh terhadap kehidupan berbagai macam organisme di perairan, karena kualitas air tersebut dapat mempengaruhi laju aktivitas metabolisme, pertumbuhan, dan perkembangan organisme yang hidup di air [10]. Klasifikasi kualitas air untuk menentukan air yang aman dan tidak merupakan suatu aspek penting untuk pengelolaan sumber daya air, Hasil dari klasifikasi ini dapat membantu dalam pengambil keputusan terkait pengolahan air, perlindungan lingkungan dan manajemen resiko [11]. Dengan diketahuinya kualitas air, maka banyak aspek

kehidupan yang dapat terjaga dari kandungan air yang berbahaya, misalnya air yang layak diminum manusia dan kualitas air yang baik bagi ekosistem.

Dalam melakukan klasifikasi kualitas air dibutuhkan sebuah data, Adapun data yang digunakan pada penelitian ini adalah sebuah dataset yang bersumber dari *repository* Kaggle dengan nama *water quality*. Dataset ini berisikan 7999 jumlah data dengan fitur sebanyak 20 dan 1 label yang berisikan 2 *class*, Tipe data fitur dalam dataset ini berupa *float* dan pada label bertipe data *binary*. *class* dalam dataset ini 0 adalah tidak aman dan 1 adalah aman. Dataset ini sudah memiliki label disetiap data nya, sehingga bisa dikatakan data ini merupakan data untuk *supervised learning*, dimana pada data ini algoritma *machine learning* akan belajar dari data pelatihan yang sudah memiliki label yang sudah diketahui. Output dari sistem ini bisa berupa prediksi *class* berdasarkan data yang sudah dipelajari [12], [13]. Dari dataset ini, machine learning akan belajar, mencari pola serta melihat korelasi antar fitur sehingga dapat memprediksi sebuah *class* berdasarkan algoritma yang digunakan.

Penelitian sebelumnya yang terkait klasifikasi kualitas air menunjukkan bahwa penggunaan algoritma *K-NN* memberikan akurasi tertinggi sebesar 85,24% [14]. Penelitian sejenis lainnya oleh Merlinda Wibowo [15] dengan objek penelitian klasifikasi deteksi stress memberikan hasil akurasi *K-NN* sebesar 59%. Berdasarkan penelitian terkait, penelitian ini melakukan klasifikasi kualitas air yang aman serta melakukan perbandingan akurasi algoritma *K-NN* dengan penelitian sebelumnya. selain itu perbedaan penelitian ini dengan penelitian sebelumnya

terdapat pada Teknik validasi serta jumlah data uji dan nilai K, yang mana hal ini dapat berpengaruh terhadap akurasi yang dihasilkan.

Dari permasalahan dan penelitian sebelumnya yang telah dijelaskan, penulis tertarik untuk melakukan penelitian ini karena terdapat potensi untuk meningkatkan dan pengembangan lebih lanjut dengan Teknik-teknik yang berbeda dalam mengatasi klasifikasi kualitas air yang aman dengan mengangkat judul penelitian “**Penerapan Algoritma KNN Dalam Klasifikasi Kualitas Air Yang Aman**”.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang yang dikemukakan di atas, maka dapat dirumuskan permasalahan yang akan diteliti yaitu:

1. Bagaimana penerapan algoritma KNN pada dataset *water quality* untuk klasifikasi dan memprediksi kondisi kualitas air yang aman dan tidak aman?
2. Bagaimana akurasi model yang dihasilkan algoritma *K-Nearest Neighbor* (KNN) dalam mengklasifikasikan kualitas air?

1.3 BATASAN MASALAH

Adapun Batasan masalah dalam penelitian ini adalah:

1. Data penelitian kualitas air diambil di *repository* Kaggle dengan nama *water quality*, memiliki jumlah data sebanyak 7999 serta memiliki 20 atribut dan 1 label yang berisi 2 *class*.

2. Algoritma klasifikasi yang dipakai pada penelitian ini yaitu *K-Nearest Neighbor* (KNN)
3. Metode untuk mengukur kinerja model menggunakan *confusion matrix*
4. Teknik pembagian data latih dan uji menggunakan 10 k-fold *cross validation*
5. Algoritma menghitung jarak vector *K-Nearest Neighbor* menggunakan *euclidean distance*

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengetahui penerapan algoritma KNN di dataset *water quality* untuk klasifikasi kualitas air.
2. Mengukur dan mengevaluasi akurasi model yang dihasilkan oleh algoritma *K-Nearest Neighbor* (KNN) dalam mengklasifikasikan kualitas air.

1.4.2 Manfaat Penelitian

Adapun manfaat penelitian ini adalah:

1. Memberikan kontribusi bagi pengembangan metode klasifikasi kualitas air.
2. Menginformasikan pengambilan keputusan terkait dengan pengelolaan dan kualitas sumber daya air.

3. Membantu dalam mengevaluasi efektivitas dan peningkatan akurasi algoritma KNN dalam klasifikasi kualitas air.

1.5 SISTEMATIKA PENELITIAN

Pada sistematika penulisan ini, penulis dapat menjelaskan secara umum apa saja yang akan dibahas dalam setiap bab pada penelitian ini. Dimana penelitian ini terdiri dari lima bab yang meliputi:

BAB I : PENDAHULUAN

Pada bab ini, penulis menjelaskan secara umum mengenai latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan.

BAB II : LANDASAN TEORI

Pada bab ini, penulis mengutip sebuah gagasan dari pakar-pakar yang berhubungan dengan topik permasalahan yang berisikan definisi-definisi yang melandasi penelitian.

BAB III : METODOLOGI PENELITIAN

Pada bab ini, menjelaskan mengenai kerangka kerja penelitian, proses machine learning serta alat-alat pembantu dalam proses penelitian.

BAB IV : HASIL DAN ANALISIS

Pada bab ini, berisi tentang hasil data analisis penelitian serta sebuah perhitungan mengenai objek yang diteliti.

BAB V : PENUTUP

Pada bab ini merupakan bab penutup dari seluruh penelitian yang berisikan kesimpulan dari seluruh bab yang ada dan saran yang bersangkutan mengenai penelitian ini.