BAB I
PENDAHULUAN

Bab ini akan berisikan latar belakang masalah yang menjadikan penulis mengerjakan penelitian ini, identifikasi masalah, tujuan, manfaat, kebaruan serta penjabaran dari kerangka penulisan.

1.1 Latar Belakang Masalah

Berdasarkan Hasil Studi Kualitas Air Minum Rumah Tangga (SKAMRT) menyatakan bahwa air isi ulang dikonsumsi oleh 31 persen rumah tangga di Indonesia, selanjutnya menggunakan dari sumur gali terlindungi untuk dikonsumsi sebanyak 15.9 persen, serta konsumsi air yang berasal dari air sumur bor atau pompa terdapat sebanyak 14.1 persen (katadata.co.id, 2021). Sumber air yang dikonsumsi oleh masyarakat Indonesia sangat beragam. Meskipun Indonesia memiliki berbagai jenis sumber air dan jumlah air yang melimpah di Indonesia, tetapi tidak semua air tersebut aman serta layak untuk dikonsumi.

SKAMRT dilaksanakan pada tahun 2020 di 34 provinsi, terdiri dari 493 kabupaten dan kota yang mewakili nasional. Dengan menggunakan kerangka sampel Susenas Kor Maret 2020, terdapat sebanyak 25.000 rumah tangga dari 2.500 Blok sensus yang ditargetkan menjadi sampel dalam penelitian ini (Puslitbang Upaya Kesehatan Masyarakat, 2020). Sampel air minum diuji menggunakan alat Sanitarian Kit atau Kesling Kit milik masing-masing dinas kesehatan atau puskesmas dari lokasi terpilih. Cara tradisional mengetahui kualitas air tersebut membutuhkan proses yang panjang dan juga menghabis biaya.

Sulit untuk membedakan apakah sumber air yang digunakan sudah tercemar atau masih layak digunakan. Selain itu, untuk mengetahui tercemar atau tidaknya air tersebut juga membutuhkan waktu dan juga biaya. Dinas Lingkungan Provinsi DKI Jakarta menyediakan dataset yang diterbitkan setahun sekali dan terbagi menjadi dua periode. Dataset yang dipublikasikan pada *website* data.jakarta.go.id ini berisikan data hasil analisa dari pengambilan contoh air sumur di Provinsi Jakarta yang mencantumkan parameter-parameter dan hasil dari

pengujian. Untuk mengolah banyaknya data yang ada dari tiap periode analisa, Dinas Lingkungan Hidup Provinsi Jakarta memerlukan sistem yang dapat melakukan pengolahan sehingga dapat menghasilkan informasi dari kumpulan data tersebut dengan cepat dan akurat.

Machine learning adalah sebuah metode yang diterapkan pada komputer agar dapat mempelajari serta memperoleh pengetahuan secara langsung. Dengan begitu komputer dapat memecahkan masalah yang diberikan menggunakan pengetahuan yang telah mereka miliki (Purwati, et al, 2021). Dalam prosesnya, *machine learning* akan menerapkan *data mining* sebagai pendukung mesin dalam mempelajari serta melatih data sehingga dapat menghasilkan informasi.

Berkembangnya *data mining* maka proses analisa untuk melakukan klasifikasi dari data yang tersedia dapat dilakukan dengan lebih cepat. *Data mining* menganalisis pola serta hubungan keterkaitkan tertentu yang ada pada data berukuran besar. Hasil analisis ini diekstraksi dan menghasilkan informasi yang berharga (Siregar & Puspabhuana, 2017). Manusia memiliki keterbatasan waktu dan tenaga, sehingga ketika harus mengelola data yang banyak cenderung menghabiskan banyak waktu. Penerapan *data mining* pada *machine learning* memungkinkan dataset dalam jumlah besar dapat diproses dalam waktu singkat dan akurat.

Dalam penelitian ini digunakan dua algoritma sebagai perbandingan yaitu *random forest* dan *naive bayes*. Pemilihan algoritma *random forest* didasarkan pada performa yang dihasilkan ketika melakukan klasifikasi terhadap kasus dengan atribut yang kompleks. Salah satunya yakni pada kasus kemungkinan diabetes tahun awal, algoritma *random forest* bekerja dengan baik dengan nilai akurasi mencapai angka 97.88% (Apriliah, et al, 2021).

Algoritma naive bayes menggunakan prinsip probabilitas bersyarat dalam proses klasifikasi sehingga mudah untuk diterapkan. Salah satu contoh penerapan algoritma naive bayes adalah pada kasus analisis sentimen dan klasifikasi teks. Dalam klasifikasi teks tersebut didapatkan nilai akurasi sebesar 98.67% dengan

presisi sebesar 93.81% serta nilai *recall* sebesar 96.67% (Deolika, Kusrini, & Luthfi, 2019).

1.2 Identifikasi Masalah

Identifikasi masalah dalam penelitian ini dibagi menjadi dua bagian yaitu rumusan masalah dan batasan penelitian. Rumusan masalah berisikan rumusan dari topik masalah yang diangkat, sedangkan batasan masalah memberikan batasan terhadap penelitian yang dilakukan agar tetap berada pada inti rumusan masalah.

1.2.1. Rumusan Masalah

Berikut beberapa rumusan masalah yang disusun oleh penulis dijabarkan sebagai berikut:

1. Bagaimana membangun model *machine learning* dengan menerapkan algoritma *random forest* dan *naïve bayes*?
2. Bagaimana perbandingan performa *machine learning* yang dibangun dengan algoritma yang berbeda dalam melakukan klasifikasi kualitas air (studi kasus *random forest* dan *naïve bayes*)?

1.2.2 Batasan Masalah

Penelitian ini menetapkan beberapa batasan agar diperoleh hasil yang maksimal. Adapun batasan masalah yang dimaksud adalah sebagai berikut:

1. Penelitian akan berfokus pada pengujian kedua algoritma yakni *random forest dan naive bayes*.
2. Data yang digunakan dalam proses klasifikasi adalah dataset air sumur DKI Jakarta pada tahun 2018 – 2019.
3. Informasi yang akan disajikan adalah hasil kualitas air di DKI Jakarta dalam bentuk klasifikasi data berdasarkan nilai akurasi, presisi, recall, dan F1-score.

4. Penelitian ini dilakukan hanya untuk mengetahui algoritma yang lebih efektif dalam melakukan klasifikasi kualitas air. Pengukuran efektivitas dilakukan menggunakan metode *confusion matrix*.

1.3 Tujuan Penelitian

Penelitian ini memiliki beberapa tujuan yang ingin dicapai sebagai berikut:

1. Memperoleh pengetahuan terkait pemodelan machine learning untuk klasifikasi data kualitas air menggunakan algoritma *random forest* dan *naïve bayes*.
2. Mengetahui algoritma yang lebih efektif dalam klasifikasi data kualitas air DKI Jakarta dengan mempertimbangkan hasil akurasi, presisi, recall, dan F1-score dari masing-masing algoritma.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat menghasilkan manfaat bagi masyarakat, peneliti, dan ilmu pengetahuan.

1.4.1 Manfaat bagi Masyarakat

Dengan mempublikasikan penelitian ini diharapkan dapat menghasilkan manfaat literasi bagi masyarakat yang ingin mengetahui kualitas air menggunakan *machine learning*. Diharapkan dengan adanya klasifikasi kualitas air menggunakan ini dapat mempermudah masyarakat dalam mengetahui apakah air yang mereka gunakan layak atau tidak. Selanjutnya penelitian ini diharapkan dapat meningkatkan minat dan keterampilan di bidang Informatika, khususnya *Data Mining* yang merupakan bagian dari kecerdasan buatan.

1.4.2 Manfaat bagi Peneliti

Diharapkan pada penelitian ini dapat menjadi sumber literasi untuk meningkatkan pengetahuan peneliti. Sehingga perkembangan klasifikasi kualitas air dapat menjadi semakin baik.

1.4.3 Manfaat bagi Ilmu Pengetahuan

Penelitian diharapkan dapat memberikan pandangan mengenai perbandingan antara dua algoritma yaitu *random forest* dan *naive bayes*. Hasil dari penelitian ini dapat memberikan kesimpulan dari kedua algoritma tersebut. Sehingga dapat dimanfaatkan sebagai sumber literasi yang kedepannya bisa dikembangkan oleh para peneliti lain di masa depan.

1.5 Kebaruan

Kebaruan penelitian ini ada pada algoritma yang digunakan dalam melakukan klasifikasi kualitas air. Pada penelitian sebelumnya Algoritma yang dilakukan oleh Rifwan Hamidi pada tahun 2017 adalah metode Learning Vector Quantization (LVQ). Pada penelitian tersebut dihasilkan nilai akurasi dengan rata-rata akhir sebesar 81.13% (Hamidi, Furqon, & Rahayudi, 2017). Sedangkan pada penelitian kali ini digunakan algoritma *random forest dan naive bayes* dalam proses klasifikasinya serta kebaruan data yang berasal dari dataset Jakarta tahun 2018-2019.

1.6 Kerangka Penulisan

Pada penelitian ini akan dibagi menjadi beberapa bab sehingga dapat mempermudah pencarian informasi yang dibutuhkan, juga memberikan penyelesaian yang sistematis. Pembagian bab yang dilakukan adalah sebagai berikut:

BAB I. Pendahuluan, bab ini berisikan latar belakang, identifikasi masalah, rumusan masalah, batasan, tujuan, manfaat, kebaruan, kerangka penulisan.

BAB II. Tinjauan Pustaka, memuat beberapa subbab diantaranya pencapaian terdahulu, dan Tinjauan teoritis yang akan memuat berbagai teori yang menjadi landasan pengetahuan bagi penulis untuk menyusun laporan dan melakukan penelitian untuk

klasifikasi kualitas air menggunakan dua algoritma yaitu *randomf* dan *naive bayes*.

BAB III. Metode Penelitian, pada bab ini diuraikan paradigma penelitian dan penelitian yang digunakan untuk meyelesaikan permasalahan yang diangkat. Paradigma penelitian digambarkan dalam sebuah diagram tulang ikan.

BAB IV. Perencanaan, bab ini berisikan uraian mengenai langkah-langkah penelitian yang meliputi pengumpulan data, pembuatan model *machine learning* dan pengujian.

BAB V. Hasil dan Pembahasan, bab berisikan pemaparan hasil pengujian beserta analisis dari penelitian yang telah dilakukan.

BAB VI. Penutup, bab ini memberikan kesimpulan dari penelitian yang telah dilakukan dan saran kelanjutan penelitian kedepannya.