IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI KUALITAS AIR YANG DAPAT DI KONSUMSI

Adrian Wisnu Saputra [1],Ade Irma Purnamasari [2], Irfan Ali [3]
[1] Program Studi Teknik Informatika, STMIK IKMI Cirebon
[2] Program Studi Teknik Informatika, STMIK IKMI Cirebon
[3] Program Studi Rekayasa Perangkat Lunak, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya Cirebon, Indonesia
adrianwisnusaputra@gmail.com

ABSTRAK

Air adalah sumber kehidupan yang penting bagi makhluk hidup termasuk manusia. Metode pemantauan dan evaluasi yang efektif dan tepat diperlukan untuk menjaga kualitas air yang aman. Sifat air yang dapat dikonsumsi sangat penting untuk kesehatan secara umum. Kontaminasi air dapat memicu berbagai infeksi, seperti mencret, kolera, dan hepatitis. Oleh karena itu, sangat penting untuk mengembangkan strategi yang dapat secara tepat meramalkan sifat air yang dapat dikonsumsi. Eksplorasi ini menerapkan perhitungan penambangan informasi Bayes yang mudah tertipu untuk mengantisipasi sifat air yang dapat dimakan. Perhitungan ini bekerja dengan menghitung kemungkinan informasi masuk ke dalam kelas tertentu dengan mempertimbangkan probabilitas kreditnya. Penelitian ini diharapkan dapat menerapkan perhitungan Naive Bayes dalam mengantisipasi sifat air yang layak dikonsumsi. Kualitas air yang baik sangat penting bagi kesehatan manusia, dan prediksi yang akurat dapat membantu orang memilih jumlah air yang tepat untuk diminum. Perhitungan Naive Bayes dipilih karena kemampuannya untuk menangani pesanan dengan variasi batas yang kompleks dalam dataset kualitas air. Informasi yang digunakan dalam penelitian ini adalah informasi kualitas air yang didapat dari situs dataset Kaggle dengan jumlah 3477 catatan informasi dengan faktor yang meliputi berbagai faktor fisik, substansi, dan alam. Hasil analisis menunjukkan bahwa model klasifikasi Naive Bayes mampu memprediksi kualitas air dengan akurasi sebesar 65,08 persen, presisi 62,08 persen, dan recall 26,47 persen. Hal ini menunjukkan bahwa algoritma Naive Bayes dapat memprediksi kualitas air minum secara akurat.

***Kata kunci**: kualitas air, data mining, naive bayes, akurasi*

1. PENDAHULUAN

Di masa komputerisasi saat ini, peningkatan pesat di bidang Informatika secara keseluruhan mempengaruhi berbagai bagian kehidupan, seperti inovasi, bisnis, pelatihan, dan lainnya. Kemampuan untuk memproses dan mengevaluasi data secara akurat dan efisien adalah salah satu manfaatnya. Hal ini memungkinkan kita untuk menyelidiki informasi dengan lebih baik dan menentukan pilihan yang lebih baik. Dalam situasi khusus ini, penelitian tentang pemanfaatan perhitungan Naive Bayes untuk menentukan sifat air yang dapat dikonsumsi menjadi sangat signifikan dan penting untuk diselesaikan.

Kehidupan manusia bergantung pada air untuk segala hal. Namun, berbagai masalah lingkungan dan kesehatan dapat terjadi akibat kualitas air yang buruk. Kebutuhan akan air bersih di Indonesia terus meningkat, namun aksesibilitasnya masih terbatas. Hal ini disebabkan karena pembangunan yang tidak memperhatikan keseimbangan wilayah dan mengurangi daerah resapan, terutama di daerah perkotaan. Akibatnya, tidak banyak sumber air bersih yang tersedia. Kontaminasi air juga merupakan masalah yang signifikan di Indonesia. Limbah rumah tangga dan rumah tangga adalah kontributor utama pencemaran air. Pencemaran air menyebabkan penurunan kualitas air dan aksesibilitas air bersih di berbagai daerah. Kualitas air yang buruk dapat menyebabkan berbagai macam infeksi, salah satunya adalah mencret. Mencret adalah penyakit yang menyebabkan kematian sekitar 800.000 orang secara keseluruhan setiap tahunnya. Mencret disebabkan oleh penggunaan air yang berantakan atau tercemar, kebersihan tangan yang kurang baik saat memoles air, dan masalah desinfeksi air minum.[1] Pertama-tama, perlu dilakukan identifikasi jenis air yang aman untuk dikonsumsi oleh masyarakat secara keseluruhan. Identifikasi ini dapat dilakukan dengan menggunakan prosedur penggalian informasi. Masalah dengan kualitas air dapat ditemukan, kategori label air dapat diprediksi, dan objek air dapat dibedakan satu sama lain dengan menggunakan metode data mining. Naive Bayes merupakan salah satu algoritma klasifikasi yang dapat digunakan dalam teknik data mining.

Studi-studi sebelumnya telah dilakukan dalam topik serupa. Penelitian pertama adalah studi yang dilakukan oleh Hardiana Said, Nurhafifah Matondang, Helena Nurramdhani Irmanda hasil dari penelitian ini menunjukkan bahwa algoritma K-Nearest Neighbor (KNN) efektif dalam memprediksi kualitas air. Berdasarkan evaluasi model menggunakan matriks kebingungan, ditemukan bahwa algoritma KNN memiliki akurasi sebesar 85,24% dalam memprediksi apakah air aman untuk dikonsumsi. Penelitian ini menyimpulkan bahwa algoritma KNN dapat digunakan untuk menentukan apakah air aman untuk dikonsumsi berdasarkan data lingkungan sekitar [1].

Penelitian kedua ditulis oleh Fauzi Yusa Rahman, Indu Indah Purnomo, Nadya Hijriana. Alasan dari eksplorasi ini adalah untuk menerapkan strategi information mining untuk mengkarakterisasi kualitas air dengan mempertimbangkan batas-batas kualitas air yang meliputi ilmu mikroba, ilmu anorganik, dan batas-batas senyawa. Beberapa perhitungan dicoba untuk dieksekusi, misalnya perhitungan pohon pilihan, innocent bayes dan k-Nearest neighbor. Teknik pengujian yang digunakan adalah k-crease cross approval dan ROC bend. Perbandingan kinerja dari ketiga algoritma yang dihasilkan oleh teknik data mining pada dataset kualitas air bertujuan untuk mengidentifikasi metode yang paling efektif untuk mengklasifikasikan kualitas air. Berdasarkan pemeriksaan perhitungan terhadap uji coba dari semua teknik yang digunakan, ditemukan bahwa perhitungan pohon pilihan merupakan perhitungan yang paling dapat diandalkan dalam mengurutkan informasi kualitas air dengan ketepatan sebesar 94,94% dan nilai AUC sebesar 0,865 sehingga termasuk ke dalam kelas pengelompokan yang baik. Dengan demikian, masih ada peluang untuk melakukan eksplorasi lebih lanjut pada poin ini.[3].

Tujuan utama dari eksplorasi ini adalah untuk mengembangkan teknik yang mahir dan tepat untuk menentukan sifat air yang dapat dikonsumsi dengan menggunakan perhitungan Innocent Bayes. Pemeriksaan ini juga diharapkan dapat menguji ketepatan dan keefektifan teknik yang dibuat dengan membandingkannya dengan strategi manual saat ini. Makna dari eksplorasi ini terletak pada peningkatan teknik yang lebih mahir dan tepat untuk menentukan sifat air yang dapat dikonsumsi. Strategi yang dibuat dapat membantu para spesialis di bidang ekologi dan kesejahteraan untuk mengantisipasi sifat air yang dapat dikonsumsi dengan lebih mahir dan tepat. Demikian juga, eksplorasi ini juga dapat menambah bidang Informatika secara keseluruhan dengan menciptakan teknik yang dapat digunakan di bidang alam dan kesejahteraan.

Penelitian ini akan menggunakan teknik kuantitatif dengan metodologi eksploratif. Dalam penelitian ini, variabel-variabel diukur secara numerik dan obyektif dengan menggunakan metode kuantitatif. Sedangkan metodologi eksploratif digunakan untuk menguji spekulasi dan memutuskan keadaan dan hasil yang logis dari hubungan antara faktor-faktor yang diteliti. Algoritma Naive Bayes akan digunakan dalam penelitian ini untuk mengklasifikasikan data kualitas air dan memastikan kualitas air minum. Karena dapat memproses data dengan cepat dan akurat, algoritma Naive Bayes dipilih. Selain itu, eksplorasi ini juga akan menggunakan metode pemeriksaan informasi faktual untuk menguraikan hasil pemeriksaan.

Temuan dari penelitian ini mungkin memiliki potensi untuk meningkatkan pemahaman kita tentang bagaimana algoritma Naive Bayes diterapkan untuk menentukan kualitas air minum yang aman. Konsekuensi dari penelitian ini dapat dilibatkan oleh para ahli di bidang alam dan kesehatan untuk menentukan kualitas air yang dapat dikonsumsi dengan lebih baik dan tepat. Selain itu, temuan dari penelitian ini dapat menjadi dasar untuk penelitian lebih lanjut mengenai penerapan algoritma Naive Bayes di bidang kesehatan dan lingkungan. Karena dapat membantu pengembangan sistem yang lebih efektif dan tepat dalam menentukan kualitas air yang dapat dikonsumsi, maka implikasi dari temuan penelitian ini juga dapat berdampak pada kemajuan teknologi di bidang kesehatan dan lingkungan.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian ini akan menggunakan teknik kuantitatif dengan metodologi eksploratif. Dalam penelitian ini, variabel-variabel diukur secara numerik dan obyektif dengan menggunakan metode kuantitatif. Sedangkan metodologi eksploratif digunakan untuk menguji spekulasi dan memutuskan keadaan dan hasil yang logis dari hubungan antara faktor-faktor yang diteliti. Algoritma Naive Bayes akan digunakan dalam penelitian ini untuk mengklasifikasikan data kualitas air dan memastikan kualitas air minum. Karena dapat memproses data dengan cepat dan akurat, algoritma Naive Bayes dipilih. Selain itu, eksplorasi ini juga akan menggunakan metode pemeriksaan informasi faktual untuk menguraikan hasil pemeriksaan.

Tinjauan pustaka adalah langkah kritis dan metodis dalam penelitian yang melibatkan pencarian, pemahaman, dan sintesis literatur yang relevan atau penelitian sebelumnya. Tujuan dari tinjauan pustaka adalah untuk mengidentifikasi kesenjangan penelitian dengan melakukan analisis dan sintesis dari penelitian sebelumnya pada subjek yang dihadapi. Teknik survei penulisan yang digunakan dalam eksplorasi ini adalah mengaudit artikel-artikel yang berhubungan dengan Eksekusi Information Mining dalam menentukan sifat air yang baik untuk dimanfaatkan dengan menggunakan Naive Bayes Algortima.

Fokus dari penelitian ini adalah pada analisis big data dan data mining. Tahap selanjutnya adalah mencari tulisan yang sesuai dengan pokok permasalahan. Langkah ini dapat memberikan garis besar harapan. Cara yang paling umum dalam mencari dan mengumpulkan artikel dengan menggunakan perangkat distribute or die, yang diambil dari basis informasi skolastik, khususnya Google Researcher, Shinta dengan semboyan "Information Mining", "Perhitungan Credulous Bayes" dan "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI KUALITAS AIR YANG DAPAT DI KONSUMSI". Evaluasi data, atau menyaring, memilih, dan menyortir artikel jurnal yang benar-benar relevan dengan topik penelitian dan mempertimbangkan kebaruannya, adalah langkah ketiga. Kemutakhiran dibatasi dengan memilih artikel dari 5 tahun terakhir, mulai dari tahun 2023, 2022, 2021, 2020, dan 2019. Sedangkan kepantasan dilihat dari kewajaran judul makalah, catatan Sinta, atau hal-

hal lain yang dianggap dapat dipercaya. Sepuluh artikel jurnal dipilih untuk ditinjau berdasarkan hasil evaluasi data.

Pada penelitian yang dilakukan oleh Sutisna, yang berjudul "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes" " Isu yang diangkat dalam catatan harian ini adalah pengaturan kualitas air dengan menggunakan teknik K-Nearst Neighbors (KNN), Naive Bayes, dan Decesion Tree. Strategi pemeriksaan yang digunakan adalah metode penggalian informasi, pengaturan, dan perhitungan Decesion Tree, KNN, dan Naive Bayes. Hasil penelitian menunjukkan bahwa teknik KNN memiliki tingkat ketepatan yang paling tinggi dalam karakterisasi kualitas air, yaitu 86,88%, diikuti oleh Decesion Tree dengan ketepatan 80,84% dan Naive Bayes dengan ketepatan 63,60%. Kesimpulan dari catatan harian ini adalah bahwa KNN dan Decesion Tree dipandang cukup memadai karena keduanya menghasilkan tingkat ketepatan di atas 80%, namun disarankan untuk melakukan penelitian lebih lanjut dengan menggunakan lebih banyak teknik dan informasi untuk mendapatkan tingkat ketepatan yang lebih tinggi.[4].

Pada penelitian yang dilakukan oleh Aldi Tangkelayuk yang berjudul "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree " dapat disimpulkan bahwa Hasil penelitian menunjukkan bahwa metode K-nearest neighbors mencapai tingkat akurasi tertinggi, yakni sebesar 86.88%, dalam mengklasifikasikan data kualitas air yang digunakan dalam penelitian ini. Sementara itu, Naïve-Bayes mencapai 63.60%, dan Decision tree mencapai 80.84%. Kesimpulan dari penelitian ini menyatakan bahwa metode Decision Tree memberikan tingkat keakuratan tertinggi, yakni 86.88%. Oleh karena itu, dapat disimpulkan bahwa metode klasifikasi Decision Tree dan KNN yang digunakan dalam penelitian ini cukup efektif, menghasilkan tingkat akurasi di atas 80%.[5].

Pada Penelitian yang dilakukan oleh Nurlindasari Tamsir yang berjudul "Algoritma Naïve Bayes untuk Prediksi Kualitas Sperma" Penerapan algoritma Naïve Bayes untuk memprediksi kualitas air yang dapat dikonsumsi telah berhasil diimplementasikan. Dari hasil penelitian dapat disimpulkan bahwa penerapan algoritma Naïve Bayes dengan Teknik Koreksi Laplace menghasilkan tingkat akurasi sebesar 92% dan tingkat kesalahan sebesar 8% dalam memprediksi kualitas sperma. Sistem yang dikembangkan menggunakan UML ini menghasilkan use case untuk pengguna dan administrator, serta 7 diagram kelas. Semua modul yang mengalami pengujian Black Box, yang meliputi validasi, pengujian batas, evaluasi kinerja, dan penilaian keamanan, menunjukkan hasil yang memuaskan.Berdasarkan wawasan yang diberikan oleh penelitian ini, disarankan bahwa pemanfaatan data mining dan kecerdasan buatan, yang dicontohkan oleh algoritma Naïve Bayes, dapat secara signifikan berkontribusi pada diagnosis dan prediksi masalah kesuburan. Selain itu, membuat aplikasi berbasis web untuk memprediksi kualitas sperma dapat menjadi alat yang berharga di bidang medis, membantu pasangan yang bergulat dengan masalah ketidaksuburan.[6]

Pada penelitian yang dilakukan oleh Baiq Nurul Azmi, yang berjudul "Analisis pengaruh PCA Pada Klasifikasi Kualitas Air menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression" Artikel ini menyelidiki pengaruh penggunaan Pemeriksaan Bagian Kepala (PCA) untuk penurunan aspek dalam proses pengaturan kualitas air, yang secara eksplisit menggunakan teknik Strategic Relapse dan K-Nearest Neighbor (kNN). Isu utama yang menjadi perhatian dalam penelitian ini adalah melemahnya tingkat kualitas air, yang dapat menyebabkan air menjadi tidak layak untuk digunakan dan menimbulkan bahaya. Oleh karena itu, kapasitas pengelompokan kualitas air yang tepat sangat penting untuk mencegah penurunan kualitas air. Prosedur pemeriksaan mengintegrasikan penggunaan Head Part Investigation (PCA) sebagai langkah prapemrosesan untuk mengurangi aspek sifat, bersama dengan Strategic Relapse dan K- Nearest Neighbor (kNN) untuk pengaturan. Penemuan ini menunjukkan bahwa strategi k- Nearest Neighbor (kNN) mengungguli kekambuhan yang dihitung, mencapai tingkat ketepatan yang luar biasa yaitu 90,8%, akurasi 90,0%, dan tinjauan 91,0%. Selain itu, hasil uji coba menunjukkan bahwa strategi k-Nearest Neighbor (kNN) mencapai ketepatan 82,42%, mengungguli ketepatan teknik Naive Bayes, yang berada di angka 70,32%.[7].

Pada penelitian yang dilakukan oleh Rita Manurung,Puji Sari Ramadhan,Moch Iswan Perangin-angin.yang berjudul "Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C.45 Dan Naive Bayes" dapat disimpulkan dari hasil: penelitian ini menggunakan Naive Bayes Clasifier (NBC) dan Algoritma C4.5, dua metode untuk klasifikasi data mining. Pengujian akan melibatkan 14 atribut. Hasil dari proses klasifikasi menunjukkan bahwa metode C4.5 memiliki tingkat akurasi 3% lebih tinggi daripada metode Naive Bayes.[8]

Pada Penelitian yang dilakukan oleh Wantoro yang berjudul "Sistem Pakar Diagnosis Penyakit Kutu Ikan Gurami (Argunus Indicus) Menggunakan Metode Naive Bayes" Ada kesimpulan bahwa fokus penelitian ini adalah menemukan penyakit pada ikan gurami berdasarkan gejalanya. Penelitian ini menggunakan metode Naive Bayes Classifier, yang merupakan metode klasifikasi probabilitas sederhana yang didasarkan pada teorema Bayes. Hasil penelitian ini mencakup evaluasi sistem melalui pengujian hasil keluaran dengan pakar, penentuan aturan dan pembobotan berdasarkan kasus yang ditemukan, dan perhitungan performa sistem menggunakan tabel confusion matrix untuk mendapatkan nilai ketepatan[9]

Pada Penelitian yang dilakukan oleh Baginda Harahap yang berjudul "Implementasi Algoritma

Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia Menggunakan Metode Naive Bayes" Dalam penelitian ini, masalah yang diidentifikasi adalah tingkat penyebaran COVID-19 di Indonesia. Tingkat penyebaran tersebut kemudian dianalisis untuk menentukan cara menyelesaikannya dan menentukan ruang lingkup masalah yang akan diteliti. Penelitian ini menggunakan analisis masalah dan studi literatur, pengumpulan data, dan penerapan algoritma Naive Bayes. Hasilnya menunjukkan bahwa metode Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output, dengan keuntungan bahwa metode lain.[10]

Pada penelitian yang dilakukan oleh Maulana Aditya Rahman , Nurul Hidayat , Ahmad Afif Supianto yang berjudul "Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)" Masalah yang diangkat dalam penelitian ini adalah pengaturan kualitas air bersih. Teknik yang digunakan untuk menangani masalah ini adalah dengan melakukan korelasi antara strategi K-Nearest Neighbor dan teknik Naive Bayes. Konsekuensi dari eksplorasi ini menunjukkan bahwa kedua teknik tersebut memiliki tingkat ketepatan yang sangat tinggi, namun korelasi antara keduanya dapat membantu dalam mengetahui strategi mana yang lebih baik dalam mengkarakterisasi sifat air bersih. Oleh karena itu, eksplorasi ini menambah pelacakan strategi yang sukses dan mahir dalam karakterisasi kualitas air bersih.[11]

Pada Penelitian ini ditulis oleh Abdi Rahim Damanik, Sumijan, dan Gunadi Widi Nurcahyo yang berjudul "Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes" Sejauh mana siswa merasa puas dengan pendidikan online selama epidemi COVID-19 adalah masalah yang ditemukan dalam penelitian ini. Metodologi penelitian ini menggunakan alat RapidMiner 5.3 dan memprediksi tingkat kepuasan siswa menggunakan algoritma klasifikasi Naïve Bayes. Prediksi tingkat kepuasan siswa dengan tingkat akurasi maksimum 96,6% adalah hasil dari penelitian ini.[12].

Pada Penelitian yang ditulis oleh Riyanto M yang berjudul "Penerapan Data Mining Dengan Metode Naïve Bayes Dan Learning Vector Quantization Credit Rating Dalam Memprediksi Kelayakan Pemberian Kredit Oleh PT. BPR Lebak Sejahtera" Masalah yang ingin diatasi dalam penelitian ini adalah penurunan kolektabilitas atau kredit macet, dengan kredit macet mulai dari 4,06% pada tahun 2019 dan diperkirakan akan mencapai 4,50% pada akhir tahun 2020. Pendekatan Learning Vector Quantization (LVQ) dan Naïve Bayes digunakan dalam penelitian ini. Berdasarkan hasil penelitian, tingkat akurasi ketika menggunakan pendekatan Naïve Bayes adalah 73,20%, namun hasil akurasi yang dihasilkan ketika menggunakan metode LVQ adalah 82,47%.Dalam tinjauan lain, strategi yang digunakan adalah teknik Credulous Bayes dengan tingkat presisi 70% hingga 92,5% menggunakan 250 indeks informasi dengan properti keamanan, pembayaran lengkap, uang muka yang berbeda, biaya absolut, jumlah kredit, status rumah, dan bermacam-macam ketukan. Selain itu, tinjauan lain menggunakan strategi K-Means Bunching untuk mengelompokkan pasien jangka pendek dan pasien rawat inap dari cakupan medis, dengan konsekuensi pengelompokan menggunakan algoritma Oneself Getting sorted out Guide (SOM). teknik yang digunakan adalah strategi Naive Bayes mengingat pilihan ke depan yang menggunakan informasi kredit di bank untuk meramalkan tingkat kesempurnaan cicilan nasabah, dengan efek memastikan ketepatan 71. Dari investigasi ini, cenderung disimpulkan bahwa teknik penambangan informasi seperti Naive Bayes, LVQ, dan K-Means Bunching digunakan untuk menangani masalah meramalkan risiko keandalan, pengelompokan pasien cakupan layanan kesehatan, dan mengantisipasi tingkat kelancaran angsuran klien. Hasil akhir dari penelitian ini menunjukkan adanya peningkatan tingkat ketepatan dalam peramalan dan pengelompokan informasi.[13].

2.2. Data Mining

Data mining adalah proses pengumpulan atau pengerukan informasi penting dari data besar (Big Data). Proses ini biasanya menggunakan metode matematika, statistika, bahkan *AI.Classification*, *clustering*, *association*, *regression*, *forecasting*, *sequence analysis*, dan *deviation analysis* adalah beberapa proses yang digunakan untuk memproses data mining.[13] Data mining dilakukan dengan menggunakan teknik dan algoritma tertentu untuk mengekstrak informasi bermanfaat dari kumpulan data yang dikumpulkan. Tujuan utama proses ini adalah untuk mendapatkan informasi yang tersembunyi di dalam data, yang dapat digunakan dalam berbagai bidang, seperti keuangan, bisnis, dan kesehatan.

2.3. Algoritma Naive Bayes

Algoritma Naive Bayes Clasifier merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes.[14]

2.4. Forecasting

Peramalan atau forecasting merupakan suatu proses yang bertujuan untuk memproyeksikan kebutuhan di masa depan, yang meliputi aspek kuantitas, kualitas, waktu, dan lokasi untuk memenuhi permintaan akan barang atau jasa.[15]Dalam dunia bisnis, forecasting digunakan untuk memprediksi permintaan konsumen, penjualan produk, atau

keuntungan perusahaan di masa depan. Dengan teknik forecasting yang tepat, kita dapat membuat keputusan yang lebih baik dalam mengelola suatu.

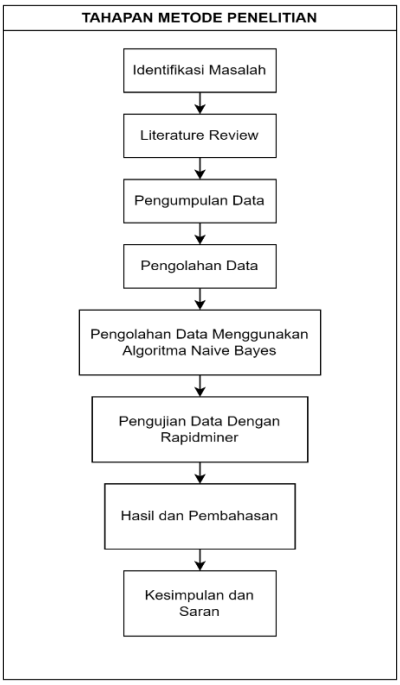
2.5. Rapidminer

RapidMiner adalah platform perangkat lunak ilmu data yang dikembangkan oleh perusahaan bernama sama dengan yang menyediakan lingkungan terintegrasi untuk persiapan data, pembelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis prediktif.

3. METODE PENELITIAN

3.1. Jenis Penelitian

Penelitian ini menerapkan metode penelitian eksperimen. Metode penelitian eksperimen adalah salah satu pendekatan yang digunakan dalam ilmu pengetahuan untuk menguji hipotesis dan mencari hubungan sebab-akibat antara variabel-variabel tertentu. Data yang diperoleh dari website kaggle dataset yang berjudul Water Potability dan akan diolah menggunakan algoritma Naive Bayes untuk memprediksi kualitas air yang baik untuk dikonsumsi dengan variabel pH Value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, dan potability sebagai variabel label. Dengan menggunakan metode penelitian eksperimen.



Gambar 1. Metode penelitian

3.1.1. Identifikasi masalah

Research Problem Mengidentifikasi masalah dengan memahami dataset terkait penelitian.

3.1.2. Literature review

Tinjauan Pustaka Mencari dan membaca literatur terkait dengan judul terkait penelitian.Menyusun ringkasan Literatur jurnal terbaru untuk memahami trend, temuan terbaru, dan isu-isu utama terkait penelitian.

3.1.3. Pengumpulan data

Research Problem Data yang digunakan dalam penelitian ini diperoleh dari webstite data open public dan Data kualitas air yang didapatkan sendiri bersumber dari website www.kaggle.com .Variabel yang digunakan pada penelitian ini untuk klasifikasi kualitas yaitu sejumlah 9 variabel dan 1 label.

3.1.4. Pengolahan data awal

(Analisis Data) Pada tahap ini akan dijelaskan mengenai data yang telah layak dilakukan proses pengolahan. Pengolahan data juga berfungsi untuk mempermudahpembentukan model.

3.1.5. Model yang diusulkan

Usulan Algoritma Data Mining Setelah melakukan pengolahan data awal, pada tahap ini akan dijelaskan mengenai model yang diusulkan yang akan digunakan terhadap data yang sudah ada.

3.2. Teknik Analisis Data

Klasifikasi dengan memanfaatkan algoritma Algoritma *Naïve Bayes*, sebagai algoritma klasifikasi yang umum dipergunakan dalam *Data Mining*. *Naive bayes* sebagai prosedur statistik untuk klasifikasi yang memberi peluang untuk menyampaikan ketidakpastian mengenai suatu model yang berprinsipkan pada menjelaskan hasil probabilitas. Prosedur ini berguna untuk menangani permasalahan diagnosis maupun prediksi. *Teorema Bayes* mempunyai bentuk seperti:

$$(H\,|X) = (X \mid H)\,(H)\,P(X) \quad (1)$$

Penjelasan:

X = Data dengan class yang belum diketahui.

H = Hipotesis Data X sebagai suatu class spesifik

P(H|X) = probabilitas hipotesis H berdasar pada kondisi x (posteriori prob.)

P(H) = Probabilitas hipotesis H (prior prob.)

P(X|H) = probabilitas X berdasar pada keadaan tersebut

P(X) = probabilitas dari X

Sumber : MENDELEY CITATION PLACEHOLDER 0

4. HASIL DAN PEMBAHASAN

4.1. Data Selection

Langkah pertama untuk mengelola data yaitu dengan menempatkan port Read excel untuk membaca dataset dalam format csv. Dataset dimasukan dengan klik dua kali pada port read csv, maka akan menampilkan lokasi data dalam folder pemilik lalu pilih dataset yang akan digunakan untuk diolah menggunakan tools read csv. Mengenali sifat dari text data yang dikumpulkan sebelumnya hingga dicoba proses labeling data. Attribute yang diberi label dalam penelitian ini merupakan potability, ialah attribute

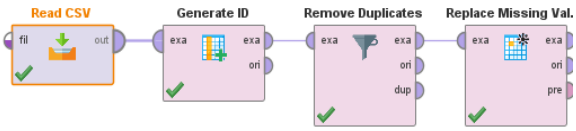
yang menampilkan apakah air aman untuk dikonsumsi manusia ataupun tidak. Air yang layak diminum Ya serta tidak bisa diminum Tidak. Proses labeling bisa dilakukan pengaturan warna pada label supaya memudahkan proses penelitian.

Gambar 2. Operator read csv

Operator Read CSV pada RapidMiner merupakan operator yang digunakan untuk membaca data dari file CSV (Comma-Separated Values). File CSV merupakan file bacaan yang menaruh informasi dalam format baris-kolom. Tiap baris dalam file CSV mewakili satu contoh informasi serta tiap kolom dalam baris mewakili satu atribut data.

4.2. Preprocessing

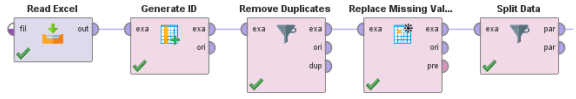
Langkah kedua adalah melakukan proses Generate ID untuk membuat ID dari masing-masing-masing atribut pada dataset dan juga melakukan Cleansing (penghapusan) data duplikat agar dataset menjadi balance untuk pemrosesan lebih lanjut dengan menggunakan port Remove Duplicates serta membuat pemodelan replace Missing Value untuk menghilangkan data yang kosong pada atribut pada dataset. Berikut adalah sub proses tersebut ditunjukan pada Gambar 3



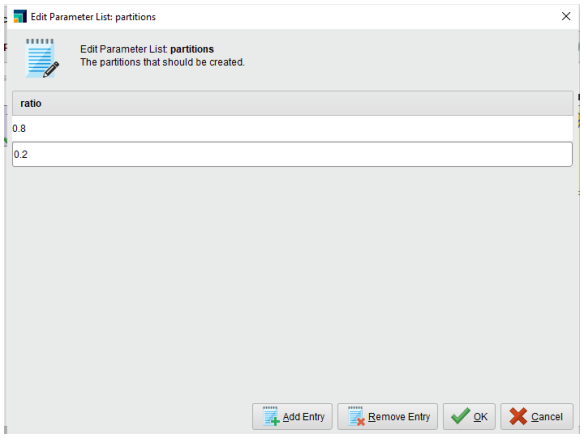
Gambar 3. Data Preprocessing

4.3. Transformation

Langkah ketiga adalah transformation menyeleksi data menggunakan port Split Data untuk membagi dataset menjadi data training dan data testing dengan parameter 0.8:0.2. Dimana 0.8 atau 80% untuk data training, sedangkan 0.2 atau 20% untuk data testing. Penelitian sebelumnya menggunakan operator split data dengan rasio 0.7:03, sehingga pada penelitian yang penulis lakukan untuk mengetahui hasil akurasi yang lebih besar. Berikut adalah sub proses tersebut:



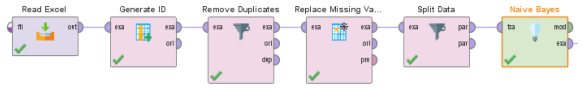
Gambar 4. Data Transformation



Gambar 5. Parameter Split Data

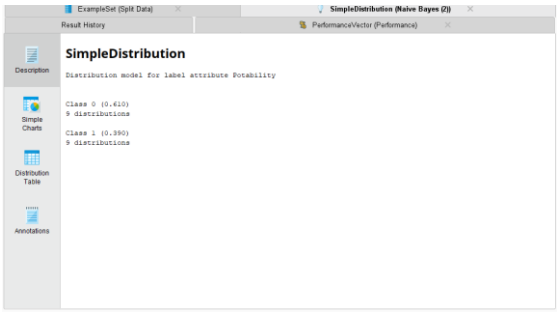
4.4. Data Mining

Langkah keempat adalah data mining menggunakan model port naïve bayes untuk klasifikasi. Kemudian menambahkan model split validation. Proses model tersebut adalah mengubungkan port naïve bayes, port apply model, dan performance. Berikut adalah sub proses tersebut:



Gambar 6. Permodelan Naive Bayes

Model Naive Bayes adalah model klasifikasi yang menggunakan teori probabilitas untuk memprediksi kelas dari suatu contoh data. Model Naive Bayes mengasumsikan bahwa setiap atribut dalam contoh data adalah independen dari atribut lainnya, diberikan kelas contoh data tersebut. Hasil dari penggunaan operator ini diperoleh informasi sebagai berikut:



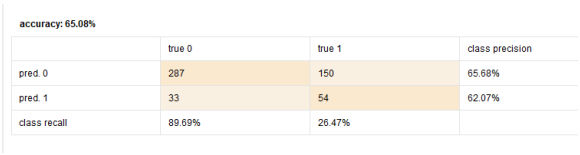
Gambar 7. Simple distribution

Berdasarkan Gambar 7 menjelaskan bahwa kelas 0 memiliki nilai klasifikasi/probabilitas 0,610 dengan 9 distribusi sedangkan kelas 1 mendapatkan nilai klasifikasi/probabilitas 0,390 dengan 9 distribusi.

4.5. Data Evaluasi

Setelah menyelesaikan proses pemodelan Naive Bayes, dilakukan pengujian akurasi untuk mengevaluasi performa model klasifikasi Naive Bayes yang telah dimodelkan sebelumnya. Akurasi diukur

menggunakan confusion matrix melalui pengunaan alat pada Rapidminer. Hasil dari confusion matrix tersebut dipresentasikan dalam bentuk gambar dengan menampilkan hasil dari confusion matrix.



accuracy: 65.08%

	true 0	true 1	class precision
pred. 0	287	150	65.68%
pred. 1	33	54	62.07%
class recall	89.69%	26.47%	

Gambar 8. Accuracy

Hasil akurasi sebesar 65.08%, dengan class precision buat pred. nol (pred. negative) merupakan 65.68% serta pred satu (pred.positive) merupakan 62.07%. Hasil accuracy didapatkan memakai persamaan 4, dimana nilai true positive sebanyak 287, true negative sebanyak 33, false negative sebanyak 150, serta false positive 33. Dari hasil percobaan tersebut, dapat disimpulkan bahwa penggunaan split data dalam proses klasifikasi berpengaruh pada tingkat akurasi.

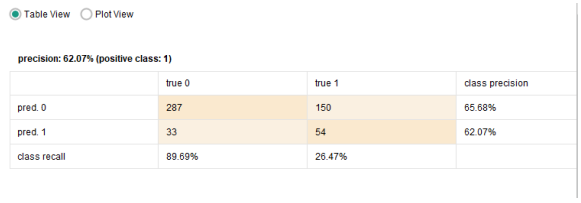


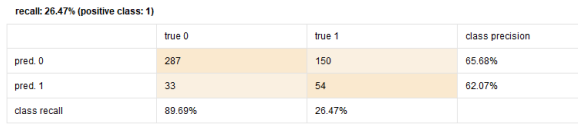
Table View Plot View

precision: 62.07% (positive class: 1)

	true 0	true 1	class precision
pred. 0	287	150	65.68%
pred. 1	33	54	62.07%
class recall	89.69%	26.47%	

Gambar 9. Precision

Dengan menghasilkan precision dengan total percentase 62.07% dari hasil algoritma Naive Bayes.



recall: 26.47% (positive class: 1)

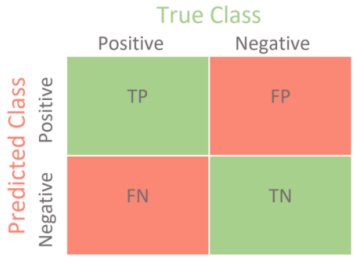
	true 0	true 1	class precision
pred. 0	287	150	65.68%
pred. 1	33	54	62.07%
class recall	89.69%	26.47%	

Gambar 10. Recall

Dengan menghasilkan class recall dengan total percentase 26.47% dari hasil pemodelan algoritma Naive Bayes. Setelah dilakukan pembuatan model dan di dapatkan hasil *accuracy, precision, dan recall* dari metode algoritma Naive bayes kemudian peneliti melakukan perhitungan dengan menggunakan metode *confusion matrix*, penjelasan *confusion matrix* di uraikan dibawah ini:

Pengukuran Akurasi Menggunakan Confusion Matrix

Confusion Matrix adalah metode klasifikasi Di mana kinerja klasifikasi dipengaruhi oleh akurasi klasifikasi, metode klasifikasi Confusion Matrix didasarkan pada hasil klasifikasi sebelumnya. Matrix confusion memberikan informasi tentang perbandingan hasil klasifikasi sistem (model) dengan hasil klasifikasi nyata. Pentingnya confusion matrix adalah bahwa itu akan menunjukkan seberapa baik model yang telah dibuat sebelumnya melalui pengukuran akurasi sebelumnya untuk menentukan seberapa akurat model tersebut. Kinerja model klasifikasi pada set data uji yang nilai sebenarnya diketahui digambarkan dengan matriks kekacauan. Tabel menunjukkan Confusion Matrix yang digunakan untuk menghitung ketepatan.



Gambar 11 Confusion Matrix
Sumber : [16]

Tabel 1. Confusion Matrix

	true positif	true negatif	Class precision
pred. positif	TP = 287	FP = 150	65.68%
pred.negatif	FN=33	TN = 54	62.07%
class recall	89.69%	26.47%	

Hasil pada tabel 1 adalah *TP(True Positif)* =287, *FN(False Negatif)* = 33, *FP(False Positif)* = 150, dan *TN(True Negatif)* = 54. Hasil *Confunsion Matrix* tersebut mempengaruhi perhitungan evaluasi klasifikasi, yaitu *accuracy, precision, dan recall*.

Rumus yang digunakan untuk menghitung *accuracy*:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Sumber : [16]

$$\text{Accuracy} = \frac{287+54}{287+54+150+33} = 0{,}65$$

Rumus yang digunakan untuk menghitung *precision*:

$$\text{Precision} = \frac{TP}{TP+TN}$$

$$\text{Precision} = \frac{TP}{TP+TN} \quad (3)$$

Rumus yang digunakan untuk menghitung *accuracy*:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Recall} = \frac{287}{287+33} = 0{,}26 \quad ()$$

Hasil dari perhitungan *Theorema Bayes* menunjukkan nilai accuracy 0.65%, precision 0.62%, dan recall 0.26%. Sehingga dapat disimpulkan bahwa algoritma Naïve Bayes ini dapat digunakan untuk memprediksi kualitas air nilai keakuratan cukup tinggi. Hal itu dibuktikan dengan nilai accuracy yang tinggi daripada penelitian sebelumnya dengan nilai *accuracy 0.63.*

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian data mining untuk memprediksi kualitas air pada dataset water potability menggunakan algoritma Naïve Bayes dengan Split Validation keakuratannya diuji menggunakan 2620 data. Dari pengujian yang telah dilakukan menghasilkan Data Mining dengan menggunakan algoritma Naïve Bayes dapat digunakan untuk memprediksi kualitas air. Data yang digunakan adalah data repository kaggle dataset. Hasil pengujian mendapatkan accuracy sebesar 65.08%, precision sebesar 62.07%, dan recall sebesar 26.47%. Pada penelitian selanjutnya peneliti menyarankan untuk perbanyak percobaan dengan penggunaan data yang lebih banyak lagi terhadap Algoritma selain Naive Bayes yang digunakan untuk mengukur tingkat akurasi dan presisinya berbeda-berbeda.

DAFTAR PUSTAKA

[1] Shylma Na'imah, "Berbagai Masalah Kesehatan Akibat Pencemaran Air di Indonesia," https://hellosehat.com/sehat/informasi-kesehatan/pencemaran-air-sebab-dan-dampak-kesehatan/.

[2] H. Said, N. Matondang, H. Nurramdhani Irmanda, and S. Informasi, "Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi Application of K-Nearest Neighbor Algorithm to Predict Consumable Water Quality." [Online]. Available: www.kaggle.com

[3] F. Yusa Rahman, I. Indah Purnomo, N. Hijriana, and I. Kalimantan Muhammad Arsyad Al Banjari, "PENERAPAN ALGORITMA DATA MINING UNTUK KLASIFIKASI KUALITAS AIR," 2022.

[4] M. Nazar Yuniar, "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes," *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 243–246, 2023, doi: 10.55338/saintek.v5i1.1383.

[5] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[6] N. Tamsir, S. Amina HUmar, and V. Rosida, "Algoritma Naïve Bayes untuk Prediksi Kualitas Sperma," vol. 6, no. 2, pp. 101–110, 2023, [Online]. Available: http://e-journal.unipma.ac.id/index.php/doubleclickAlgoritmaNaïveBayesuntukPrediksiKualitasSperma…

[7] B. N. Azmi, A. Hermawan, and D. Avianto, "Jurnal Sistem dan Teknologi Informasi Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression," vol. 7, no. 2, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO

[8] R. Manurung, P. Sari Ramadhan, and M. Iswan Perangin-angin, "PERBANDINGAN AKURASI KLASIFIKASI TINGKAT KEMISKINAN ANTARA ALGORITMA C.45 DAN NAIVE BAYES Program Studi Sistem Informasi, STMIK Triguna Dharma *** Program Studi Sistem Informasi, STMIK Triguna Dharma," *Jurnal CyberTech*, vol. x. No.x, [Online]. Available: https://ojs.trigunadharma.ac.id/

[9] A. Wantoro, H. Sulistiyani, Y. Yuniarthe, A. Setya Putra, A. Candra Widyawati, and N. Putra Wicaksono, "Sistem Pakar Diagnosis Penyakit Kutu Ikan Gurami (Argunus Indicus) Menggunakan Metode Naive Bayes 1,*)."

[10] A. Felicia Watratan, A. B. Puspita, D. Moeis, S. Informasi, and S. Profesional Makassar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," 2020. [Online]. Available: http://journal.isas.or.id/index.php/JACOST

[11] M. A. Rahman, N. Hidayat, and A. A. Supianto, "Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[12] A. R. Damanik, S. Sumijan, and G. W. Nurcahyo, "Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes," *Jurnal Sistim Informasi dan Teknologi*, pp. 88–94, Aug. 2021, doi: 10.37034/jsisfotek.v3i3.49.

[13] M. Rianto, R. Rusdiah, and H. Ichwan, "Penerapan Data Mining Dengan Metode Naïve Bayes Dan Learning Vector Quantization Credit Rating Dalam Memprediksi Kelayakan Pemberian Kredit Oleh PT. BPR Lebak Sejahtera," vol. 1.

[14] A. A. Saputro, "Sistem Pendukung Keputusan Penerimaan Bantuan Sosial Program Keluarga Harapan (PKH) dengan Menggunakan Metode Naïve Bayes Classifier (Studi Kasus di Balai Desa Bendungan Kraton Pasuruan)," *Jurnal Ilmiah Edutic : Pendidikan dan Informatika*, vol. 9, no. 1, pp. 40–48, Nov. 2022, doi: 10.21107/edutic.v9i1.12232.

[15] A. Lusiana and P. Yuliarty, "PENERAPAN METODE PERAMALAN (FORECASTING) PADA PERMINTAAN ATAP di PT X," *Industri Inovatif : Jurnal Teknik Industri*, vol. 10, no. 1, pp. 11–20, Jun. 2020, doi: 10.36040/industri.v10i1.2530.

[16] K. Ting, "Confusion Matrix," 2017, p. 260. doi: 10.1007/978-1-4899-7687-1_50.