INOTEK, Vol.8 Agustus 2024
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Perbandingan Metode Algoritma *Decission Tree* dan *K-Nearest Neighbors* untuk Memprediksi Kualitas Air yang dapat dikonsumsi

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1*]Deri Fitriono, [2]Saylendra Arga Wardani, [3]M Nizar Bahri Al Varuq, [4]Aidina Ristyawan, [5] Erna Daniati
[1-5]Universitas Nusantara PGRI Kediri
[1]derifitriono70@gmail.com, [2]argasaylendra@gmail.com, [3]nizarqw8@gmail.com, [4]adinaristi@unpkediri.ac.id, [5]ernadaniati@unpkediri.ac.id

Abstrak—Air merupakan kebutuhan yang sangat penting bagi mahkluk hidup termasuk manusia, namun tidak semua air aman untuk dikonsumsi, sehingga perlu adanya identifikasi terkait kualitas air yang baik untuk dikonsumsi. Oleh karena itu sangat penting mengembangkan strategi yang tepat untuk memprediksi atau meramlkan kualitas air yang dapat dikonsumsi. Pada penelitian ini akan menggunakan perhitungan Decission Tree dan K-Nearest Neigbors untuk klasifikasi sifat air yang layak dikonsumsi. Kualitas air yang baik sangat penting untuk kesehatan manusia, dan prediksi yang akurat dapat membantu orang memilih jumlah air yang tepat untuk diminum. Kedua 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 Decision Tree sebesar 75.69%, sedangkan metode K-nearest Neighbors memiliki tingkat akurasi sebesar 79,39%, yang merupakan metode yang paling baik untuk klasifikasi data

Kata Kunci—Air;Decission Tree;K-Nearest Neigbors

Abstract Water is a very important need for living things including humans, but not all water is safe for consumption, so it is necessary to identify the quality of water that is good for consumption. Therefore, it is very important to develop the right strategy to predict or predict the quality of water that can be consumed. This research will use Decission Tree and K-Nearest Neigbors calculation for the classification of water properties that are suitable for consumption. Good water quality is essential for human health, and accurate prediction can help people choose the right amount of water to drink. These two algorithms will be compared in the data classification process to find out which method is the most accurate, judging by the highest accuracy rate. The results of this study show that the Decision Tree method is 75.69%, while the K-nearest Neighbors method has an accuracy rate of 79.39%, which is the best method for data classification.

Keywords*—water;Decision Tree;K-Nearest Neighbors*

Penulis Korespondensi:

Aidina Ristyawan,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: adinaristi@unpkediri.ac.id,
ID Orcid: [https://orcid.org/0009-0003-2712-1507]
Handphone: 081232624460

I. PENDAHULUAN

Air adalah sumber daya alam yang sangat penting bagi manusia untuk bertahan hidup. Semua organ masyarakat berusaha semaksimal mungkin untuk mendapatkan sumber air terbaik untuk memenuhi kebutuhan sehari-hari mereka. Air digunakan sebagai sumber daya untuk berbagai tujuan, termasuk minum, perumahan, irigasi, peternakan, perikanan, pembangkit listrik, transportasi, industri, dan tempat rekreasi[1]. Sementara disisi lain kebutuhan air bersih terus meningkat setiap tahunnya, ketersediaan air bersih sangat terbatas. Hal ini disebabkan oleh pembangunan yang semakin meningkat tanpa mempertimbangkan lingkungan sekitar dan daerah resapan yang sempit, terutama di daerah perkotaan. Akibatnya, kurangnya ketersediaan sumber air bersih menjadi masalah besar di Indonesia. Pencemaran air adalah sumber utama pencemaran air di Indonesia.

Masalah air yang terkontaminasi ini dapat membawa berbagai penyakit menular, seperti diare, kolera, dan penyakit lain yang terkait dengan infeksi bakteri dan virus, oleh karena itu, pemantauan dan klasifikasi kualitas air sangatlah penting Dalam menghadapi masalah ini. Teknologi dan metode ilmiah sangat penting untuk memantau dan memprediksi kualitas air. Pengguanaan teknologi berbasis data untuk analisis dan prediksi kualitas air dapat membantu dalam pengambilan keputusan yang lebih cepat dan akurat. Misalnya, algoritma pembelajaran mesin dan data mining dapat digunakan untuk mengolah data kualitas air dan membuat prediksi yang akurat tentang layak atau tidaknya air untuk dikonsumsi[2].

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam *data mining* sangat bervariasi[3].

Penelitian tentang klasifikasi untuk memprediksi kualitas air yang dapat dikonsumsi sudah banyak dilakukan, terutama dalam hal klasifikasi air minum menggunakan pembelajaran mesin. Yang dilakukan oleh Hariana Said, Nurhafifah Matondang, Helena Nurramdhani Irmanda dengan judul "Penerapan Algoritma *K-Nearest Neighbor* Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi" dari hasil penelitian tersebut diperoleh pengukuran model menggunakan table *confusion matrix* yaitu memiliki Tingkat akurasi tertinggi sebesar 85,24%[4]. Dan pada penelitian yang dilakukan oleh Fauzi Yusa Rahman, Indu indah Purnomo, Nadya Hijriana dengan judul "PENERAPAN ALGORITMA DATA MINING UNTUK KLASIFIKASI KUALITAS AIR" menggunakan algoritma *Decission Tree* diperoleh nilai akurasi yang cukup tinggi yaitu sebesar 94,94% Dengan nilai AUC sebesar 0,865 sehingga termasuk golongan klasifikasi yang baik[5] Selanjutnya pada penelitian yang dilakukan Adrian Wisnu Saputra, Ade Irma Purnamasari, Irfan Ali menggunakan *Naive Bayes* dengan judul

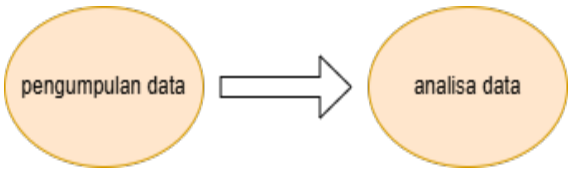
"IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI KUALITAS AIR YANG DAPAT DI KONSUMSI" didapatkan hasil yang cukup kecil yaitu hasil analisis menunjukkan bahwa model klasifikasi *Naive Bayes* mampu memprediksi kualitas air dengan akurasi sebesar 65,08 persen, presisi 62,08 persen, dan recall 26,47 persen[6].

Pada penelitian ini memanfaatkan data kualitas air dari dataset Kaggle dengan judul *Water quality and Potability* dalam format CSV, yang mencakup berbagai nilai pada sepuluh atribut. Atribut *"potability"* berfungsi sebagai label yang menunjukkan apakah air tersebut aman atau tidak aman untuk diminum, berdasarkan persentase nilai dari masing-masing atribut dalam dataset.

Berdasarkan permasalahan yang ada, yaitu untuk membandingkan kedua metode *Naïve Bayes* dan *K-Nearest Neighbors (KNN),* penelitian ini dilakukan dengan judul "Perbandingan Metode Algoritma *Naïve Bayes* dan *K-Nearest Neighbors* untuk Memprediksi Kualitas Air yang Dapat Dikonsumsi". Penelitian ini menggunakan software RapidMiner untuk mengetahui nilai akurasi tertinggi dari kedua metode yang akan diimplementasikan dalam klasifikasi data. Analisis perbandingan akurasi kualitas air akan dilakukan dengan menggunakan klasifikasi data *Naïve Bayes* dan *K-Nearest Neighbors*, yang bertujuan untuk menentukan metode yang paling efektif dalam klasifikasi kualitas air dengan hasil akurasi yang paling optimal.

II. METODE

Dalam penelitian ini menggunakan pendekatan kuantitatif yang menekankan analisisnya pada data numerikal, juga dikenal sebagai angka-angka, yang diproses dengan teknik statistik [7]. Sebagai bahan pengolah data mining, data numerik diubah menjadi data kualitatif selama proses pengumpulan data. Data yang digunakan berasal dari Dataset *Water Quality and Potability Kaggle*[8]. Selanjutnya, data akan dipelajari dengan menggunakan metode algoritma *Decision Tree* dan *K-Nearest Neighbors*. Pada penelitian ini juga digunakan fase penelitian berikut:



Gambar 1. Fase Riset

2.1. Pengumpulan Data

Pengumpulan ini mengunakan metode kuantitatif. Tahap pengumpulan data dilakukan dengan teknik pengumpulan data yang menggunakan data dari kaggle sebesar 2620 record data

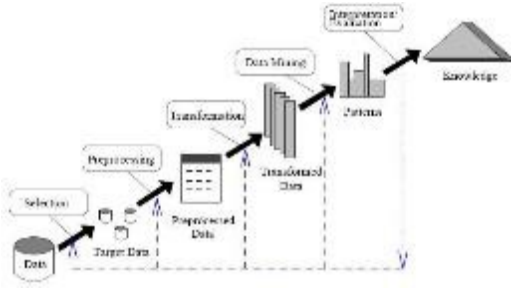
tentang *Water Quality and Potability. Kaggle* merupakan situs/platform yang resmi untuk mengadakan perlombaan-perlombaan di bidang *Data Science*, dimana situs ini merupakan sumber pembelajaran data science[9].

Tabel 1. *Dataset Water Quality and Potability*

Feature	Deskripsi
pH	Menunjukkan keasaman atau kebasaan air
Hardness	kandungan mineral kalsium dan magnesium dalam air
Solids	Menunjukkan jumlah zat padat terlarut dalam air.
Chloramines	Kandungan kloramin dalam air
Sulfate	Kandungan sulfat dalam air.
Conductivity	Mengukur kemampuan air untuk menghantarkan listrik,
Organic_carbon	Kandungan karbon organik dalam air.
Trihalomenthanes	Kandungan trihalomethane dalam air, yang terbentuk sebagai hasil sampingan dari desinfeksi air dengan klorin.
Turbidity	Mengukur kekeruhan air.
Potability	Indikator apakah air aman untuk diminum atau tidak (1 = Aman, 0 = Tidak Aman).

2.2. Analisa Data

Penelitian ini mengunakan tehnik Knowledge Discovery Data (KDD) yaitu proses untuk meringkas sebuah informasi yang bermanfaat yang sebelumnya belum diketahui dan tidak terlihat [10]. untuk menganalisa data teknik terdiri dari beberapa tahap,yaitu :



Gambar 2. Tahapan Penelitian

1. *Data Selection*

   Data Selection adalah pengambilan dan pemilihan sampel data yang ingin diolah untuk menjadi informasi penting[11]. Hasil yang diperoleh dari teknik data mining disimpan

dalam file yang terpisah dari database produksi. Pada tahap ini, data diambil dari Dataset *Water Quality and Potability Kaggle*

2. *Data Preprocessing*

   Preprocessing adalah prosedur penggeledahan data yang bertujuan untuk mengenali, mencium, dan memulihkan (maupun menghilangkan) peringatan yang cacat atau tidak cermat dari sejumlah peringatan, grafik, maupun database. Preprocessing data penting dalam analisis data mining untuk membersihkan, mengubah format, dan mempersiapkan data agar lebih mudah dan akurat.[12]

3. *Data Transformation*

   Tahapan *preprocessing* data mining yang sangat penting adalah persiapan data. Ini karena kualitas input data sangat memengaruhi kualitas output analisis yang dihasilkan. Salah satu teknik *preprocessing* data adalah data *transformation*, yang mengubah atau mengkonsolidasi data ke bentuk yang sesuai untuk mining [13].

4. *Data Mining*

   Data mining adalah metode untuk menemukan pola dan data menarik dalam kumpulan data tertentu dengan menggunakan teknik khusus. Data mining memiliki banyak metode dan algoritma yang berbeda. Metode dan algoritma yang tepat sangat bergantung pada tujuan dan teknik KDD secara totalitas. Pada penelitian ini, prosedur penggalian data menggunakan algoritma Naïve Bayes dan K-Nearest Neighbors.

5. *Evaluation*

   interpretasi atau evaluasi adalah hasil dari data mining di analisis untuk memastikan bahwa apakah pola atau Informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya

III. HASIL DAN PEMBAHASAN

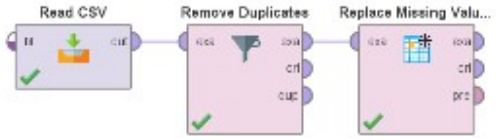
1. Data Selection

Langkah pertama dalam mengelola data adalah memasukkan port Read Excel untuk membaca dataset dalam format csv ditunjukan pada gambar 3 dibawah. Dalam penelitian ini, atribut yang disebut potability menunjukkan apakah air aman untuk dikonsumsi manusia. Labeling dapat dilakukan dengan mengubah warna label untuk memudahkan penelitian. Ada dua jenis air: yang boleh diminum dan yang tidak boleh diminum .

Gambar 3. *Operator CSV*

2. Preprocessing

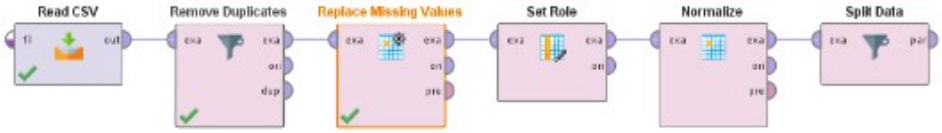
Pada langkah kedua dilakukan *preprocessing*. Pada tahap ini, data duplikat dibersihkan atau dihilangkan. Ini dilakukan untuk membuat dataset imbang untuk pemrosesan berikutnya. Selain itu, port Remove Duplicates digunakan untuk membuat pemodelan replace Missing Value untuk menghilangkan data yang kosong pada atribut dataset. Berikut adalah sub proses tersebut ditunjukan pada Gambar 3.



Gambar 4. *Data Preprocessing*

3. Transformation

Langkah ketiga adalah set role yaitu mengatur peran dari setiap kolom dalam dataset serta mengatur peran kolom target (potabilitas air) sebagai label, pada tahap selanjutnya menggunakan normalize untuk menormalkan data agar memastikan semua fitur atau data berada pada skala yang sama. Selanjutnya, menggunakan port Split Data untuk mengubah data seleksi. Ini membagi dataset menjadi data pelatihan dan data pengujian dengan parameter 0,7:0,3, yang berarti 0,7 atau 70% untuk data pelatihan dan 0,3 atau 30% untuk data pengujian, Pada gambar dibawah Ini adalah langkah-langkah yang dilakukan :



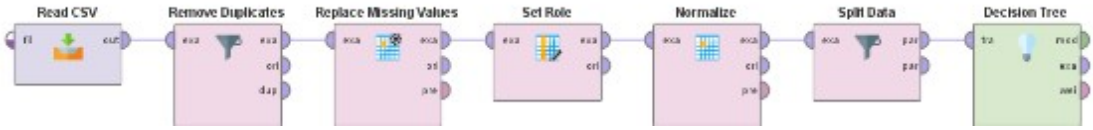
Gambar 5. *Data Transformation*

Gambar 6. Parameter *Split Data*

4. Data Mining

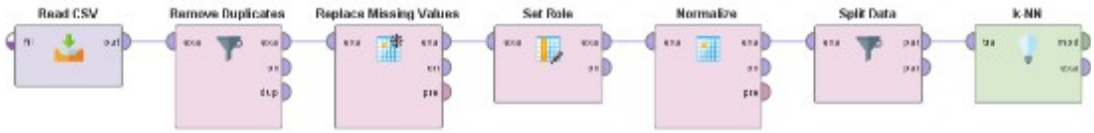
Dalam langkah keempat penelitian ini, data mining menggunakan model Algoritma Decision Tree dan Algoritma K-nearest neighbors untuk klasifikasi, kemudian dilanjutkan

untuk membandingkan hasil dari kedua model algoritma tersebut. Selanjutnya, model validasi split ditambahkan, proses ini menghubungkan port Decision Tree dan Algoritma K-nearest neighbors, port apply model, dan port performance. Berikut Ini adalah langkah-langkah yang dilakukan :



Gambar 7. Pemodelan *Decision tree*

Decision tree adalah sebuah struktur yang digunakan untuk memecah kumpulan data besar menjadi kelompok-kelompok record yang lebih kecil dengan menggunakan serangkaian aturan Keputusan [14].



Gambar 8. Pemodelan K-nearest neighbors

K-Nearest Neighbor (KNN) adalah sebuah metode untuk mengklasifikasikan sebuah objek berdasarkan data pembelajaran yang secara keseluruhan paling dekat dengan objek yang dimaksud. Data pembelajaran diproyeksikan ke ruang berdimensi, dimana parang memproyeksikan fitur data masing-masing. Rentang ini diklasifikasikan sebagai bagian atas berdasarkan klasifikasi data pembelajaran[15].

5. Data Evaluasi

Pengujian akurasi dilakukan secara berurutan setelah proses pemodelan Decision tree dan K-nearest neighbors selesai. Tujuannya adalah untuk mengevaluasi bagaimana model klasifikasi Decision tree dan K-nearest neighbors yang telah dimodelkan sebelumnya berfungsi.

3.1. Pemodelan Decision Tree

accuracy: 75.89%

	true 0	true 1	class precision
pred. 0	1358	597	72.90%
pred. 1	40	425	81.40%
class recall	97.50%	41.58%	

Gambar 9. *Accuracy Model Decision Tree*

Hasil akurasi sebesar 75.69%, dengan class precision buat pred. nol (pred. negative) merupakan 72.30% serta pred satu (pred.positive) merupakan 91.40%. Hasil accuracy didapatkan memakai persamaan 4, dimana nilai true positive sebanyak 1558, true negative sebanyak 40, false negative sebanyak 597, serta false positive 425. Dari hasil percobaan tersebut, dapat disimpulkan bahwa penggunaan split data dalam proses klasifikasi berpengaruh pada tingkat akurasi.

precision: 91.40% (positive class: 1)

	true 0	true 1	class precision
pred. 0	1558	597	72.30%
pred. 1	40	425	91.40%
class recall	97.50%	41.59%	

Gambar 10. *Precision Model Decision Tree*

Dengan menghasilkan precision dengan total persentase 91.40% dari hasil algoritma Decision Tree.

recall: 41.59% (positive class: 1)

	true 0	true 1	class precision
pred. 0	1558	597	72.30%
pred. 1	40	425	91.40%
class recall	97.50%	41.59%	

Gambar 11. Recall Model Decision Tree

Dengan menghasilkan class recall dengan total percentase 41.59% dari hasil pemodelan algoritma Decision Tree. Setelah dilakukan pembuatan model dan di dapatkan hasil accuracy, precision, dan recall dari metode algoritma Decision Tree kemudian peneliti melakukan perhitungan dengan menggunakan metode confusion matrix.

3.2. Pemodelan K-nearest neighbor

accuracy: 79.39%

	true 0	true 1	class precision
pred. 0	1502	444	77.18%
pred. 1	96	578	85.76%
class recall	93.99%	56.56%	

Gambar 12. Accuracy Model K-nearest neighbor

Hasil akurasi sebesar 79.39%, dengan class precision buat pred. nol (pred. negative) merupakan 77.18% serta pred satu (pred.positive) merupakan 85.76%. Hasil accuracy didapatkan memakai pendekatan K-10, dimana nilai true positive sebanyak 1502, true

negative sebanyak 578, false negative sebanyak 578, serta false positive 444. Dari hasil percobaan tersebut, dapat disimpulkan bahwa penggunaan pendekatan K di port K-nearest neighbor dalam proses klasifikasi berpengaruh pada tingkat akurasi.

precision: 85.76% (positive class: 1)

	true 0	true 1	class precision
pred. 0	1502	444	77.18%
pred. 1	96	578	85.76%
class recall	93.99%	56.56%	

Gambar 13. Precision Model K - Nearest Neighbor

Dengan menghasilkan precision dengan total persentase 85.76% dari hasil algoritma K-nearest neighbor

recall: 56.56% (positive class: 1)

	true 0	true 1	class precision
pred. 0	1502	444	77.18%
pred. 1	96	578	85.76%
class recall	93.99%	56.56%	

Gambar 14. Recall Model K - Nearest Neighbor

Dengan menghasilkan class recall dengan total percentase 56.56% dari hasil pemodelan algoritma K-nearest neighbor. Setelah dilakukan pembuatan model dan di dapatkan hasil accuracy, precision, dan recall dari metode algoritma K-nearest neighbor kemudian peneliti melakukan perhitungan dengan menggunakan metode confusion matrix.

3.3. Hasil Akurasi

Tabel 6. Perbandingan Akurasi

Decision Tree	K - Nearest Neighbors
75.69%	79.39%

Analisis perbandingan akurasi klasifikasi kualitas air menggunakan metode K-Nearest Neighbors (KNN) dan Decision Tree menunjukkan bahwa K – Nearest Neighbors menghasilkan akurasi tertinggi sebesar 79,39% dalam penelitian ini, sementara Decision Tree memiliki akurasi sebesar 75,69%.

IV. KESIMPULAN

Tujuan dilakukan penelitian ini untuk mengetahui hasil perbandingan tingkat keakuratan dengan menggunakan record sebesar 2620 data dari metode penelitian yang digunakan yaitu Decision Tree dan K-Nearest Neighbors. Dilihat dari Class Recall dan Class Precision metode yang menghasilkan tingkat keakuratan yang paling tinggi adalah K - Nearest Neighbor yaitu sebesar 79.39%. sedangkan pada metode Decission Tree

menghasilkan akurasi sebesar 75.69% Metode klasifikasi Decision Tree dan KNN pada penelitian ini cukup baik digunakan karena menghasilkan tingkat akurasi diatas 75%, namun untuk mendapatkan hasil akurasi yang lebih maksimal untuk penelitian selanjutnya bisa menggunakan metode yang lain

DAFTAR PUSTAKA

[1] E. W. Annisa Weningtyas, "Pengelolaan Sumber Daya Air Berbasis Kearifan Lokal Sebagai Modal Untuk Pembangunan Berkelanjutan".

[2] L. Savitri and R. Nursalim, "Klasifikasi Kualitas Air Minum menggunakan Penerapan Algoritma Machine Learning dengan Pendekatan Supervised Learning." [Online]. Available: https://ejournal.unib.ac.id/diophantine,

[3] P. Bidang Komputer Sains dan Pendidikan Informatika, D. Akademi Perekam dan Informasi Kesehatan Iris Padang Jl Gajah Mada No, and S. Barat, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4.5 Yuli Mardi".

[4] 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

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

[6] A. W. Saputra, A. I. Purnamasari, and I. Ali, "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI KUALITAS AIR YANG DAPAT DI KONSUMSI," 2024.

[7] C. Riang, S. U. Al, and A. Mandar, "Teknik Data Mining Menggunakan Classification Dalam Sistem Penunjang Keputusan Peminatan SMA Negeri 1 Polewali," Online.

[8] LAKSIKA THARMALINGAM, "Water Quality and Potability," kaggle.com.

[9] K. Akmal, A. Faqih, and F. Dikananda, "PERBANDINGAN METODE ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENYAKIT STROKE," 2023. [Online]. Available: www.researchgate.net

[10] D. Fitrianah, S. Dwiasnati, H. H. H, and K. A. Baihaqi, "Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes," *Faktor Exacta*, vol. 14, no. 2, p. 92, Aug. 2021, doi: 10.30998/faktorexacta.v14i2.9297.

[11] A. Yoga Pratama *et al.*, "Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," 2021.

[12] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[13] H. Junaedi *et al.*, "Prosiding Konferensi Nasional 'Inovasi dalam Desain dan Teknologi'-IDeaTech 2011 DATA TRANSFORMATION PADA DATA MINING".

[14] A. Muzakir and R. A. Wulandari, "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," *Scientific Journal of Informatics*, vol. 3, no. 1, 2016, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[15] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," 2012.