JURNAL SISFOTENIKA
Vol. 14 No. 2, Juli 2024
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/sisfotenika.v14i2.430

Prediksi Kualitas Susu Menggunakan Metode K-Nearest Neighbors

Milk Quality Prediction Using The K-Nearest Neighbors Method

Nazori Suhandi[1] Rendra Gustriansyah[2] Abel Destria*[3], Marshanda Amalia[4], Via Kris[5]
[1,2,3,4,5]Teknik Informatika, Fakultas Ilmu Komputer dan Sains, Palembang, Universitas Indo Global Mandiri; Jl. Jend. Sudirman Km.4 No. 62, Palembang
e-mail: [1]nazori@uigm.ac.id, [2]rendra@uigm.ac.id, *[3]2021110084@students.uigm.ac.id, [4]2021110097@students.uigm.ac.id , [5]2021110020@students.uigm.ac.id

Abstrak

Susu adalah sumber nutrisi yang kaya akan kalsium dan laktosa, yang memiliki peran penting dalam mengatasi kekurangan gizi. Kualitas susu ditentukan oleh pH dan proses pasteurisasi. Penelitian ini bertujuan untuk memprediksi kualitas susu menggunakan Metode K-Nearest Neighbors (K-NN). Analisis dilakukan melalui serangkaian langkah, termasuk pra-pemrosesan data yang meliputi encoding data kategorikal, penanganan nilai yang hilang, dan pembersihan data. Selanjutnya, K optimal dipilih menggunakan metode elbow, dengan nilai K=3. Data kemudian dipisahkan menjadi data pelatihan dan data uji untuk menghindari overfitting dan memvalidasi performa model, dan hasil pengujian penggunaan K-NN untuk memprediksi kualitas susu dengan menguji tiga skema pembagian data yang berbeda: 80-20, 70-30, dan 60-40. Dengan memanfaatkan Confusix untuk menghitung precision, recall, dan accuracy, kita dapat menilai proporsi kasus positif yang terklasifikasi dengan benar, diidentifikasi dengan benar. Dimana Hasil accuracy terbaik diperoleh dari skema satu sebesar 0,94, recall 0,8, dan precision mencapai 1. Penelitian ini memberikan kontribusi penting dalam memahami, memprediksi, dan memantau kualitas susu, ini mencakup pemahaman yang mendalam tentang faktor-faktor yang memengaruhi kualitas susu, pengembangan model prediktif yang maju. Secara keseluruhan, penelitian ini memperkuat dasar ilmiah untuk industri susu secara menyeluruh.

Kata kunci*— Susu, Kualitas, K-NN, Ellbow, Klasifikasi.*

Abstract

Milk is a nutrient-rich source abundant in calcium and lactose, playing a crucial role in addressing nutritional deficiencies. Milk quality is determined by pH levels and pasteurization processes. This research aims to predict milk quality using the K-Nearest Neighbors (K-NN) Method. The analysis is conducted through a series of steps, including data preprocessing involving categorical data encoding, handling missing values, and data cleansing. Subsequently, the optimal K value is selected using the elbow method, with a value of K=3. The data is then divided into training and testing sets to avoid overfitting and validate model performance, and the testing results of using K-NN to predict milk quality are evaluated using three different data splitting schemes: 80-20, 70-30, and 60-40. By utilizing Confusion Matrix to calculate precision, recall, and accuracy, we can assess the proportion of correctly classified positive cases, accurately identified. The best accuracy result is obtained from scheme one at

0,94, with a recall of 0.8, and precision reaching 1. This research provides a significant contribution to understanding, predicting, and monitoring milk quality, encompassing a profound understanding of factors influencing milk quality and the development of advanced predictive models. Overall, this study strengthens the scientific foundation for the dairy industry comprehensively.

Keywords—*Milk, Quality, K-NN, Elbow, Classification*

1. PENDAHULUAN

Susu sebagai salah satu dari sedikit makanan alami yang hampir tak tertandingi, memiliki ciri khas yang menonjol dan memiliki kandungan kalsium yang amat tinggi dan laktosa yang mendukung. Ditengah tantangan global terkait masalah gizi yang terus menjadi perhatian dalam kesehatan masyarakat, kebutuhan akan solusi yang praktis untuk meningkatkan keseimbangan nutrisi, terutama saat menghadapi kekurangan gizi, semakin mendesak. Dalam hal ini, susu menjadi salah satu pilihan penting untuk meningkatkan asupan nutrisi. Karakteristiknya yang unik, seperti warna kekuningan yang tidak tembus cahaya, aroma sapi yang khas, dan rasa manisnya, membuatnya diminati. Namun, tidak hanya nutrisi yang diperhatikan; standar kualitas juga menjadi fokus, dengan pH yang ideal dan proses pasteurisasi pada suhu yang tepat. Dengan situasi di mana keamanan dan kualitas makanan menjadi perhatian utama, pemahaman mendalam tentang karakteristik dan persyaratan standar menjadi kunci dalam memastikan susu yang aman dan bermutu tinggi bagi konsumen [1-2].

Penelitian ini membahas penggunaan metode K-NN dalam mengklasifikasikan kualitas susu pasteurisasi. Meskipun accuracy mencapai 0,97 untuk K = 3, namun perlu diperhatikan beberapa kelemahan, seperti pendekatan manual dalam menentukan nilai optimal K, yang dapat menyebabkan potensi untuk mengabaikan nilai K yang lebih optimal tanpa menggunakan metode elbow. Selain itu, pengujian dilakukan hanya tiga kali dengan nilai K yang tetap, dan pembagian dataset hanya dalam satu skenario 80:20, tanpa mengeksplorasi rasio pembagian yang berbeda [3].

Penelitian lainnya mengimplementasikan tiga metode Machine Learning, yaitu Random Forest, K-NN, dan Neural Network untuk mengklasifikasikan sampel susu. Hasilnya menunjukkan performa terbaik Random Forest dengan accuracy 0,96, precision 0,98, dan recall 0,94. Sementara K-NN memiliki accuracy 0,82, precision 0,81, dan recal 0,87, dan Neural Network memiliki accuracy 0,54, precision 0,82, dan recal 0,20. Namun, penelitian ini memiliki kekurangan dalam skema pembagian dataset, hanya menggunakan split data 70:80 untuk pelatihan tanpa validasi silang dan tidak menggunakan metode elbow untuk menentukan nilai K yang optimal untuk K-NN[4].

Penelitian terakhir ini melibatkan penggunaan algoritma Backpropagation. Dataset yang terdiri dari tujuh fitur (pH, Suhu, Rasa, Bau, Lemak, Kekeruhan, Warna) dibagi menggunakan K-fold Cross Validation dengan 3 fold. Dari penentuan nilai parameter optimal, ditemukan bahwa learning rate 0.5, hidden weight 7, dan epoch 750 merupakan konfigurasi terbaik. Implementasi ini berhasil mencapai accuracy sebesar 0,97,923, menunjukkan keefektifan metode Backpropagation dalam klasifikasi kualitas susu sapi[5].

Pada penelitian berikutnya, akan dilakukan peningkatan untuk mengatasi kekurangan yang ada pada penelitian sebelumnya. Ini mencakup penerapan metode elbow secara sistematis untuk menentukan nilai optimal K pada metode K-NN, serta variasi pembagian dataset untuk memahami lebih baik pengaruh rasio pembagian terhadap performa model. Selain itu, akan dimasukkan penggunaan metode validasi silang dan dataset pengujian terpisah untuk evaluasi

model yang lebih akurat dan dapat diandalkan. Selain menampilkan accuracy, evaluasi model akan mencakup precision dan recall untuk memberikan pemahaman yang lebih komprehensif tentang performa model klasifikasi susu sapi.

Investasi dalam pengembangan sistem pemantauan yang canggih penting bagi produsen susu untuk mempertahankan reputasi dan memastikan kepuasan pelanggan. Metode K-NN digunakan untuk meramalkan kualitas susu dengan akurat berdasarkan analisis data. K-NN adalah metode sederhana namun efektif dalam klasifikasi, di mana data yang akan diklasifikasikan dibandingkan dengan data pelatihan terdekat dalam ruang fitur. Jarak antar data dalam metode K-NN biasanya dihitung menggunakan metode jarak Euclidean [6].

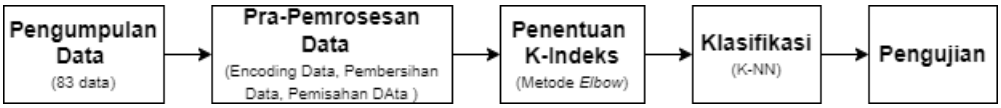
Metode ini diterapkan karena kemudahan implementasinya dan kemampuannya dalam menangani data yang kompleks serta non-linear. Kelebihan K-NN terletak pada prinsip kerja yang sederhana, di mana metode ini menentukan klasifikasi berdasarkan jarak terdekat antara sampel uji dan sampel latih, tanpa perlu mempertimbangkan distribusi kelas yang ada [7],

Metode K-NN memberikan kerangka kerja yang handal untuk mengidentifikasi pola-pola dalam dataset susu yang luas dan beragam, serta memberikan fleksibilitas dalam memodelkan hubungan antar variabel.

Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi seberapa efektifnya metode K-NN dalam mengklarifikasikan kualitas susu dan menguji kemampuan K-NN dalam memprediksi kualitas susu berdasarkan atribut- atribut yang ada. Diharapkan hasil dari penelitian ini dapat memberikan wawasan yang lebih mendalam tentang potensi metode K-NN dalam meningkatkan pengawasan dan pengendalian kualitas susu, serta berperan dalam pengembangan sistem pemantauan yang lebih canggih bagi para produsen susu..

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini terdiri dari beberapa langkah yang dirancang untuk mengumpulkan, mengolah, dan menganalisis data dengan tujuan untuk memahami dan memprediksi kualitas susu yang dapat dilihat pada Gambar 1.



Gambar 1 Tahapan Penelitian

Pada Gambar 1 merupakan tahapan pada penelitian yang dimulai dari mengumpulkan data dari sumber publik yang tersedia, dengan dataset terdiri dari sejumlah data dengan atribut-atribut yang mencakup informasi tentang kualitas susu. Langkah selanjutnya melibatkan pra-pemrosesan data, di mana langkah-langkah seperti mengonversi atribut yang bersifat deskriptif menjadi bentuk numerik, menghapus duplikat, dan menangani nilai yang hilang. Tujuan dari langkah ini adalah untuk memastikan keakuratan data serta mempermudah analisis lebih lanjut.

Pengumpulan data

Pada tahap pengumpulan data, analisis kualitas susu akan dilakukan menggunakan K-NN dengan memanfaatkan dataset yang merupakan sampel data susu yang diperoleh dari peternakan susu segar. Dataset ini terdiri dari 100 data dengan 7 atribut yaitu, pH, Temperature, Taste, Odor, Fat, Turbity dan Colour. Adapun sampel data dari kualitas susu dapat dilihat pada Tabel 1.

Tabel 1. Sampel Data Kualitas Susu

No	pH	Temperature	Taste	Odor	Fat	Turbidity	Colour	Grade

1	6.6	35	1	0	1	0	254	high
2	6.6	36	0	1	0	1	253	high
3	8.5	70	1	1	1	1	246	low
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	8.6	55	0	1	1	1	255	low

Pra-Pemrosesan Data

Data yang digunakan tidak selalu dalam kondisi optimal untuk diproses, karena dapat mengandung berbagai masalah seperti nilai yang hilang, data yang berlebihan, outlier, atau format data yang tidak sesuai. Masalah-masalah tersebut dapat mempengaruhi hasil dari proses analisis data itu sendiri, seperti proses klasifikasi data. Oleh karena itu, pra-pemrosesan menjadi langkah yang diperlukan untuk mengatasi masalah-masalah tersebut dan memastikan data siap untuk diproses lebih lanjut [9].

Tahap ini bertujuan untuk mengubah data mentah menjadi data yang terstruktur dan bersih agar dapat diolah secara efisien, sehingga menghasilkan hasil yang optimal. Langkah-langkah yang dilakukan dalam pra-pemrosesan data dapat dilihat pada gambar 2.

Gambar 2 Tahapan Pra-Pemrosesan Data

a) Encoding Data,
   Encoding data dilakukan ketika dataset mengandung variabel kategorikal seperti kelas kualitas susu (contohnya: high, medium, low), diperlukan konversi variabel tersebut menjadi bentuk numerik agar bisa dimasukkan ke dalam model analisis.
b) Pembersihan Data,
   Pembersihan data melibatkan eliminasi data duplikasi dan data yang tidak valid. Tujuannya adalah untuk memastikan bahwa dataset yang digunakan bersih dan tidak mengandung data yang tidak valid atau tidak relevan.. Dengan demikian, proses pembersihan tersebut menghasilkan dataset yang relevan dan valid untuk analisis lebih lanjut.
c) Pemisahan Data
   Pemisahan data yaitu membagi dataset menjadi dua bagian: data latih dan data uji. Data latih akan digunakan untuk melatih model metode K-NN, sementara data uji akan digunakan untuk menguji kinerja model. Pemisahan data ini penting untuk mengukur seberapa baik model dapat menggeneralisasi pada data baru yang belum pernah dilihat sebelumnya.

Tahap-tahap ini memastikan bahwa data yang kita gunakan siap untuk diproses lebih lanjut dan memberikan hasil analisis yang akurat dan relevan. Dengan encoding data, pembersihan data, dan pemisahan data, kita dapat membangun model prediksi yang lebih handal dan efektif, yang pada akhirnya akan membantu dalam memahami dan memprediksi kualitas susu dengan lebih baik. Proses ini tidak hanya meningkatkan kualitas data, tetapi juga meningkatkan kemampuan model dalam memberikan hasil yang lebih baik dan dapat diandalkan.

Penentuan K-Indeks

Penentuan K-Indeks dilakukan dengan menggunakan metode elbow untuk menentukan jumlah optimal dari klaster dengan melihat sudut yang terbentuk pada kurva perbandingan

antara jumlah klaster. Jika nilai dari klaster pertama dengan klaster kedua memiliki penurunan nilai terbesar, maka nilai klaster tersebut adalah nilai klaster terbaik. Jumlah klaster yang terbaik, 'k', akan dipilih pada titik tersebut. Metode ini merupakan metode visual yang melihat variasi intra-klaster total atau jumlah Kuadrat Sum dari Klaster Dalam (WSS) sebagai fungsi dari jumlah klaster. Semakin besar jumlah klaster k, nilai WSS akan semakin kecil atau sebaliknya [10]. Rumus metode elbow dapat dinyatakan sebagai berikut:

$$WSS = \sum_{k=1}^{k} \sum_{i=1}^{n} || \; x_i^{(j)} - c_j ||^2 \quad (1)$$

k adalah jumlah total cluster, n merupakan jumlah objek dalam setiap cluster, dimana xI menunjukkan elemen i dalam suatu cluster, dan c_j adalah titik tengah cluster ke-j.

Klasifikasi

Klasifikasi dilakukan mengunakan metode K-NN yang digunakan untuk mengklasifikasikan objek berdasarkan data pembelajaran yang memiliki jarak paling dekat dengan objek yang akan diklasifikasikan. Metode ini melakukan klasifikasi dengan membandingkan kemiripan atau kedekatan data baru dengan data lainnya. Dengan menggunakan similarity atau kesamaan, metode ini dapat mengklasifikasikan data ke dalam kelas yang sesuai[11]. Adapun Langkah-langkah untuk perhitungan K-NN sebagai berikut:

1. Menetapkan nilai parameter K dengan metode elbow.
2. Menghitung jarak antara setiap data dalam data latih dan data uji. Jarak biasanya diukur menggunakan metode Euclidean, yang dirumuskan sebagai berikut:
   $$de = \sqrt{\sum_{k=1}^{n} (X_i - Y_i)^2}$$
   de merepresentasikan jarak Euclidean, X_i merujuk pada data uji ke-i, dimana Y_i menandakan data latih ke-i, dan *n* menunjukkan jumlah keseluruhan data yang digunakan.
3. Memilih K jarak terdekat.
4. Menetapkan kelas yang sesuai dengan data terdekat.
5. Menghitung jumlah kelas dari tetangga terdekat dan menetapkan kelas tersebut sebagai kelas data yang dievaluasi.

Pengujian

Pengujian dilakukan menggunakan Confusion Matrix dengan cara memberikan gambaran tentang keputusan yang dihasilkan selama proses pelatihan dan pengujian, serta memberikan estimasi kinerja klasifikasi berdasarkan kebenaran atau ketidakbenaran objek. Confusix adalah sebuah tabel yang memperlihatkan klasifikasi data uji yang benar dan yang salah. Sebagai contoh, confusix digunakan dalam klasifikasi biner [12].

Precision merupakan perhitungan terhadap perkiraan proporsi kasus positif yang benar dan dirumuskan dalam persamaan 3.

$$precision = \frac{TrueP}{TrueP + FalseP} \quad (3)$$

Hasil precision merupakan jumlah data yang benar dan tepat (true positive) dibagi dengan data yang benar dan tepat dan data hasil tidak terduga (true positive) dan false positive).

Recall merupakan perhitungan terhadap perkiraan proporsi kasus positif yang diidentifikasi benar dan dirumuskan dalam persamaan 4.

$$recall = \frac{TrueP}{TrueP + FalseN} \quad (4)$$

Recall merupakan jumlah data yg benar dan tepat(true positive) dibagi dengan data yang benar dan tepat dan data hasil hilang (true positive dan false negative).

Accuracy merupakan perhitungan terhadap proporsi dari jumlah total prediksi yang benar dan dirumuskan dalam persamaan 5.

$$\text{Accuracy} = \frac{TrueP+TrueN}{TrueP+FalseN+FalseP+TrueN} \quad (5)$$

Hasil accuracy klasifikasi merupakan jumlah data yang tepat (true positive dan true negative) dibagi dengan total data.

Keterangan:

a. TrueP (True Positive) merupakan banyaknya data di kelas aktualnya positif dan kelas prediksinya juga positif.
b. FalseN (False Negative) merupakan banyaknya data di kelas aktualnya positif sedangkan kelas prediksinya negatif.
c. FalseP (False Positive) merupakan banyaknya data di kelas aktualnya adalah kelas negatif sedangkan kelas prediksinya positif.
d. TrueN (True Negative) merupakan banyaknya data di kelas aktualnya adalah kelas negatif dan kelas prediksinya juga negatif.

3. HASIL DAN PEMBAHASAN

Hasil Pra-Pemrosesan Data

Encoding Data

Pada proses ini, variabel kategorikal dalam dataset diubah menjadi bentuk numerik untuk mempersiapkannya dalam analisis. Konversi ini meningkatkan kesiapan dataset untuk diproses oleh metode analisis seperti K-NN dengan lebih efisien. Hasil dari proses ini dapat dilihat pada tabel 2.

Tabel 2 Hasil Encocing Data

ATRIBUT	NILAI	*Hasil Encoding*
Grade	*High*	3
	Medium	2
	Low	1

Tabel 2 menunjukkan hasil encoding atribut grade yang awalnya merupakan variabel kategorikal (high, medium, low) diubah menjadi bentuk numerik: high menjadi 3, medium menjadi 2, dan low menjadi 1.

Proses encoding ini sangat krusial dalam pengolahan data karena algoritma pembelajaran mesin umumnya bekerja lebih optimal dengan data numerik dibandingkan dengan data kategorikal. Data numerik memungkinkan algoritma untuk memanfaatkan informasi dengan lebih baik selama proses analisis. Sebagai contoh, dalam metode K-NN, penggunaan nilai numerik memungkinkan perhitungan jarak antar data dilakukan dengan lebih akurat. Hal ini karena K-NN mengandalkan perhitungan jarak antar titik data untuk menentukan klasifikasi atau prediksi. Dengan konversi atribut 'Grade' menjadi angka numerik, model K-NN dapat menghitung jarak antar titik data dengan lebih tepat, yang pada gilirannya meningkatkan akurasi prediksi yang dihasilkan.

Selain itu, encoding data ini juga membantu dalam menghindari potensi kesalahan interpretasi oleh algoritma. Data kategorikal dalam bentuk teks dapat menimbulkan ambigu bagi algoritma yang tidak dirancang untuk menangani tipe data tersebut secara langsung. Dengan mengubahnya menjadi format numerik, kita memastikan bahwa informasi dari variabel kategorikal dapat diproses dan dianalisis secara maksimal oleh berbagai algoritma pembelajaran mesin, tidak hanya K-NN.

Secara keseluruhan, proses encoding ini merupakan langkah awal yang penting dalam pra-pemrosesan data yang memastikan bahwa dataset siap untuk analisis lebih lanjut dan dapat memberikan hasil yang lebih akurat dan reliabel dalam model pembelajaran mesin yang digunakan.

Pembersihan Data

Selanjutnya, dalam tahap pembersihan data, dataset awal yang terdiri dari 100 data. Selama proses pembersihan data, terdeteksi bahwa 17 data memiliki nilai yang hilang, yang mempengaruhi integritas data secara keseluruhan. Setelah mengidentifikasi dan menghapus data yang tidak lengkap ini, jumlah total data yang tersisa menjadi 83 data. Proses pembersihan ini sangat penting karena memastikan bahwa dataset yang digunakan untuk analisis lebih lanjut adalah relevan dan valid

Pemisahan Data

Pada proses pemisahan data, dataset dibagi menjadi dua bagian: data latih dan data uji. Penelitian ini mencakup pengujian dengan tiga skema yang ditunjukkan dalam Tabel 3.

Tabel 3 Skema Pembagian Dataset

Skema	Data Latih	Data Uji
1	80% (66 data)	20% (17 data)
2	70% (58 data)	30% (25 data)
3	60% (49 data)	40% (34 data)

Dalam Tabel 3, kita dapat melihat tiga skema pembagian dataset yang berbeda. Skema pertama membagi data menjadi 80% untuk data latih dan 20% untuk data uji, dengan total 66 data latih dan 17 data uji. Skema kedua menggunakan 70% data latih dan 30% data uji, menghasilkan 58 data latih dan 25 data uji. Skema ketiga membagi data menjadi 60% data latih dan 40% data uji, dengan 49 data latih dan 34 data uji.

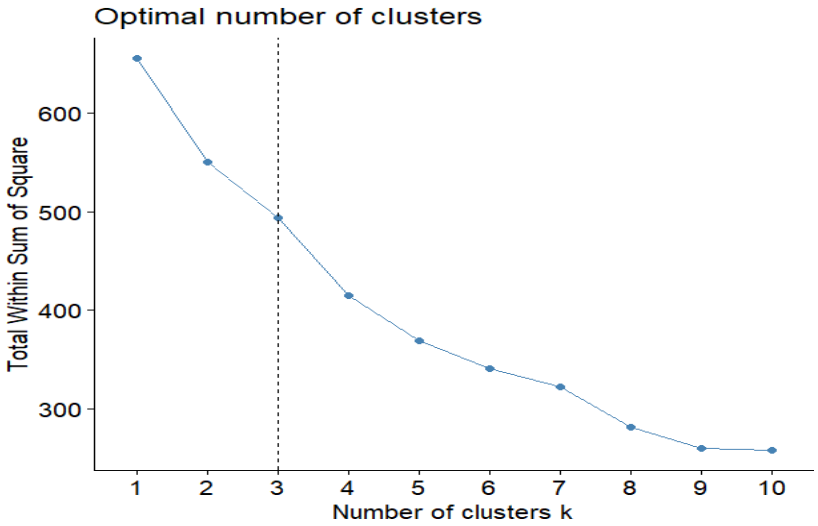
Melakukan pra-pemrosesan data secara teliti sangat penting untuk memastikan bahwa data yang digunakan dalam algoritma K-NN bersih, terstruktur, dan siap digunakan. Proses pra-pemrosesan ini mencakup pembersihan data dari kesalahan dan inkonsistensi, serta penyusunan data agar sesuai dengan kebutuhan analisis. Tindakan ini tidak hanya meningkatkan kualitas data secara keseluruhan, tetapi juga berkontribusi secara signifikan terhadap kualitas analisis yang dilakukan oleh model.

Data yang bersih dan terstruktur memungkinkan model K-NN untuk mengenali pola dan hubungan yang ada dalam data dengan lebih efektif. Hal ini berdampak langsung pada akurasi prediksi yang dihasilkan oleh model. Prediksi yang akurat dan andal sangat penting dalam konteks analisis data, karena memberikan dasar yang kuat untuk pengambilan keputusan yang berbasis data. Selain itu, data yang siap digunakan juga mengurangi risiko terjadinya kesalahan selama proses analisis, yang pada gilirannya meningkatkan kepercayaan terhadap hasil akhir yang diperoleh dari model.

Dengan demikian, pemisahan data yang tepat dan pra-pemrosesan data yang teliti adalah langkah-langkah krusial dalam proses analisis data yang memastikan hasil prediksi yang dihasilkan oleh model K-NN menjadi lebih akurat dan dapat diandalkan.

Penentuan K-Indeks

Penentuan jumlah klaster optimal (K) dalam sebuah dataset dapat dilakukan dengan menggunakan metode elbow. Metode ini membantu mengidentifikasi titik optimal di mana penambahan jumlah klaster tidak lagi memberikan penurunan yang signifikan pada WSS (Within-Cluster Sum of Squares). Hasil dari optimasi dengan metode elbow menunjukkan bahwa nilai optimum untuk K adalah 3. Hal ini ditunjukkan oleh titik di mana penambahan klaster setelah K=3 tidak menghasilkan penurunan WSS yang berarti. Hasil optimasi metode elbow ini dapat dilihat pada Gambar 3.



Gambar 3 Grafik Optimasi Metode Elbow

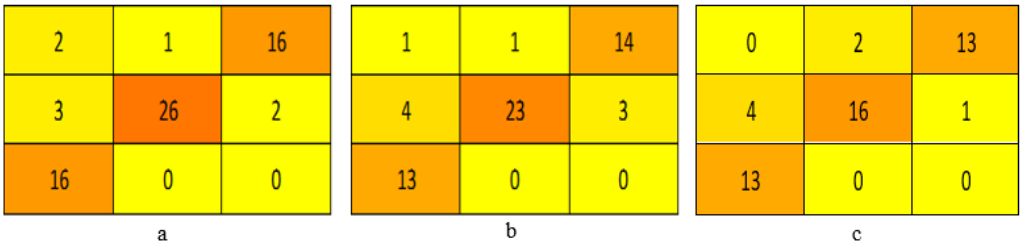
Gambar 3 memperlihatkan hasil optimasi metode elbow, di mana terlihat jelas bahwa penurunan WSS mulai melambat setelah K=3. Grafik tersebut membantu dalam memahami alasan pemilihan K=3 sebagai jumlah klaster optimal. Pada titik ini, penambahan klaster lebih lanjut tidak memberikan manfaat signifikan dalam mengurangi WSS, sehingga K=3 dianggap sebagai jumlah klaster yang lebih efisien dan efektif untuk analisis lebih lanjut pada dataset ini. Metode elbow digunakan untuk menentukan jumlah klaster optimal dengan cara memplot WSS terhadap jumlah klaster yang berbeda. Awalnya, penambahan jumlah klaster akan menyebabkan penurunan WSS yang signifikan karena setiap klaster tambahan mampu menangkap variasi data dengan lebih baik. Namun, setelah mencapai titik tertentu, penambahan klaster berikutnya tidak memberikan penurunan WSS yang berarti, menciptakan bentuk seperti siku (elbow) pada grafik. Titik ini adalah jumlah klaster optimal yang diidentifikasi oleh metode elbow.

Dalam kasus ini, setelah dilakukan optimasi dengan metode elbow, K=3 ditemukan sebagai jumlah klaster optimal. Grafik pada Gambar 3 menunjukkan bahwa penurunan WSS signifikan terjadi hingga K=3, dan setelah itu, laju penurunan melambat secara drastis. Hal ini menunjukkan bahwa penambahan klaster setelah K=3 tidak meningkatkan kualitas klasterisasi secara signifikan. Dengan demikian, metode elbow memberikan dasar yang kuat untuk menetapkan K=3 sebagai jumlah klaster yang paling efisien dan efektif untuk analisis lebih lanjut pada dataset ini. Keputusan ini memastikan bahwa model klasterisasi yang digunakan akan memiliki keseimbangan yang baik antara kompleksitas dan efektivitas dalam menangkap struktur data yang ada.

Klasifikasi

Klasifikasi menggunakan metode K-NN adalah tahap penting dalam evaluasi model, yang bertujuan untuk menentukan dengan tepat jumlah sampel yang terklasifikasi sebagai True Positive (TP), mencerminkan prediksi yang akurat terhadap kelas positive. Proses ini juga bertujuan untuk memastikan bahwa True Negative (TN) dikenali secara optimal, menunjukkan kemampuan model dalam mengidentifikasi sampel yang sebenarnya negatif.

Selain itu, penekanan juga diberikan pada pengenalan kasus False Positive (FP), di mana model memberikan klasifikasi yang tidak tepat terhadap sampel negative, dan False Negative (FN), yang menunjukkan kekurangan dalam model dalam mengenali sampel positive, mengisyaratkan adanya batasan dalam pemahaman model terhadap dataset yang digunakan. Oleh karena itu, evaluasi kinerja model dapat dilakukan secara komprehensif, mendukung accuracy dan ketepatan prediksi model.

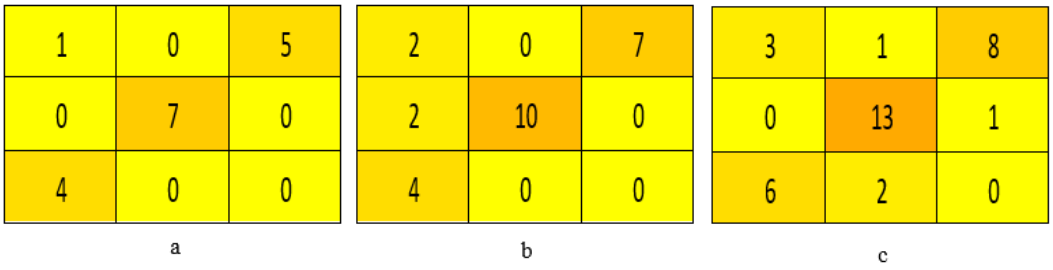


Gambar 4 *Confusix* Data Latih (a) Skema 1, (b) Skema 2, (c) Skema 3

Pada Gambar 4 (a) Skema 1 yang mana data latih terdiri dari 66 sampel. Hasil confusion matrix menunjukkan 16 True Negative (TN), 3 False Positive (FP), 26 True Positive (TP), dan 0 False Negative (FN). Dengan 16 TN, model mampu mengenali sampel negative dengan baik, sementara 26 TP menunjukkan model mengklasifikasikan sampel positive dengan accuracy tinggi. Adanya 3 FP berarti ada beberapa sampel negative yang salah diklasifikasikan sebagai positive, namun jumlah ini relatif kecil dibandingkan total data. Tidak adanya FN (0) mengindikasikan bahwa model berhasil mengidentifikasi semua sampel positive dengan benar pada data latih ini.

Lalu pada Gambar 4 (b) Skema 2 dalam skema ini, data latih terdiri dari 58 sampel. Hasil confusix menunjukkan 13 TN, 3 FP, 23 TP, dan 0 FN. Meski jumlah data latih lebih sedikit, model tetap menunjukkan kinerja yang baik tanpa adanya FN, yang berarti semua sampel positive dikenali dengan benar. Jumlah FP tetap sama dengan skema 1, menunjukkan bahwa model memiliki konsistensi dalam kesalahan klasifikasi terhadap sampel negative. Dengan 13 TN dan 23 TP, model menunjukkan performa yang stabil dalam mengenali sampel negative dan positive.

Dan terakhir pada Gambar 4 (c) Skema 3 Dalam skema ini, data latih terdiri dari 49 sampel. Hasil confusix menunjukkan 12 TN, 3 FP, 16 TP, dan 0 FN. Meskipun data latih lebih sedikit dibandingkan skema sebelumnya, model tetap konsisten tanpa adanya FN, yang berarti tidak ada sampel positive yang salah diklasifikasikan sebagai negative. Jumlah FP tetap sama, menunjukkan bahwa kesalahan dalam mengklasifikasikan sampel negative masih konsisten. Dengan 12 TN dan 16 TP, model tetap efektif dalam mengenali sampel negative dan positive.



Gambar 5 Confusix Data Uji (a) Skema 1, (b) Skema 2, (c) Skema 3

Pada Gambar 5 (a) Skema 1 Dalam skema ini, data uji terdiri dari 17 sampel. Hasil confusix menunjukkan 4 TP, 7 TN, 0 FP, dan 6 FN. Ini berarti model mampu mengklasifikasikan 4 sampel positive dengan benar dan 7 sampel negative dengan benar. Tidak adanya FP (0) mengindikasikan bahwa model sangat akurat dalam mengidentifikasi sampel negative. Namun, adanya 6 FN menunjukkan bahwa model masih memiliki kelemahan dalam mengenali sampel positive, di mana 6 sampel positive salah diklasifikasikan sebagai negative.

Pada Gambar 5 (b) Skema 2 yang mana skema ini, data uji terdiri dari 25 sampel. Hasil confusix menunjukkan 4 TP, 10 TN, 2 FP, dan 9 FN. Meskipun ada peningkatan jumlah data uji, model tetap menunjukkan kemampuan yang baik dalam mengklasifikasikan sampel

negative dengan 10 TN. Namun, adanya 2 FP menunjukkan bahwa ada beberapa sampel negative yang salah diklasifikasikan sebagai positive. Adanya 9 FN mengindikasikan bahwa model masih memiliki kesulitan dalam mengenali sampel positive, di mana lebih banyak sampel positive yang salah diklasifikasikan sebagai negative dibandingkan skema 1.

Dan terakhir Gambar 5 (c) Skema 3 dapat dilihat pada skema ini, data uji terdiri dari 34 sampel. Hasil confusix menunjukkan 6 TP, 13 TN, 0 FP, dan 15 FN. Meski tidak ada FP (0), yang berarti tidak ada sampel negatif yang salah diklasifikasikan sebagai positive, adanya 15 FN menunjukkan bahwa model memiliki kesulitan besar dalam mengenali sampel positive. Meskipun model berhasil mengklasifikasikan 6 sampel positif dan 13 sampel negative dengan benar, jumlah FN yang tinggi mengindikasikan bahwa banyak sampel positive yang salah diklasifikasikan sebagai negative.

Evaluasi menggunakan confusix memberikan gambaran yang mendetail tentang kinerja model K-NN dalam mengklasifikasikan sampel data. Dalam ketiga skema untuk data latih, model menunjukkan konsistensi tanpa adanya False Negatives (FN = 0), menunjukkan bahwa model sangat baik dalam mengenali sampel positive. Namun, adanya beberapa False Positives (FP) menunjukkan bahwa masih ada beberapa kesalahan dalam mengklasifikasikan sampel negative.

Untuk data uji, performa model bervariasi. Pada skema 1, model menunjukkan kinerja yang baik dengan tidak adanya FP dan jumlah FN yang cukup rendah. Namun, pada skema 2 dan 3, jumlah FN meningkat, menunjukkan bahwa model memiliki kelemahan dalam mengenali sampel positive pada data uji yang lebih besar. Meskipun tidak ada FP pada skema 3, jumlah FN yang tinggi menunjukkan bahwa model mungkin memerlukan penyesuaian lebih lanjut untuk meningkatkan accuracy.

Secara keseluruhan, analisis confusix ini menunjukkan bahwa model K-NN memiliki potensi yang baik dalam klasifikasi sampel positive dan negative, namun masih ada ruang untuk perbaikan, terutama dalam mengurangi jumlah False Negatives pada data uji. Evaluasi ini sangat penting untuk memastikan bahwa model yang digunakan dapat memberikan prediksi yang akurat dan dapat diandalkan dalam berbagai skenario analisis data.

Pengujian

Dalam proses pengujian analisis kualitas susu menggunakan metode K-NN, dilakukan tiga skema pengujian yang berbeda dengan memanfaatkan 83 data setelah melalui tahap pra-pemrosesan data. Nilai optimal K yang ditentukan dengan metode elbow adalah 3. Dan hasil Pengujian dapat dilihat pada Tabel 4

Tabel 4 Hasil Pengujian Metode metode

Skema	*Split Data*	*Accuracy*	*Precision*	*Recall*
1	80:20	0,94	1	0.8
2	70:30	0,84	1	0.5
3	60:40	0,79	0.75	0.6666667

Hasil pengujian menunjukkan bahwa skema 1 dengan pembagian data 80:20 merupakan pilihan terbaik dalam penelitian ini. Dalam skema ini, di mana 80% dari data digunakan sebagai data latih dan 20% sebagai data uji, model K-NN mencapai tingkat akurasi tertinggi, yaitu 0,94. Ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam memprediksi kelas-kelas yang benar. Selain itu, nilai precision yang sempurna (1) menunjukkan bahwa model hampir tidak menghasilkan hasil positif palsu, sementara nilai recall yang tinggi, 0,8, menandakan kemampuan model dalam mendeteksi sebagian besar kelas positive dengan efektif.

Analisis dari penelitian sebelumnya menggunakan penerapan Machine Learning dalam memeriksa kualitas susu, terutama melalui metode K-NN, Random Forest, dan Neural Network.

Jurnal pertama menekankan penggunaan K-NN untuk mengklasifikasikan susu pasteurisasi dengan tingkat akurasi yang tinggi, meskipun terdapat kekurangan dalam optimasi nilai K dan variasi dalam pembagian dataset. Jurnal kedua juga menggunakan K-NN tetapi dengan tingkat akurasi yang rendah, tanpa optimasi nilai K dan variasi dataset. Jurnal ketiga memperbaiki kekurangan sebelumnya dengan menggunakan Backpropagation dan K-fold Cross Validation, meskipun tidak menggunakan K-NN secara khusus.

Namun demikian, hasil pengujian pada penelitian ini menunjukkan bahwa dalam analisis kualitas susu menggunakan metode K-NN, dengan pembagian data 80:20 menghasilkan tingkat accuracy tertinggi, mencapai 0,94. Model ini berhasil memprediksi kelas dengan sangat baik, yang ditunjukkan oleh nilai precision yang tinggi dan nilai recall yang memuaskan.

Hasil penelitian ini menegaskan bahwa dengan menggabungkan pendekatan-pendekatan ini, pemahaman tentang kinerja model K-NN dalam menganalisis kualitas susu dapat ditingkatkan secara signifikan. Hal ini memberikan dasar yang lebih kuat untuk pengambilan keputusan yang lebih baik dalam menggunakan metode K-NN untuk analisis kualitas susu. Dengan demikian, penelitian ini memberikan kontribusi penting dalam bidang analisis kualitas susu menggunakan pembelajaran mesin, khususnya dengan metode K-NN, dan menyoroti pentingnya pra-pemrosesan data dan optimasi parameter dalam meningkatkan performa model.

Penelitian ini juga menggarisbawahi pentingnya melakukan pembagian dataset yang optimal dan penggunaan metode optimasi seperti elbow untuk menentukan nilai K yang terbaik. Dengan demikian, penelitian ini tidak hanya menunjukkan keefektifan metode K-NN dalam analisis kualitas susu tetapi juga memberikan panduan praktis bagi penelitian-penelitian selanjutnya dalam bidang yang sama. Penggunaan pembagian data 80:20 terbukti memberikan hasil yang paling baik dalam konteks penelitian ini, menyoroti pentingnya memilih proporsi data latih dan uji yang sesuai untuk mendapatkan hasil prediksi yang paling akurat.

4. KESIMPULAN

Hasil Pengujian metode K-NN pada dataset kualitas susu dengan berbagai skema pembagian data (80-20, 70-30, dan 60-40) telah menunjukkan hasil yang menjanjikan. Penggunaan metode elbow untuk menentukan jumlah klaster optimal (K) menghasilkan nilai K=3, yang konsisten dengan struktur data yang diamati. Dalam skema pembagian data 80-20, model K-NN mencapai kinerja terbaik dengan accuracy mencapai 0,9, recall sebesar 0,8, dan precison mencapai 1,0.

Hasil ini menunjukkan bahwa metode K-NN secara keseluruhan mampu memberikan prediksi kualitas susu yang akurat dan dapat diandalkan. Kinerja yang baik ini mencerminkan kemampuan model K-NN dalam mengenali pola dan struktur dalam dataset, sehingga mampu melakukan klasifikasi dengan tingkat kesalahan yang rendah. Dengan demikian, metode K-NN dapat dianggap sebagai alat yang efektif untuk mengevaluasi kualitas susu berdasarkan data yang tersedia.

5. SARAN

Meskipun metode K-NN telah memberikan hasil yang memuaskan, ada beberapa langkah yang dapat diambil untuk memperkuat hasil prediksi. Pertama, dengan mempertimbangkan kurangnya kuantitas data susu sebagai batasan dalam penelitian, penambahan kuantitas data dapat meningkatkan generalisasi model. Lebih banyak data dapat membantu dalam melatih model untuk mengenali pola yang lebih kompleks dan representatif dari kualitas susu. Selain itu, untuk mendapatkan pemahaman yang lebih holistik tentang kinerja model, disarankan untuk melakukan perbandingan dengan metode lain seperti Decision Tree, Naive Bayes, atau Support Vector Machine. Perbandingan ini dapat memberikan wawasan tambahan tentang keunggulan dan kelemahan masing-masing pendekatan dalam menganalisis

kualitas susu, serta membantu peneliti memilih pendekatan yang paling sesuai dengan tujuan dan kebutuhan penelitian.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada bapak Dr.Rendra Gustriansyah., M.Kom, bapak Nazori Suhandi., M.M dan teman-teman atas dukungan dan bimbingan yang mereka berikan selama proses penelitian dan penyusunan paper ini.

DAFTAR PUSTAKA

[1] G. Suranto, "-bsn-bijak-memilih-susu-gunakan-sni-sebagai-acuan @ bsn.go.id." Badan Standarisasi Nasional, Jakarta, 2021.

[2] D. Fitriati and M. Fahrudin, "Perangkingan Jenis Susu Untuk Balita Non-Asi Dengan Metode Simple Additive Weighting (Saw)," *J. Teknol. Terpadu*, vol. 5, no. 1, 2019, doi: 10.54914/jtt.v5i1.188.

[3] N. Wakhidah and S. N. Rochmah, "Klasifikasi kualitas mutu susu pasteurisasi menggunakan metode klasifikasi k-Nearest Neighbor," *AITI J. Teknol. Inf.*, vol. 21, no. 1, pp. 58–71, 2024, doi: https://doi.org/10.24246/aiti.v21i1.58-71..

[4] P. Sadeghi Vasafi and B. Hitzmann, "Comparison of various classification techniques for supervision of milk processing," *Eng. Life Sci.*, vol. 22, no. 3–4, pp. 279–287, 2022, doi: 10.1002/elsc.202100098.

[5] F. Adiba, A. A. Nur Risal, and M. Tahir, "Implementasi Algoritma Backpropagation untuk Klasifikasi Kualitasi Susu Sapi," *J. Mediat.*, vol. 6, no. 2, p. 42, 2023, doi: 10.26858/jmtik.v6i2.46013.

[6] Nikmatun, I. Alvi, Waspada, and Indra, "Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.

[7] A. Budianto, R. Ariyuana, and D. Maryono, "Perbandingan K-Nearest Neighbor (K-NN) Dan Support Vector Machine (Svm) Dalam Pengenalan Karakter Plat Kendaraan Bermotor," *J. Ilm. Pendidik. Tek. dan Kejuru.*, vol. 11, no. 1, p. 27, 2019, doi: 10.20961/jiptek.v11i1.18018.

[8] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[9] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis based on k-means," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 1, pp. 470–477, 2019, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[10] I. N. Abrar and A. Abdullah, "Klasifikasi Penyakit Liver Menggunakan Metode Elbow Untuk Menentukan K Optimal pada Algoritma K-Nearest Neighbor (K-NN)," *J. SISFOKOM (Sistem Inf. dan Komputer)*, vol. 12, no. 2, pp. 218–228, 2023, doi: 10.32736/sisfokom.v12i2.1643.

[11] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," Indones. J. Data Sci., vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[12] M. León and D. Ossa, "Machine Learning Applied to Milk Sample Classification," *Proc. First Aust. Int. Conf. Ind. Eng. Oper. Manag.*, pp. 2390–2398, 2022, doi: 10.46254/au01.20220511.