INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 1 Tahun 2025 Page 422-435
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Seleksi Fitur Information Gain Untuk Klasifikasi Kualitas Susu Sapi Menggunakan Metode K-Nearest Neighbor Dan Naïve Bayes

Fauzi Wardah Ali[1✉], I Wayan Sumarjaya[2], Eka N. Kencana[3]
Program Studi Matematika, Universitas Udayana
Email: fauziw2713@gmail.com[1✉]

Abstrak

Kemajuan dalam komputasi modern, khususnya klasifikasi, telah membantu manusia dalam mengklasifikasikan berbagai tugas yang memakan waktu dan komputasi yang mahal, salah satunya adalah klasifikasi kualitas susu sapi. Pengklasifikasian ini penting dilakukan untuk mengurangi kemungkinan beredarnya susu dengan kualitas buruk di masyarakat. Data pada penelitian ini terdiri dari 429 susu kualitas rendah, 374 susu kualitas menengah, dan 256 susu kualitas tinggi. Penelitian ini menguji performa algoritma klarifikasi K-nearest neighbor dan naïve Bayes dengan menggunakan teknik seleksi fitur information gain dan tanpa seleksi fitur information gain. Hasil penelitian ini menunjukkan performa KNN mengalami peningkatan rata-rata akurasi sebesar 1,04 persen dengan perhitungan jarak Euclid dan 0,92 persen dengan perhitungan jarak Manhattan ketika menggunakan lima fitur. Sedangkan performa naïve Bayes mengalami penurunan akurasi sebesar 3,48 persen. Perlakuan yang berbeda tersebut memiliki perbedaan akurasi yang signifikan.
Kata Kunci: *kualitas susu sapi, KNN, naive Bayes, information gain.*

Abstract

Recent advances in modern computation, especially classification, have helped human to classify various task which were time consuming and compputationally-expensive, one such task is classification cow's quality milk. The classification is important as away to reduce the possibility of poor quality milk circulating in the community. The data in this study consists of 429 low quality milk, 374 medium quality milk, and 256 high quality milk. This study tested the performance of K nearest neighbor (KNN) and naïve Bayes clarification algorithms using information gain feature selection techniques and without information gain feature selection. The results of this research show that KNN performance have an average increase in accuracy of 1.04 percent when calculating the Euclid distance and 0.92 percent when calculating the Manhattan distance when using five features. Meanwhile, the performance of naïve Bayes have a decrease in accuracy of 3.48 percent. These different treatments have significant differences in accuracy.
Keyword: *cows milk quality, KNN, naive Bayes, information gain.*

PENDAHULUAN

Perkembangan komputer pada zaman sekarang terutama dalam komputasi untuk melakukan klasifikasi memungkinkan dapat mempermudah peran manusia untuk mengklasifikasikan berbagai hal, salah satunya dapat membantu untuk mengklasifikasikan kualitas susu sapi sehingga dapat mengurangi kemungkinan susu dengan kualitas buruk beredar di masyarakat. Klasifikasi adalah bentuk analisis data yang mengekstraksi model untuk menempatkan objek ke dalam kelas atau label tertentu (Han et al., 2012). Ada beberapa metode klasifikasi yaitu algoritma C4.5, *neural network*, *naïve Bayes*, *logistic regression* dan *K-nearest neighbor* (KNN).

Susu sapi adalah salah satu bahan pangan yang dianjurkan untuk dikonsumsi setiap hari karena kandungan nutrisinya memiliki banyak manfaat untuk tubuh manusia. Susu sapi mengandung banyak nutrisi, seperti protein, lemak, vitamin, kalsium, mineral, serta asam amino esensial dan non-esensial (Multamiah et al., 2013). Kandungan nutrisi tersebut sangat berguna untuk tumbuh kembang anak yang salah satunya dapat membantu mengembangkan kecerdasan anak, karena persentase penyerapan susu dalam tubuh sebesar 98 persen sampai 100 persen (Hidayat et al., 2016).

Nutrisi yang terkandung di dalam susu sapi dapat dipengaruhi oleh kualitasnya. Susu dapat menyebabkan gangguan kesehatan manusia apabila susu yang dikonsumsi memiliki kualitas yang buruk, seperti dapat menimbulkan penyakit pencernaan (Wiranti et al., 2022). Oleh karena itu, pengujian terhadap kualitas susu harus dilakukan untuk menghindari kualitas susu yang buruk dikonsumsi oleh manusia. Syarat susu segar berdasarkan Standar Nasional Indonesia (SNI) 3141-1: 2011 yang dilakukan dengan salah satu uji organoleptik adalah konsistensi susu 2,75 (agak encer), bau susu segar 5,63 (khas susu) dan tidak mengalami perubahan bau, dan warna susu 4,63 (putih kekuningan) (Wasito, 2017).

K-nearest neighbor (KNN) merupakan salah satu metode klasifikasi yang melabeli objek baru berdasarkan tetangga terdekat objek baru tersebut (Gorunescu, 2011). Menurut Bhatia & Vandana (2010) kelebihan dari algoritma KNN yaitu algoritma ini efektif jika menggunakan data yang jumlahnya besar dan pelatihannya cepat. *Naïve* Bayes merupakan salah satu algoritma klasifikasi probabilitas sederhana yang menentukan probabilitas dengan menganalisis frekuensi dalam sekumpulan data. Menurut Syarli & Muin (2016) kelebihan dari algoritma *naïve* Bayes yaitu metode ini mudah diimplementasikan dan memberikan hasil yang baik untuk banyak kasus.

Sari (2016) menggunakan teknik seleksi fitur *information gain* pada algoritma *machine learning* untuk prediksi performa akademik siswa dengan menggunakan algoritma klasifikasi *decision tree*, *random forest*, *neural network*, SVM, dan *naïve* Bayes. Hasil penelitiannya

menunjukkan adanya peningkatan akurasi klasifikasi dari penggunaan teknik seleksi fitur *information gain*. Implementasi seleksi fitur *information gain* juga dilakukan oleh Dewantoro et al (2019) pada *word sense disambiguation* Bahasa Indonesia yang menggunakan teknik klasifikasi *decision list*. Berdasarkan penelitiannya diperoleh peningkatan akurasi dan presisi pada teknik klasifikasi *decision list* yang menggunakan teknik seleksi fitur *information gain*. Penelitian oleh Mutmainnah et al (2019) tentang pengaruh seleksi fitur *information gain* pada algoritma *K-nearest neighbor* untuk klasifikasi tingkat kelancaran pembayaran kredit kendaraan menunjukkan seleksi fitur *information gain* dapat meningkatkan akurasi algoritma KNN.

Performa dari suatu metode klasifikasi dapat dipengaruhi oleh jumlah atribut dari objek yang akan diklasifikasi. Jumlah atribut yang banyak dapat membebani kinerja algoritma klasifikasi sehingga dapat menurunkan akurasi (Hafidzullah et al., 2019). Dengan demikian, diperlukan seleksi fitur untuk mendapatkan atribut yang optimal sehingga dapat meningkatkan efisiensi proses klasifikasi dan memperbaiki kinerja dari suatu algoritma klasifikasi (Nabella et al., 2019). *Information gain* merupakan salah satu teknik seleksi fitur yang cara kerjanya dengan mengurutkan atribut-atribut yang relevan atau atribut yang memiliki paling banyak informasi dari kelas tertentu (Aini et al., 2018).

METODE PENELITIAN

Penelitian ini menggunakan jenis data sekunder. Data yang digunakan merupakan dataset kualitas susu sapi yang datanya dikumpulkan di India dan diunggah oleh Shrijayan (2022) di situs Kaggle. Variabel yang digunakan pada penelitian adalah pH, temperatur, rasa, bau, *fat,* kekentalan, dan warna.

Pada penelitian ini, pembuatan model algoritma klasifikasi menggunakan bantuan bahasa pemrograman Python dengan menggunakan *library scikit-learn*. Langkah – langkah yang akan dilakukan dalam penelitian ini sebagai berikut.

1. Melakukan seleksi fitur menggunakan metode *information gain*. Menurut Shaltout et al (2014) *information gain* mampu mendeteksi fitur yang memiliki informasi terbanyak berdasarkan kelas tertentu. Tahapan seleksi fitur *information gain* sebagai berikut.
   i. Menghitung nilai *entropy* dari masing-masing fitur. *Entropy* merupakan suatu ukuran yang digunakan untuk mengukur ketidakpastian. Teknik seleksi fitur information gain bertujuan untuk mengurangi ketidakpastian atribut tertentu. Oleh karena itu, fitur dengan nilai *information gain* yang besar memiliki informasi yang banyak untuk melakukan klasifikasi (Leung et al., 2011). Perhitungan *entropy* menggunakan persamaan

$$E(S) = \sum_{i}^{n} -P_i \log_2 P_i \quad (1)$$

Nilai n menunjukkan jumlah kelas dan P_i menunjukkan rasio sampel i terhadap total kelas.

ii. Menghitung nilai *information gain* setiap fitur menggunakan persamaan

$$Gain(S, A) = E(S) - \sum_{v \in value\ A} -\frac{S_v}{S} E(S_v) \quad (2)$$

Dengan A menyatakan atribut, $E(S)$ menyatakan nilai entropy atribut, v menyatakan nilai yang ada pada atribut A, S_v jumlah sampel untuk nilai v, S menyatakan jumlah seluruh sampel, data dan $E(S_v)$ menyatakan nilai entropy untuk sampel yang memiliki nilai v. Fitur diurutkan berdasarkan nilai *information gain* tertinggi, dan fitur yang digunakan untuk klasifikasi adalah fitur dengan urutan teratas (Chormunge & Jena, 2016).

2. Melakukan klasifikasi menggunakan algoritma *K-nearest neighbor* (KNN). Menurut Gorunescu (2011) algoritma KNN bekerja dengan menggunakan prinsip bahwa semua objek yang berada di dalam satu kelas yang sama cenderung memiliki jarak yang kecil. Tahapan algoritma KNN sebagai berikut

i. Menentukan nilai parameter k untuk menentukan jumlah tetangga objek. Menurut Larose & Larose (2014) tidak ada cara terbaik untuk memilih nilai k dalam klasifikasi KNN. Ketika memilih nilai k terlalu kecil maka algoritma KNN sangat dipengaruhi oleh pencilan. Oleh sebab itu, untuk mendapatkan nilai k terbaik dibutuhkan proses *cross validation* sehingga dapat mengurangi kemungkinan eror pada algoritma KNN.

ii. Menghitung jarak setiap objek data uji ke semua objek data latih. Ketika menghitung matriks jarak, terdapat atribut tertentu yang memiliki nilai yang tidak setara. Untuk mengatasi hal tersebut maka diperlukan normalisasi data (Larose & Larose, 2014). Normalisasi data dapat dilakukan menggunakan Z-*score standardization* yang perhitungannya sebagai berikut.

$$Z = \frac{X - \mu}{\sigma} \quad (3)$$

Nilai X menyatakan *observed value*, μ menyatakan nilai rata-rata, dan σ menyatakan nilai simpangan baku. Perhitungan jarak sangat tidak tepat jika atributnya memiliki jenis data kategorik. Sebagai gantinya perhitungan matriks jarak dengan atribut kategorik dapat dilakukan dengan mendefinisikan fungsi '*different from*" (Larose & Larose, 2014).

$$Different(x_1, y_1) = \begin{cases} 0 & if\ x_i = y_i \\ 1 & otherwise \end{cases} \quad (4)$$

Pada penelitian ini perhitungan matriks jarak menggunakan perhitungan jarak Euclid dan Manhattan. Perhitungan jarak Euclid menghitung selisih akar kuadrat antara pasangan titik data sedangkan perhitungan jarak Manhattan dihitung dengan menjumlahkan selisih nilai mutlak antara dua titik (Mulak & Talhar, 2015). Perhitungan jarak Euclid dan Manhattan ditunjukkan pada (5) dan (6)

$$d_i = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{6}$$

iii. Menentukan tetangga objek yang diklasifikasi dari nilai k yang telah ditentukan dan menentukan kelas objek berdasarkan mayoritas kelas tetangganya.

3. Melakukan klasifikasi menggunakan algoritma *naïve* Bayes. Algoritma *naïve* Bayes menggunakan teorema Bayes yang mengasumsikan semua atributnya independen (Patil & Sherekar, 2013). Tahapan algoritma *naïve* Bayes sebagai berikut.

i. Menghitung probabilitas bersyarat dengan menggunakan persamaan berikut

$$P(x_k|C_i) = \frac{P(x_k \cap C_i)}{P(C_i)} \tag{7}$$

Jika atribut bernilai kontinu maka perhitungan $P(x_i|C_i)$ menggunakan fungsi kepadatan peluang Gauss

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

Untuk menghindari perhitungan yang menghasilkan probabilitas nol, maka perlu menggunakan *Laplacian corection*, yaitu setiap pasangan kategori pada atribut mempunyai tambahan satu *tuple* (Han et al., 2012).

ii. Menghitung probabilitas *likelihood* $P(X|C_i)$. Perhitungan probabilitas *likelihood* menjadi komputasi yang mahal karena kumpulan dataset dengan atribut yang banyak. Oleh karena itu, untuk menyederhanakan perhitungan probabilitas likelihood dapat mengasumsikan atribut independen satu sama lain (Han et al., 2012). Sehingga perhitungan probabilitas *likelihood* dapat dilakukan dengan persamaan berikut

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{9}$$

iii. Menghitung probabilitas posterior $P(C_i|X)$ menggunakan persamaan

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}. \tag{10}$$

Algoritma *naive* Bayes mengklasifikasikan X sebagai kelas C_i berdasarkan nilai $P(C_i|X)$ tertinggi.

4. Evaluasi model menggunakan *confussion matrix* dan *k-fold cross validation* dengan menghitung akurasi menggunakan persamaan berikut

$$akurasi = \frac{TP+NP}{TP+NP+TF+NF} \quad (11)$$

HASIL DAN PEMBAHASAN

Seleksi Fitur *Information Gain*

Teknik seleksi fitur *information gain* diawali dengan menghitung nilai *entropy* total, kemudian mencari nilai *entropy* setiap atributnya.

Tabel 1 Nilai *Entropy* Atribut Rasa

Rasa	Rendah	Medium	Tinggi	*Entropy*
0	175	219	86	480
1	254	155	170	579
Total	429	374	256	1059

Dalam penelitian ini dataset yang digunakan memiliki tiga kelas, sehingga nilai *entropy* maksimum sebesar $log_2 3 = 1{,}58496$. Nilai $entropy_{rasa=0}$ sebesar $1{,}491684$ dan nilai $entropy_{rasa=1}$ sebesar $1{,}549576$ artinya atribut rasa dengan kategori 0 dan kategori 1 memiliki ketidakpastian yang besar dalam memprediksi kelas, karena nilai *entropy* mendekati nilai maksimumnya.

Tabel 2 Nilai *Information Gain* dari Dataset Kualitas Susu Sapi

Atribut	Nilai *Information Gain*
pH	0,811141
Temperatur	0,686365
Warna	0,320326
Kekentalan	0,234075
Lemak	0,223821
Bau	0,160863
Rasa	0,031713

Atribut pH memiliki nilai *information gain* terbesar dibandingkan enam atribut lainnya. Nilai maksimum *information gain* sama dengan nilai $entropy_{total}$. Pada penelitian ini nilai maksimum *information gain* adalah $1{,}553612$. Semakin mendekati nilai maksimumnya berarti atribut tersebut sangat membantu dalam memprediksikan kelas karena memiliki nilai ketidakpastian yang minim.

Klasifikasi *K-Nearest Neighbor*

Klasifikasi *K-Nearest Neighbor* diawali dengan mencari jarak antara data latih dan data uji. Perhitungan jarak yang dilakukan dalam penelitian ini menggunakan perhitungan jarak Euclid dan Manhattan.

Tabel 3 *Confusion Matrix* KNN Tanpa Seleksi Fitur

Kualitas	Prediksi rendah	Prediksi medium	Prediksi tinggi
Aktual rendah	86	0	0
Aktual medium	2	113	1
Aktual tinggi	0	1	115

Berdasarkan Tabel 4 ditunjukkan bahwa 88 susu yang diprediksi memiliki kualitas rendah pada kenyataannya terdapat 2 susu berkualitas menengah, dari 114 susu yang diprediksi menengah pada kenyataannya terdapat 1 susu berkualitas tinggi, dan dari 116 susu yang diprediksi memiliki kualitas tinggi pada kenyataannya terdapat 1 susu berkualitas medium. Dari tabel tersebut dapat diperoleh nilai akurasinya sebagai berikut.

$$accuracy = \frac{TP_{rendah}+TP_{medium}+TP_{tinggi}}{Total\ Sampel}$$
$$= \frac{86+113+115}{318} = 0{,}9874.$$

Model KNN tanpa seleksi fitur yang menggunakan perhitungan jarak Euclid dan nilai $k = 3$ memiliki tingkat akurasi sebesar 98,74 persen. Artinya 98,74 persen data uji yang kelasnya diprediksi dengan metode KNN sesuai dengan kelas sebenarnya

Tabel 4 Nilai Iterasi 10 *Fold Cross Validation*

Iterasi	Akurasi
Iterasi ke-1	0,96875
Iterasi ke-2	0,96875
Iterasi ke-3	0,90625
Iterasi ke-4	1,00000
Iterasi ke-5	0,93750
Iterasi ke-6	0,90625
Iterasi ke-7	1,00000
Iterasi ke-8	1,00000
Iterasi ke-9	0,93548
Iterasi ke-10	0,87096
Rata-rata	0,94935

Berdasarkan nilai akurasi pada sepuluh iterasi, didapatkan nilai simpangan baku sebesar 0,04551 dan nilai rata-ratanya sebesar 0,94935. Nilai standar deviasi dan rata-rata

tersebut menunjukkan bahwa variasi nilai setiap iterasinya berada pada kisaran 0,94935 ± 0,04551. Hal tersebut menunjukkan algoritma KNN memiliki kinerja yang stabil terhadap data baru karena memiliki variasi yang cenderung kecil.

Tabel 5 Akurasi KNN dengan Perhitungan Jarak Euclid dan Manhattan Tanpa Seleksi Fitur

k	Akurasi KNN dengan jarak Euclid	Akurasi KNN dengan jarak Manhattan
1	0,984274194	0,981149194
2	0,981048387	0,977923387
3	0,949395161	0,940020161
4	0,940020161	0,936895161
5	0,930645161	0.918044355
6	0,908669355	0,899294355
7	0,889919355	0,880544355
8	0,880443548	0,874193548
9	0,877318548	0,871068548
10	0,874193548	0,849092742
Rata-rata	0,9216	0,9128

Berdasarkan rata-rata akurasi, algoritma KNN dengan menggunakan perhitungan jarak Euclid memiliki performa yang lebih baik dibandingkan dengan menggunakan perhitungan jarak Manhattan. Selisih akurasi rata-ratanya sebesar 1,12 persen.

Klasifikasi Naive Bayes

Klasifikasi menggunakan algoritma *naïve* Bayes diawali dengan menghitung probabilitas setiap kelasnya $P(kualitas)$. Kemudian menghitung probabilitas bersyarat setiap atribut terhadap kelasnya $P(x_k|C_i)$ yang digunakan untuk mencari nilai probabilitas *likelihood* $P(X|C_i)$. Probabilitas *likelihood* dan probabilitas kelas C_i digunakan untuk mencari probabilitas posterior $P(C_i|X)$. Algoritma naïve Bayes mengklasifikasikan X adalah kelas C_i berdasarkan nilai $P(C_i|X)$ tertinggi.

Tabel 6 *Confusion Matrix Naïve* Bayes tanpa Seleksi Fitur

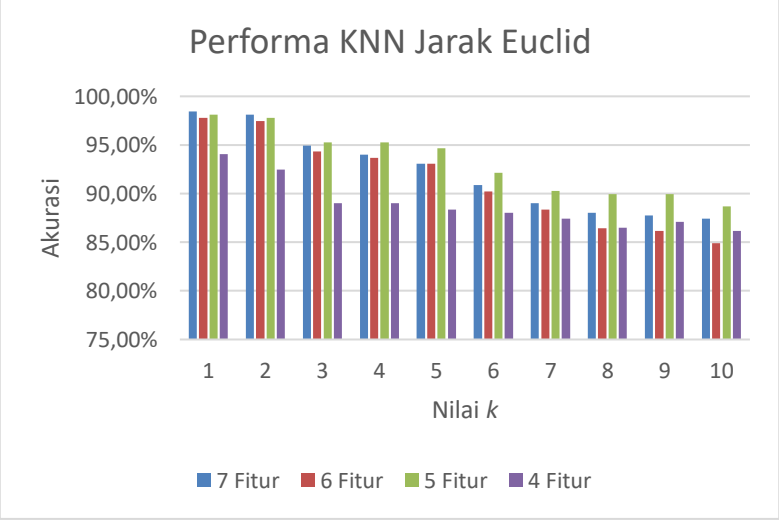
Kualitas	Prediksi rendah	Prediksi medium	Prediksi tinggi
Aktual rendah	85	1	0
Aktual medium	6	108	2
Aktual tinggi	52	0	64

Dari Tabel 6 menunjukkan bahwa 143 susu yang diprediksi memiliki kualitas rendah pada kenyataannya terdapat 6 susu dengan kualitas menengah dan 52 susu dengan kualitas tinggi, dari 109 susu yang diprediksi memiliki kualitas medium pada kenyataannya terdapat 1 susu yang memiliki kualitas rendah, dan dari 68 susu yang diprediksi memiliki kualitas tinggi pada kenyataannya terdapat 2 susu yang memiliki kualitas medium. Dari tabel tersebut dapat diperoleh nilai akurasinya sebagai berikut.

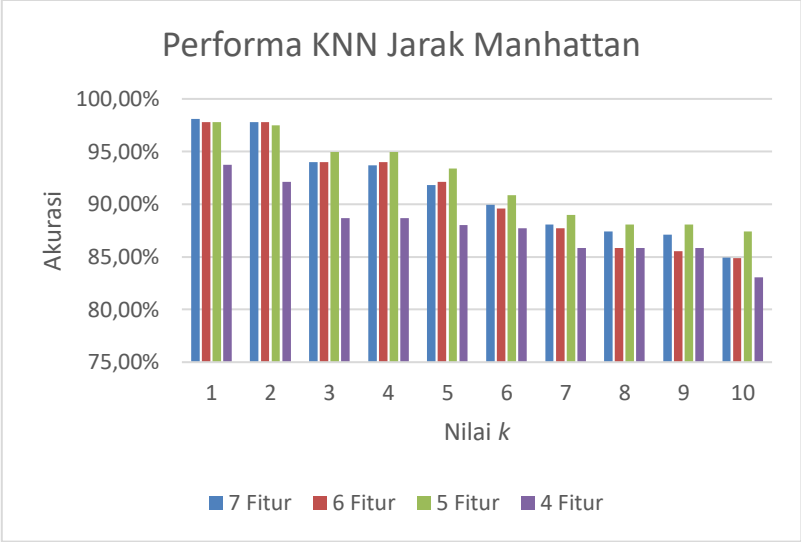
$$accuracy = \frac{TP_{rendah}+TP_{medium}+TP_{tinggi}}{Total\ Sampel}$$
$$= \frac{85+108+64}{318} = 0{,}8081$$

Model *naïve* Bayes tanpa seleksi fitur memiliki tingkat akurasi sebesar 80,81 persen, artinya algoritma *naïve* Bayes dapat memprediksi 80,81 persen kelas data uji sesuai dengan kelas sebenarnya. Selanjutnya dilakukan validasi silang menggunakan 10 *fold cross validation* dan menghasilkan rata-rata akurasi sebesar 89,61 persen.

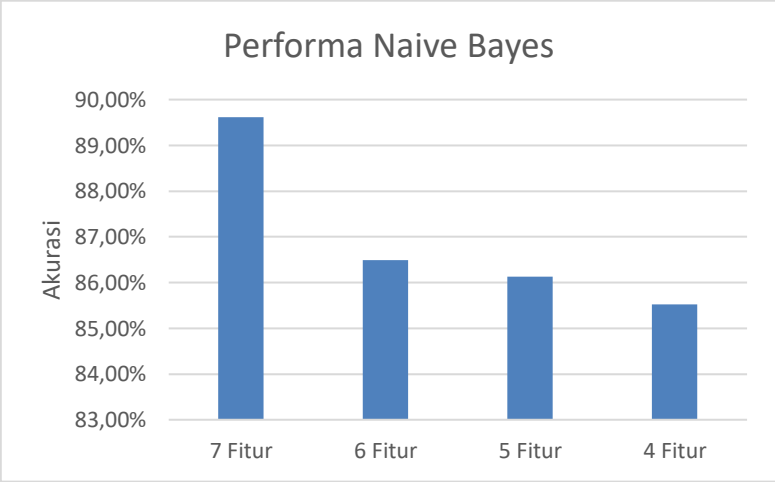
Perbandingan menggunakan seleksi fitur *information gain* dan tanpa seleksi fitur



Gambar 1 Grafik Akurasi Metode KNN Jarak Euclidean



Gambar 2 Grafik Akurasi Metode KNN Jarak Manhattan



Gambar 3 Grafik Akurasi Metode KNN *Naïve* Bayes

Berdasarkan Gambar 1 dan Gambar 2, nilai akurasi KNN mengalami penurunan ketika fitur dieliminasi, tetapi ketika menggunakan nilai $k = 3$ sampai $k = 10$ performa KNN dengan menggunakan lima fitur memiliki akurasi yang lebih tinggi dibandingkan dengan menggunakan tujuh fitur (tanpa seleksi fitur). Sedangkan performa metode naïve Bayes (Gambar 3) mengalami penurunan nilai akurasi ketika menggunakan seleksi fitur, semakin banyak fitur yang dieliminasi nilai akurasi dari algoritma tersebut semakin kecil.

Selanjutnya perbandingan performa KNN dan *naïve* Bayes dilakukan menggunakan lima fitur dan tujuh fitur karena saat menggunakan lima fitur ada peningkatan dalam akurasi algoritma KNN.

Tabel 7 Perbandingan Akurasi

Algoritma		Nilai Akurasi Tanpa Seleksi Fitur		Menggunakan Seleksi Fitur *Information Gain*	
KNN	K	Jarak Euclid	Jarak Manhattan	Jarak Euclid	Jarak Manhattan
	1	98,43%	98,11%	98,11%	97,80%
	2	98,10%	97,79%	97,79%	97,48%
	3	94,94%	94,00%	95,27%	94,96%
	4	94,00%	93,69%	95,27%	94,96%
	5	93,06%	91,80%	94,65%	93,39%
	6	90,87%	89,93%	92,13%	90,87%
	7	88,99%	88,05%	90,25%	88,99%
	8	88,04%	87,42%	89,94%	88,05%
	9	87,73%	87,11%	89,94%	88,05%
	10	87,42%	84,91%	88,67%	87,43%
Rata-rata KNN		92,16%	91,28%	93,20%	92,20%
Naïve Bayes		89,61%		86,13%	

Tabel 7 menunjukkan bahwa algoritma KNN tanpa seleksi fitur memiliki tingkat akurasi yang lebih baik daripada menggunakan teknik seleksi fitur *information gain* untuk nilai $k = 1$ dan $k = 2$ yaitu sebesar 98,43% dan 98,10%. Ketika algoritma KNN menggunakan nilai $k > 2$ tingkat akurasi dengan menggunakan 5 fitur memiliki tingkat akurasi yang lebih baik dibandingkan tanpa seleksi fitur. Ketika akurasi KNN $k = 1$ sampai $k = 10$ dirata-ratakan, algoritma KNN memiliki nilai akurasi yang lebih besar ketika menggunakan teknik seleksi fitur information gain. Pada algoritma naïve Bayes tingkat akurasi tanpa seleksi fitur memiliki nilai yang lebih baik dibandingkan menggunakan teknik seleksi fitur information gain. Untuk mengetahui apakah ada perbedaan yang signifikan terhadap dua algoritma yang memiliki perlakuan berbeda, maka dilanjutkan dengan uji signifikansi dengan hipotesis berikut

$H_0 : \mu_1 = \mu_2$ tidak ada perbedaan akurasi antara algoritma klasifikasi yang menggunakan teknik seleksi fitur information gain dan tanpa teknik seleksi fitur

$H_0 : \mu_1 \neq \mu_2$ ada perbedaan akurasi yang signifikan antara algoritma klasifikasi yang menggunakan teknik seleksi fitur information gain dan tanpa teknik seleksi fitur

Tabel 8 Nilai t hitung dan t tabel

t_{hitung}	$t_{0,025,20}$
2,78034	2,086

Tabel 8 menunjukkan $t_{hitung} > t_{tabel}$, yang artinya tidak cukup bukti untuk menolak hipotesis H_0. Hal tersebut menyatakan adanya perbedaan yang signifikan antara algoritma yang menggunakan teknik seleksi fitur information gain dan tanpa teknik seleksi fitur pada tingkat signifikan 0,05

SIMPULAN

Berdasarkan hasil dan pembahasan yang telah dipaparkan pada bab sebelumnya, akurasi KNN dengan menggunakan teknik seleksi fitur *information gain* memiliki rata-rata akurasi sebesar 93,20 persen dengan perhitungan jarak Euclid dan 92,20 persen dengan perhitungan jarak Manhattan. Algoritma *naïve* Bayes dengan teknik seleksi fitur *information gain* memiliki tingkat akurasi sebesar 86,13 persen. Selisih dari algoritma KNN dengan perhitungan jarak Euclid dan algoritma *naïve* Bayes untuk dataset kualitas susu sapi sebesar 7,07 persen.

Hasil perbandingan algoritma dengan seleksi fitur *information gain* (5 fitur) dan tanpa teknik seleksi fitur menunjukkan adanya perbedaan tingkat akurasi. Algoritma KNN dengan perhitungan jarak Euclid mengalami kenaikan tingkat akurasi sebesar 1,04 persen dan dengan perhitungan jarak Manhattan mengalami kenaikan tingkat akurasi sebesar 0,92 persen. Sedangkan algoritma *naïve* Bayes mengalami penurunan akurasi sebesar 3,48 persen ketika menggunakan teknik seleksi fitur *information gain*. Perlakuan yang berbeda tersebut memiliki perbedaan akurasi yang signifikan pada tingkat $\alpha = 0{,}05$.

DAFTAR PUSTAKA

Aini, S. H. A., Sari, Y. A., & Arwan, A. (2018). Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Komputer*, *2*(9), 2546–2554.

Bhatia, N., & Vandana. (2010). Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security*, *8*(2), 302–305.

Chormunge, S., & Jena, S. (2016). Efficient feature subset selection algorithm for high dimensional data. *International Journal of Electrical and Computer Engineering*, *6*(4), 1880–1888. https://doi.org/10.11591/ijece.v6i4.9800

Dewantoro, S., Herdiani, A., & Puspandari, D. (2019). Implementasi Information Gain sebagai Feature Selection pada Word Sense Disambiguation Bahasa Indonesia dengan Teknik

Klasifikasi Decision List. *E- Proceeding of Engineering*, *6*(3), 10425–10435.

Gorunescu, F. (2011). *Data Mining: Concept, Models, and Techniques*. Springer.

Hafidzullah, M., Sutrisno, & Marji. (2019). Seleksi Fitur dengan Information Gain pada Identifikasi Jenis Attention Deficit Hyperactivity Disorder Menggunakan Metode Modified K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(11), 10444–10452.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concept and Techniques*. Morgan Kaufmann.

Hidayat, L. R., Setiawan, B. D., & Nurwasito, H. (2016). *Pengklasifikasian Kualitas Susu Sapi dengan Algoritma Fuzzy K-Nearest Neighbor (FK-NN)*.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge In Data An Introduction to Data Mining*. Wiley.

Leung, K. S., Lee, K. H., Wang, J. F., Ng, E. Y. T., Chan, H. L. Y., Tsui, S. K. W., Mok, T. S. K., Tse, P. C. H., & Sung, J. J. Y. (2011). Data mining on DNA sequences of hepatitis B virus. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(2), 428–440. https://doi.org/10.1109/TCBB.2009.6

Mulak, P., & Talhar, N. (2015). Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. *International Journal of Science and Research*, *4*(7), 2101–2104.

Multamiah, L., Utami, S., & Sudewo, A. T. A. (2013). Kajian Kadar Lemak dan Bahan Kering Tanpa Lemak Susu Kambing Sapera di Cilacap dan Bogor. *Jurmal Ilmiah Peternakan*, *1*(3), 874–880.

Mutmainnah, U., Darma Setiawan, B., & Dewi, C. (2019). Pengaruh Seleksi Fitur Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kelancaran Pembayaran Kredit Kendaraan. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(9), 8882–8888. http://j-ptiik.ub.ac.id

Nabella, F. Y., Sari, Y. A., & Wihandika, R. C. (2019). Seleksi Fitur Information Gain Pada Klasifikasi Citra Makanan Menggunakan Hue Saturation Value dan Gray Level Co-Occurrence Matrix. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(2), 1892–1900.

Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*, *6*(2), 256–261.

Sari, B. N. (2016). Implementasi Teknik Seleksi Fitur Information Gain pada Algoritma Klasifikasi Machine Learning untuk Prediksi Performa Akademik Siswa. *Seminar*

Nasional Teknologi Informasi Dan Multimedia, *2*(9), 55–60.

Shaltout, N. A., El-Hefnawi, M., Rafea, A., & Moustafa, A. (2014). Information Gain as a Featrure Selection Method for the Efficient Classification of Influenza Based on Viral Hosts. *Proceedings of the World Congress on Engineering*, *1*.

Shrijayan. (2022, August 3). *Milk Quality Prediction*. https://www.kaggle.com/datasets/cpluzshrijayan/milkquality

Syarli, & Muin, A. A. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan. *Jurnal Ilmiah Ilmu Komputer*, *2*(1), 22–26.

Wasito. (2017). Persepsi Dan Adopsi SNI 3141-1: 2011 Keluarga Peternak Sapi Perah Kawasan Usaha Peternakan(KUNAK) Kabupaten Bogor. *Jurnal Standardisasi*, *19*(3), 241–254.

Wiranti, N., Wanniatie, V., Husni, A., & Qisthon, A. (2022). Kualitas Susu Sapi Segar pada Pemerahan Pagi dan Sore. *Jurnal Riset Dan Inovasi Peternakan*, *6*(2), 123–128.

.