Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 12, No. 6, Desember 2025, hlm. 1395-1404 e-ISSN: 2528-6579

KLASIFIKASI RASA BUAH JERUK PONTIANAK (*CITRUS NOBILIS VAR. MICROCARPA*) MENGGUNAKAN METODE K-CLUSTER CLASSIFICATION TREE (K-CT)

Jimmy Tjen[*1]

[1]Program studi Informatika, Fakultas Teknologi Informasi, Universitas Widya Dharma Pontianak, Pontianak
Email: [1]jimmy.tjen@mathmods.eu
[*]Penulis Korespondensi

(Naskah masuk: 07 Januari 2025, diterima untuk diterbitkan: 10 Desember 2025)

Abstrak

Revolusi Industri 4.0 telah mendorong penggunaan teknologi di berbagai aspek kehidupan, seperti industri makanan, dengan konsep *machine learning* digunakan untuk mengidentifikasi kualitas dan rasa dari bahan makanan. Perkembangan teknologi ini mendorong pengembangan dari metode baru yang lebih efisien dalam waktu komputasi, namun memiliki kemampuan model prediktif yang akurat. Penelitian ini bertujuan untuk memperkenalkan sebuah metode *ensemble* baru yang mengkombinasikan metode *classification tree* (CT) dan metode k-means, yang disebut sebagai metode *k-cluster classification tree* atau k-CT. metode k-CT merupakan metode yang dirancang untuk mengefisiensikan waktu komputasi dari metode *CT* tanpa mengurangi kemampuan prediktif dari metode tersebut. Pada penelitian ini, metode k-CT divalidasi menggunakan data primer yang diambil dari pengamatan sifat fisik dari buah jeruk Pontianak. Dari 457 sampel data yang ada, 80% data digunakan untuk melatih model pohon, sedangkan 20% yang tersisa digunakan untuk memvalidasi kualitas prediksi dari model. Berdasarkan pada percobaan yang dilakukan, diperoleh 3 temuan. Pertama, metode k-CT dapat mengklasifikasikan rasa dari buah jeruk Pontianak dengan akurasi sebesar 92%. Hasil ini menunjukan bahwa metode k-CT memiliki performa model prediktif yang lebih baik jika dibandingkan dengan metode CT, *random forest* dan *gradient boosting.* Kedua, ditemukan bukti lemah secara statistik ($\alpha < 0{,}1$) bahwa metode k-CT memiliki kompleksitas waktu yang lebih singkat daripada metode *CT*, sesuai dengan Lema yang dibuktikan. Ketiga, berdasarkan pada aturan jika – maka yang dibentuk oleh metode k-CT, diketahui bahwa warna jeruk bukanlah faktor dominan yang menentukan rasa dari buah jeruk, melainkan diameter buah jeruk yang merupakan faktor dominan untuk menentukan rasa buah jeruk.

Kata kunci: *Classification Tree, Decision Tree Learning, Jeruk Pontianak, K-means, Metode Klasifikasi.*

TASTE CLASSIFICATION OF CITRUS NOBILIS VAR. MICROCARPA USING K-CLUSTER CLASSIFICATION TREE (K-CT) METHOD

Abstract

The Industrial Revolution 4.0 has driven the integration of technology into various aspects of life, including the food and beverage industry, where machine learning methods are employed to evaluate food quality and taste. Consequently, the development of efficient machine learning techniques that provide accurate predictions with reduced computational complexity has become increasingly important. This research introduces a novel classification tree (CT)-based algorithm, termed the k-cluster classification tree (k-CT). The k-CT enhances the CT method by offering faster computations while preserving its predictive accuracy. The proposed methodology was validated using a primary dataset comprising the physical properties of "jeruk Pontianak" (Citrus nobilis var. microcarpa) oranges. Of the 457 available samples, 80% were utilized for training the tree-based models, while the remaining 20% were reserved for validating predictive accuracy. The experiments yielded three key findings. First, the k-CT achieved an accuracy of 92% in classifying the taste of "jeruk Pontianak," outperforming CT, random forest, and gradient boosting methods. Second, there is weak evidence ($\alpha < 0.1$) suggesting that the k-CT performs faster than the CT method. Lastly, based on the if-then rules derived from the k-CT tree structure, it was observed that the skin color of "jeruk Pontianak" does not significantly influence its taste. Instead, the diameter of the fruit has a strong impact on its taste.

Keywords: *Classification Tree, Decision Tree Learning, Jeruk Pontianak, K-means, Classification Methods*

1. PENDAHULUAN

Perkembangan teknologi terutama pada era industri 4.0 mengakibatkan teknologi digunakan di berbagai aspek kehidupan, seperti pada identifikasi rasa dan kualitas makanan (Singh et al., 2022; Hassoun et al., 2023; Konur et al., 2023). Rasa

makanan mengacu pada perasaan atau sensasi yang dirasakan oleh indra pengecap (reseptor lidah) dan indra penciuman ketika makanan masuk ke dalam tubuh (Ji et al., 2023; Rojas et al., 2023; Zeng et al., 2023). Proses identifikasi ini dapat dilakukan dengan cara menghasilkan data dari makanan seperti ciri fisik dari makanan dan pengujian rasa yang kemudian dapat diproses dengan menggunakan berbagai jenis algoritma *machine learning* (Bhargava dan Bansal, 2020; Hemamalini et al., 2022). Salah satu algoritma yang digunakan untuk mengidentifikasi kualitas dan rasa dari makanan adalah metode *classification tree* (CT) (Jiménez-Carvelo et al., 2019; Nilashi et al., 2021; Bhattacharyya and Pal, 2024).

Metode CT merupakan metode *machine learning* dari rumpun *decision tree learning*, yang membangun struktur pohon biner untuk mengklasifikasikan data ke dalam kelas tertentu (Hand, 2020). Metode CT akan membagi data ke dalam sub-himpunan tertentu berdasarkan pada kemiripan fitur data, sehingga akan terkumpul data yang serupa pada tiap sub-himpunan yang disebut sebagai titik daun (*leaf node*), yang mempermudah proses klasifikasi data (Charbuty and Abdulazeez, 2021; Aghaei, Gómez and Vayanos, 2024). Metode CT telah terbukti mampu mengklasifikasikan kualitas dari bahan pangan seperti yang ditunjukan pada penelitian berikut: (Aulia, Wijaya dan Hidayat, 2021; Farah et al., 2021; Mascellani et al., 2021; Averós, Lavín dan Estevez, 2024; Tjen, 2024a)

Penelitian terkait: Pada tahun 2024, (Tjen, 2024) melakukan penelitian untuk mengklasifikasikan kualitas dari minuman *wine* dengan menggunakan metode CT dan *entropy-based subset selection* (E-SS). Berdasarkan pada percobaan yang dilakukan, diketahui bahwa metode E-SS CT mampu mengklasifikasikan kualitas dari minuman *wine* dengan akurasi di atas 95%. Penelitian yang dilakukan oleh (Mascellani et al., 2021) menunjukan bahwa metode CT dapat digunakan untuk mengklasifikasikan kualitas minuman wine asal Ceko dengan akurasi di atas 93%. Penelitian yang dilakukan oleh (Aulia, Wijaya dan Hidayat, 2021) menunjukan bahwa metode *gradient boosting* (GB) yang merupakan metode *ensemble* dari CT mampu mengklasifikasikan kualitas beras yang ada di Indonesia dengan akurasi sebesar 96%. Penelitian oleh (Farah et al., 2021) menyatakan bahwa metode *decision tree* seperti *random forest* (RF), GD dan CT dapat digunakan untuk menentukan keaslian dan kualitas dari susu sapi. Terakhir, pada tahun 2024, (Averós, Lavín and Estevez, 2024) menunjukan bahwa metode CT dapat mengklasifikasikan kesehatan dari ayam broiler atau ayam potong, dengan akurasi di atas 80%.

Penelitian di atas telah menunjukan potensi dari algoritma CT dan turunannya dalam mengidentifikasi kualitas dan rasa dari bahan pangan. Dalam kasus ini, metode CT beserta turunannya dalam bentuk RF dan GD terbukti mampu meningkatkan akurasi dari model CT. Namun, terlepas dari kelebihan metode CT, terdapat sebuah kelemahan, yakni waktu pemrosesan dari metode CT berada pada $O(n \cdot m \cdot log(m))$, dengan m menyatakan jumlah fitur dan n menyatakan jumlah sampel (Smarra, Tjen and D'Innocenzo, 2022). Sehingga, untuk sampel data yang besar, metode ini akan membutuhkan waktu pemrosesan yang lama. Terkait dengan permasalahan ini, maka dibutuhkan solusi untuk dapat mengontrol kompleksitas waktu dari metode CT tanpa mengurangi performa dari algoritma tersebut.

Kontribusi: penelitian ini diformulasikan untuk menyelesaikan permasalahan komplesitas waktu dari metode CT dengan cara melakukan partisi sampel data menggunakan metode k-means. Metode k-means merupakan metode *clustering*, yang menghasilkan klaster berdasarkan pada perbedaan jarak antar data (sebagai contoh jarak kuadratik atau cosinus) (Ahmed, Seraj dan Islam, 2020; Ikotun et al., 2023). Dengan menggunakan metode k-means, maka data dapat dibagi menjadi beberapa klaster dan berpotensi untuk mempersingkat waktu pelatihan model CT tanpa mengurangi kemampuan prediktif model. adapun yang menjadi kontribusi dari penelitian ini adalah sebagai berikut:

1. Penelitian ini menurunkan konsep matematis terkait dengan metode CT yang dipartisi datanya dengan metode k-means. Metode ini disebut sebagai *k-cluster CT* (k-CT).
2. Penelitian ini menurunkan algoritma k-CT yang merupakan metode *ensemble* baru, untuk memprediksi rasa dari sebuah makanan.
3. Metode pada penelitian ini divalidasi dengan menggunakan data primer yang diperoleh dari observasi sifat fisik dari buah jeruk Pontianak (*citrus nobilis var. microcarpa*).

Keterbaharuan topik: penelitian memperkenalkan metode k-CT yang merupakan metode *ensemble* baru berbasis pohon keputusan. Metode ini akan digunakan untuk menentukan rasa dari buah jeruk Pontianak berdasarkan pada ciri fisik. Buah Jeruk Pontianak dipilih sebagai sampel penelitian, karena buah jeruk Pontianak merupakan buah endemik yang ada di Kalimantan Barat. Selain itu, buah jeruk juga merupakan salah satu produk pangan yang dikirim ke berbagai provinsi di Indonesia.

Ciri fisik yang ditinjau pada penelitian ini meliputi warna kulit, ukuran pori, bentuk jeruk, keberadaan dari bercak pada kulit jeruk akibat infeksi virus *thrips* dan tekstur kulit. lebih lanjut, penelitian ini akan membandingkan performa dari metode k-CT dengan metode berbasis *decision tree* lainnya seperti RF dan GB, mengacu pada penelitian sebelumnya.

Artikel penelitian ini terdiri dari 5 bagian. Bagian pertama membahas mengenai latar belakang, penelitian terkait, kontribusi dan keterbaharuan topik. Pada Bagian ke-2 akan dibuktikan secara matematis bahwa dimungkinkan untuk metode k-CT memiliki kompleksitas waktu yang lebih singkat dari CT.

Lebih lanjut, akan dibahas pula algoritma dari metode k-CT. Bagian ke-3 akan menjelaskan mengenai data yang digunakan, serta konfigurasi penelitian (*experiment setup*). Bagian ke-4 akan menunjukan performa dari metode yang digagas, beserta perbandingannya terhadap metode berbasis pohon lainnya. Terakhir, pada Bagian ke-5 akan diuraikan kesimpulan dari penelitian dan arah penelitian lanjutan untuk metode k-CT.

2. METODE K-CT

Pada bagian ini, akan dibahas terlebih dahulu pembuktian matematis dari metode k-CT. dalam kasus ini, akan ditunjukan bahwa dimungkinkan untuk metode k-CT memiliki kompleksitas waktu yang lebih singkat atau sama dengan metode CT asalkan jumlah klaster k dipilih cukup kecil untuk n yang besar. Terkait konsep dasar yang digunakan pada penelitian ini, silahkan merujuk pada (Sarker, 2021) terkait dengan konsep dari metode k-means dan (James et al., 2023) untuk metode berbasis *decision tree*.

2.1. Uraian Matematis Metode k-CT

Untuk membuktikan bahwa metode k-CT dapat diselesaikan lebih cepat daripada metode CT, maka perlu ditunjukan bahwa kompleksitas waktu dari CT lebih besar daripada k-CT. Secara teknis, metode k-CT terdiri dari 2 proses utama: pembagian klaster data dan metode CT.

Untuk sebuah himpunan data $X = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots \boldsymbol{x_n}]; X \in \mathbb{R}^{m\times n}$ dengan jumlah sampel m, jumlah fitur sebanyak n, dan $\boldsymbol{x_i} = [x_i(1)\ x_i(2)\ \dots x_i(m)]^{\intercal}; \boldsymbol{x_i} \in \mathbb{R}^m$, metode k-means memiliki kompleksitas waktu sebesar $O(m \times n \times k)$ dengan k menyatakan jumlah klaster (Ahmed, Seraj and Islam, 2020). asumsikan bahwa metode k-CT membagi data menjadi k klaster. Misalkan $X_k = [\boldsymbol{x_{1,k}}\ \boldsymbol{x_{1,k}}\ \dots \boldsymbol{x_{n,k}}] \in \mathbb{R}^{m_k\times n}$ adalah himpunan data baru yang dihasilkan dari proses *clustering* dengan menggunakan k-means, sehingga $X_1 \cup x_2 \cup \dots \cup X_k = X$, dengan m_k menyatakan banyak sampel pada sub himpunan data ke-k.

Diketahui bahwa metode CT memiliki kompleksitas waktu sebesar $O(n \cdot m \cdot log(m))$. Jika diasumsikan bahwa $m_1 \approx m_2 \approx \dots \approx m_k$, maka setiap klaster memiliki banyak sampel sekitar $\frac{m}{k}$. Dalam kasus ini, maka kompleksitas waktu untuk metode k-CT akan berada pada $O(m \cdot n \cdot k + n \cdot \sum_{i=1}^{k} m_i \cdot \log(m_i))$. Sehingga, untuk membuktikan bahwa metode k-CT dapat diselesaikan lebih cepat daripada metode CT, maka cukup membuktikan bahwa pertidaksamaan:

$$O(n \cdot m \cdot \log(m)) \geq O\left(m \cdot n \cdot k + n \cdot \sum_{i=1}^{k} m_i \cdot \log(m_i)\right) \quad (1)$$

bernilai benar untuk k dan n tertentu.

Pembuktian pertama. Pembuktian pertama dilakukan dengan menggunakan aturan dari kompleksitas waktu. Berdasarkan pada asumsi sebelumnya, bahwa jika semua klaster dianggap memiliki ukuran yang kurang lebih sama, dengan ukuran m_c, maka

$$n \cdot \sum_{i=1}^{k} m_i \cdot log(m_i) = n \cdot (k \cdot (m_c \cdot log(m_c)). \quad (2)$$

Namun, karena $m_c \cdot k = m$, maka persamaan (2) dapat disederhanakan menjadi

$$n \cdot \sum_{i=1}^{k} m_i \cdot log(m_i) = n \cdot m \cdot log\left(\frac{m}{k}\right). \quad (3)$$

Dengan mensubstitusikan persamaan (3) ke dalam ruas kanan dari persamaan (1), maka dapat diperoleh

$$O(n \cdot m \cdot \log(m)) \geq O\left(m \cdot n \cdot k + n \cdot m \cdot \log\left(\frac{m}{k}\right)\right)$$
$$\Rightarrow O(n \cdot m \cdot log(m)) \geq O\left(m \cdot n \cdot \left(k + log\left(\frac{m}{k}\right)\right)\right) \quad (4)$$

Dalam kasus ini, perhatikan bahwa jika k dipilih sedemikian rupa, sehingga

$$k + log(k) \leq log(m), \quad (5)$$

maka nilai dari k akan kurang lebih setara dengan $\log\left(\frac{m}{k}\right)$. Sehingga,

$$O\left(m \cdot n \cdot \left(k + log\left(\frac{m}{k}\right)\right)\right) = O(2 \times m \cdot n \cdot log\, m). \quad (6)$$

Dengan,

$$O(2 \times m \cdot n \cdot log(m)) \in O(m \cdot n \cdot log(m), \quad (7)$$

akan mengakibatkan persamaan (4) menjadi benar.

Dalam kasus ini, pembuktian dengan kompleksitas waktu mengisyaratkan bahwa k harus dipilih sedemikian rupa, sehingga memenuhi kondisi pada persamaan (5). Dengan pemilihan ini, maka metode k-CT dapat memiliki kompleksitas waktu yang setara atau lebih cepat daripada metode CT.

Pembuktian kedua. Pembuktian kedua dapat dilakukan dengan menggunakan prinsip *superadditivity* dari sebuah fungsi konveks (Dragomir, 2021). Tinjau persamaan (1). Dalam kasus ini, untuk menunjukan bahwa metode k-CT memiliki kompleksitas waktu yang lebih singkat atau setara dengan CT maka cukup menyelesaikan:

$$n \cdot m \cdot log(m) \geq m \cdot n \cdot k + n \cdot \sum_{i=1}^{k} m_i \cdot log(m_i), \quad (8)$$

atau menunjukan bahwa $\exists k \in \mathbb{Z}^+/\{1\}$, sehingga:

$$m \cdot log(m) - \sum_{i=1}^{k} m_i \cdot log(m_i) - m \cdot k \geq 0 \quad (9)$$

Persamaan (9) secara khusus dapat diselesaikan dengan menggunakan Lema 1.

Lema 1:
" Untuk $m, m_1, m_2, \dots m_k, k \in \mathbb{Z}^+; m_1 + m_2 + \cdots + m_k = m$, *dan sebuah fungsi* $f(n) = n \cdot log\,(n)$; $f: \mathbb{R} \to \mathbb{R}^+$ *berlaku bahwa:*

$$f(m) \geq \sum_{i=1}^{k} f(m_i)\ "$$

Pembuktian:
Pembuktian dapat dilakukan menggunakan induksi dan prinsip *superadditivity* dari fungsi konveks: untuk dua buah bilangan $a, b \in \mathbb{R}$, berlaku bahwa:

$$f(a+b) \geq f(a) + f(b). \quad (10)$$

Secara spesifik, persamaan (10) dapat diterapkan pada fungsi f, karena untuk $x \in \mathbb{R}$; fungsi f merupakan fungsi konveks, yang dibuktikan dengan:

$$\frac{d^2}{dx^2} f(x) = \frac{d^2}{dx^2} x \cdot \log x$$
$$\Rightarrow \frac{d^2}{dx^2} f(x) = \frac{d}{dx}\left(\frac{1}{\ln(10)} + \log(x)\right)$$
$$\Rightarrow \frac{d^2}{dx^2} f(x) = \frac{1}{x \cdot ln(10)} = \frac{1}{ln(10)} \cdot \frac{1}{x} > 0 \quad (11)$$

Algoritma 1: Metode k-CT

Masukan: $[X\ y], k$
Keluaran: $\hat{y}$
Proses:
$\boldsymbol{c_a} =$ k-means$(X, k, max\ ("silhouette"))$;
$X_d = [X\ c_a\ y]$;

```
# Tahap pertama
untuk i = 1:k
    [C_i, y_i] = [X, y: c_a = i];
akhiri untuk i

# Tahap kedua
untuk i = 1:k
    f_CTi = pohon_klasifikasi(C_i, y_i);
akhiri untuk i

# prediksi hasil
x̃_t = X(t, :)
ŷ(t) = { f_CT1(x̃_t) jika x̃_t ∈ c_1
       {        ⋮
       { f_CTn(x̃_t) jika x̃_t ∈ c_k
proses selesai
```

$\forall x > 0$.

Sekarang, tinjau $k = 2$. Dalam kasus ini, berlaku bahwa $f(m) \geq f(m_1) + f(m_2)$, berdasarkan pada persamaan (10). Untuk kasus $k = 3$, terlihat bahwa: $f(m_1) + f(m_2) + f(m_3) \leq f(m_1 + m_2) + f(m_3)$. Lebih lanjut, berlaku pula $f(m_1 + m_2) + f(m_3) \leq f(m_1 + m_2 + m_3)$, dan seterusnya, sehingga Lema 1 terbukti.

Dengan menggunakan Lema 1, maka selisih dari bentuk pertama dan kedua dari ruas kiri persamaan (9) sudah dipastikan bernilai ≥ 0. Dalam kasus ini, terlihat bahwa dengan memilih k sesuai dengan persamaan (5), maka persamaan (9) akan terbukti benar. Hal ini menunjukan bahwa dimungkinkan untuk metode k-CT memiliki kompleksitas waktu yang lebih singkat atau setara dengan metode CT. pembuktian secara numeris menggunakan simulasi akan ditunjukan pada Bagian 4.

2.2. Algoritma Metode k-CT

Tinjau himpunan data X seperti pada sub bagian 2.1. Misalkan $\boldsymbol{y} = [y(1)\ y(2) \dots y(m)]^\intercal$; $\mathbf{y} \in \mathbb{Z}^m$ merupakan vektor yang menyatakan kualitas atau rasa dari bahan makanan. Sebagaimana sehingga setiap vektor data $\tilde{\boldsymbol{x}}_i = [x_1(i)\ x_2(i) \dots x_n(i)]^\intercal; \mathbb{R}^n$ akan berkorespondensi dengan $y(i)$. Berdasarkan pada definisi tersebut, tujuan dari metode k-CT adalah untuk membangun k buah klaster dengan setiap $\boldsymbol{c_i} = [c_i(1)\ c_i(2) \dots c_i(n)]; c_i \in \mathbb{R}^n; i = 1,2, \dots, k$ menyatakan centroid dari klaster ke-i. Lebih lanjut, setiap klaster akan dilengkapi dengan sebuah model CT, sehingga model prediksi kualitas atau rasa dari bahan makanan dapat dinyatakan sebagai:

$$\hat{y}(j) = \begin{cases} f_{CT1}(\tilde{\boldsymbol{x}}_j)\ jika\ \tilde{\boldsymbol{x}}_j \in \boldsymbol{c_1} \\ \vdots \\ f_{CTk}(\tilde{\boldsymbol{x}}_j)\ jika\ \tilde{\boldsymbol{x}}_j \in \boldsymbol{c_k} \end{cases}. \quad (12)$$

Dengan $\hat{\boldsymbol{y}} = [y(1)\ y(2) \dots y(m)]^\intercal; \hat{\boldsymbol{y}} \in \mathbb{Z}^m$ menyatakan prediksi dari kualitas atau rasa bahan makanan dengan menggunakan metode k-CT. secara khusus, persamaan (12) dapat dibentuk dalam 2 langkah: bangun sub himpunan data, prediksi dengan model CT.

Langkah pertama. Langkah pertama dari metode k-CT adalah membangun sub himpunan data berdasarkan pada algoritma k-means. Asumsikan bahwa akan dibangun k buah klaster. Misalkan $K_a = \{k_a(1), k_a(2), \dots k_a(k)\}; k_a \in \mathbb{Z}^k$ merupakan himpunan nama klaster. Lebih lanjut, misalkan $\boldsymbol{c_a} = [c_a(1)\ c_a(2) \dots c_a(m)]^\intercal; \boldsymbol{c_a} \in K_a^m$ merupakan sebuah vektor yang menunjukan klaster mana yang ditempati oleh data tertentu, sebagai contoh $c_a(i)$ menunjukan nama klaster untuk $\tilde{\boldsymbol{x}}_i$. Berdasarkan pada definisi tersebut, tujuan dari langkah ini adalah untuk menghasilkan sebuah himpunan data teraugmentasi $X_d = [X\ \boldsymbol{c_a}]; X_d \in \mathbb{R}^{m \times (n+1)}$ yang berfungsi untuk menunjukan klaster dari setiap sampel.

Perlu diketahui bahwa metode k-means merupakan metode yang memerlukan pemilihan titik awal. Terkait dengan hal tersebut, maka klaster yang dipilih merupakan klaster yang menghasilkan nilai rerata *silhouette* tertinggi. Misalkan $\bar{\boldsymbol{s}}$ menyatakan rerata dari vektor *silhouette*, dengan $\boldsymbol{s} = [s(1)\ s(2) \dots s(m)]^\intercal; \boldsymbol{s} \in [-1,1]^m$ dapat ditentukan sebagai:

Gambar 1. Jeruk Pontianak yang digunakan untuk proses identifikasi rasa.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}. \quad (13)$$

Dengan:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\substack{c_a(j)=c_a(i) \\ i \neq j}} d(i,j), \quad (14)$$

menyatakan kohesi atau rerata jarak dari dari sampel ke-i terhadap sampel lain yang berada pada klaster yang sama dengan sampel ke-i, dan

$$b(i) = \min_{c_a(i) \neq c_a(j)} \frac{1}{|C_j|} \sum_{c_a(j) \neq i} d(i,j) \quad (15)$$

menyatakan separasi atau rerata jarak dari sampel ke-i terhadap sampel lain yang terletak pada klaster yang berbeda, dengan $|C_i|$ menyatakan kardinalitas atau jumlah sampel yang memiliki klaster yang sama dengan sampel ke-i, dan $d(i,j)$ menyatakan jarak antara sampel i dan j.

Hasil akhir dari langkah pertama akan menghasilkan sekelompok sub himpunan dengan nilai *silhouette* tertinggi. Pada proses berikutnya, untuk setiap klaster akan dibangun sebuah model CT untuk dapat menghasilkan prediksi.

Tahap kedua. Proses ini akan berfokuskan untuk membangun model CT untuk setiap sub himpunan data. Untuk klaster ke-i, misalkan bahwa C_i menyatakan sub himpunan data dari X dengan $\boldsymbol{c_a} = i$. lebih lanjut, misalkan pula $\boldsymbol{y_i}$ sebagai sebuah vector yang berkorespondensi dengan C_i. dalam kasus ini, setiap model dari CT akan dibangun dengan mengikuti persamaan:

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = f_{CTi}(\boldsymbol{C_i}). \quad (16)$$

Dengan f_{CTi} menyatakan model CT untuk klaster ke-i. dengan mgnulangi proses untuk setiap kalster, maka akan dihasilkan model CT untuk setiap klaster, dan dimungkinkan untuk melakukan prediksi sesuai dengan persamaan (12). Algoritma untuk metode k-CT ditunjukan oleh Algoritma 1.

3. METODE PENELITIAN

Pada bagian ini, akan dijelaskan data yang digunakan untuk memvalidasi performa model dan beberapa pengaturan terkait dengan percobaan yang dilakukan. Pada penelitian ini, data yang digunakan merupakan data primer yang diperoleh berdasarkan pada observasi dan pengujian rasa dari buah jeruk Pontianak.

3.1. Data Penelitian

Data yang digunakan merupakan data yang berasal dari pengamatan fisik dan pengujian rasa dari buah jeruk Pontianak (lihat Gambar 1). Buah jeruk dipilih secara acak, dari 4 orang penjual jeruk yang ada di Kalimantan Barat, dan kemudian dipilih 10 orang responden untuk menguji rasa dari buah jeruk.

Data yang dihasilkan terdiri dari 457 sampel dengan 7 buah fitur: warna dominan dari kulit jeruk (hijau atau kuning), ukuran pori dari kulit jeruk (besar atau kecil), keberadaan *thrips* atau bercak kehitaman pada jeruk (ada atau tidak ada), bentuk buah jeruk (bulat atau elips), permukaan kulit jeruk (halus atau keriput), diameter dari buah jeruk dalam satuan cm dan rasa dari buah jeruk (asam atau manis). Dalam kasus ini, rasa dari buah dapat dipilih oleh setiap responden sesuai dengan rasa dominan yang mereka rasakan. Pengisian data dilakukan secara tertutup, artinya responden tidak mengetahui dan tidak terpengaruh oleh jawaban dari responden lain.

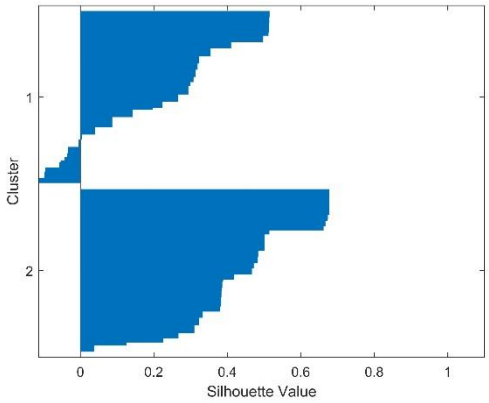
Tabel 1. Representasi numerik untuk fitur dan kelas data.

Nama Fitur	Representasi Numerik	Representasi kelas
Warna	$\boldsymbol{x_1}$	Hijau = 0 Kuning = 1
Ukuran pori	$\boldsymbol{x_2}$	Kecil = 1 Besar = 0
Thrips	$\boldsymbol{x_3}$	Ada = 1 Tidak ada = 0
Bentuk	$\boldsymbol{x_4}$	Bulat = 1 Elips = 0
Tekstur kulit	$\boldsymbol{x_5}$	Halus = 0 Keriput = 1
Diameter	$\boldsymbol{x_6}$	-

Pada penelitan ini, 80% data akan digunakan untuk melatih model, dan 20% yang tersisa akan digunakan untuk memvalidasi performa model. dalam penelitan ini, semua fitur data digunakan untuk melatih model k-CT. namun, diameter buah jeruk tidak akan digunakan untuk membangun klaster. Hal ini disebabkan karena fitur lain yang tersisa berasal dari jenis data diskrit, sedangkan diameter merupakan fitur yang bersifat kontinyu.

Klaster untuk k-CT dibangun dengan menggunakan jarak Hamming, yang dalam kasus ini untuk 2 buah titik sampel: i dan j dan himpunan data X, jarak dari kedua sampel dinyatakan sebagai:

$$d(i,j) = \sum_{k=1}^{n} x_{ij}(k),$$
$$x_{ij}(k) = \begin{cases} 1 \text{ jika } x_k(i) \neq x_k(j) \\ 0 \text{ jika sebaliknya} \end{cases}. \quad (17)$$



Gambar 2. Grafik silhouette untuk $k = 2$

Secara umum, semakin besar jarak dari dua buah sampel, semakin tidak mirip sampel tersebut. Untuk proses *clustering*, jumlah klaster ideal yang dapat digunakan adalah 2. Hal ini karena terdapat 365 data latih yang dengan $\log(365) = 2{,}55$. Sehingga nilai k yang dapat dipilih berdasarkan pada persamaan (5) adalah 2.

Untuk mempersingkat penyebutan nama fitur, maka setiap fitur akan dikodekan ke dalam bentuk yang sesuai dengan penjabaran algoritma pada Bagian 2. Tabel 1 menyatakan representasi numerik untuk setiap fitur dari buah jeruk Pontianak.

3.2. Konfigurasi Penelitian (Experiment Setup)

Proses pengujian dari metode akan dilakukan dengan menggunakan data sesuai pada Sub-bagian 3.1. Secara khusus, performa dari metode k-CT akan dibandingkan dengan metode lain seperti CT, RF dan GB sesuai dengan yang telah digunakan pada penelitian terdahulu. Untuk metode RF dan GB, digunakan variasi 10, 50 dan 100 pohon. Lebih lanjut, metode *adaptive boosting* (ada boost) digunakan untuk proses optimasi metode GB.

Terkait dengan kualitas model prediktif, terdapat 8 besaran yang ditinjau: *true positive rate* (TPR), *true negative rate* (TNR), *false positive rate* (FPR), *false negative rate* (FNR), akurasi ($A\%$), skor F1 seperti yang digunakan pada (Tjen, 2024) serta presisi ($P\%$) dan *recall* ($R\%$). Presisi mengacu kepada kemampuan dari model untuk dapat dengan benar menebak kelas positif relatif terhadap jumlah tebakan kelas positif yang dihasilkan. Secara matematis, presisi dinyatakan sebagai:

$$P\% = \frac{TP}{TP + FP} \times 100\%. \tag{18}$$

Dengan TP atau *true positive* menyatakan jumlah tebakan kelas positif yang benar dan FP atau *false positive* menyatakan jumlah tebakan positif yang salah diprediksi (seharusnya negatif namun diprediksi positif). Sedangkan *recall* menyatakan kemampuan dari model untuk menebak semua kelas positif yang ada di dalam data. Persamaan matematis untuk recall dinyatakan dalam persamaan (19).

$$R\% = \frac{TP}{TP + FN} \times 100\%. \tag{19}$$

Dengan, FN atau *false negative* menyatakan banyaknya tebakan negatif yang salah prediksi.

Pada penelitian ini, rasa manis akan diasosiasikan dengan kelas positif dan asam sebagai kelas negatif. Hal ini didasarkan pada secara normal, manusia lebih menginginkan memakan buah jeruk yang berasa manis ketimbang yang berasa asam. Sehingga rasa asam diasosiasikan ke dalam kelas negatif. Namun pemilihan jenis kelas ini bergantung pada tujuan penelitian dan dapat ditentukan sesuai dengan tujuan yang ingin dicapai dalam penelitian.

Metode terbaik dalam penelitan ini dipilih sebagai metode dengan tingkat akurasi tertinggi dengan waktu tersingkat berdasarkan pada proses pengujian menggunakan mesin dengan CPU i7-7th Gen, GPU GTX 1050 Ti, dan RAM 16 GB. Untuk menghasilkan perbandingan yang adil, maka pengukuran waktu akan dilakukan tanpa memperhitungkan waktu preprocessing data dan prediksi data yang dalam kasus ini akan merugikan metode RF dan GB. Lebih lanjut, percobaan akan diulang sebanyak 30 kali per metode untuk memperoleh hasil yang valid.

Tabel 4. Hasil uji T untuk metode k-CT dan CT

Parameter	k-CT	CT
rerata	0,0145	0,0152
varians	$3{,}6 \times 10^6$	$1{,}19 \times 10^6$
Asumsi varians sama		
t-hitung	1,661	
t-kritis	1,672	
p-value (searah)	0,051	
Asumsi varians tidak sama		
t-hitung	1,661	
t-kritis	1,678	
p-value (searah)	0,051	

Setelah metode terbaik diperoleh, tahap berikutnya adalah memunculkan hasil *confusion matrix* dan struktur pohon untuk metode tersebut. Tujuannya adalah agar dapat dihasilkan alur identifikasi dari rasa makanan berdasarkan hasil dari metode terbaik.

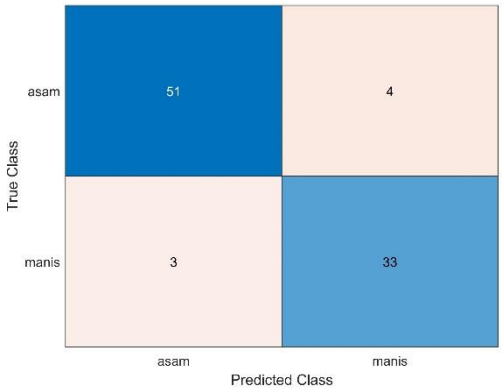
4. HASIL DAN PEMBAHASAN

Pada bagian ini akan dibahas performa dari metode k-CT beserta 3 metode berbasis pohon lainnya. Bahasan akan diawali dengan proses menentukan akurasi model prediktif, dan kemudian dilanjutkan dengan mengukur waktu yang dibutuhkan oleh setiap algoritma untuk menyelesaikan 1 kali proses.

Tabel 2. Perbandingan Kualitas model prediktif antara metode k-CT, CT, RF dan GB

parameter	k-CT	CT	RF-10	RF-50	RF-100	GB-10	GB-50	GB-100
Akurasi	92,31%	89,01%	90,11%	90,11%	91,21%	80,22%	89,01%	89,01%
Presisi	89,19%	82,50%	81,40%	84,62%	83,33%	66,67%	82,50%	82,50%
Recall	91,67%	91,67%	97,22%	91,67%	97,22%	100,00%	91,67%	91,67%
F1-Score	90,41%	86,84%	88,61%	88,00%	89,74%	80,00%	86,84%	86,84%
TPR	91,67%	91,67%	97,22%	91,67%	97,22%	100,00%	91,67%	91,67%
TNR	92,73%	87,27%	85,45%	89,09%	87,27%	67,27%	87,27%	87,27%
FPR	7,27%	12,73%	14,55%	10,91%	12,73%	32,73%	12,73%	12,73%
FNR	8,33%	8,33%	2,78%	8,33%	2,78%	0,00%	8,33%	8,33%

*hasil terbaik diantara semua model



Gambar 3. *Confusion Matrix* untuk metode k-CT

Berdasarkan pada proses *clustering* dengan $k = 2$ diperoleh bahwa *silhouette* maksimum untuk metode k-CT adalah sebesar 0,32. Gambar 2 menunjukan grafik *silhouette* untuk tiap sampel. Terlihat bahwa beberapa sampel pada klaster 1 sebenarnya tidak berada pada klaster yang benar (nilai *silhouette* negatif). Hal ini disebabkan bahwa sebenarnya jumlah klaster yang ideal untuk data ini bukanlah 2. Namun, untuk menunjukkan bahwa algoritma k-CT dapat diproses dengan kecepatan yang kurang lebih sama dengan metode CT, maka nilai *silhouette* ini dapat diabaikan.

Tabel 3. Perbandingan waktu komputasi antara metode k-CT, CT, RF dan GB dalam satuan sekon (s)

No.	k-RT	RT	RF-10	RF-50	RF-100	GB-10	GB-50	GB-100
1	0,0152	0,0152	0,0282	0,0898	0,1688	0,5688	0,5403	0,4476
2	0,0159	0,0148	0,0258	0,0935	0,1762	0,4557	0,4661	0,4415
3	0,0134	0,0165	0,0320	0,0899	0,1635	0,4739	0,4920	0,4651
4	0,0140	0,0142	0,0354	0,0948	0,1890	0,4372	0,7641	0,4434
5	0,0135	0,0145	0,0337	0,1086	0,2301	0,5471	0,8126	0,5061
6	0,0148	0,0155	0,0308	0,1133	0,2353	0,5141	0,8115	0,5335
7	0,0143	0,0141	0,0260	0,0966	0,2366	0,5616	0,5475	0,5245
8	0,0201	0,0189	0,0262	0,0885	0,2820	0,5046	0,5265	0,5285
9	0,0204	0,0163	0,0279	0,0864	0,2099	0,4982	0,5180	0,4962
10	0,0148	0,0176	0,0309	0,0968	0,1544	0,4664	0,4490	0,4598
11	0,0135	0,0169	0,0298	0,0872	0,1610	0,4324	0,4533	0,4559
12	0,0138	0,0149	0,0304	0,0889	0,1726	0,4335	0,4450	0,6333
13	0,0132	0,0148	0,0264	0,0868	0,2575	0,4254	0,4646	0,5195
14	0,0133	0,0154	0,0284	0,1192	0,1694	0,4582	0,4489	0,5095
15	0,0134	0,0148	0,0334	0,1307	0,1549	0,4350	0,4658	0,4804
16	0,0137	0,0144	0,0362	0,1126	0,1607	0,4293	0,4467	0,4794
17	0,0145	0,0148	0,0397	0,1191	0,1564	0,4407	0,4865	0,4710
18	0,0137	0,0144	0,0427	0,1112	0,1608	0,4413	0,5553	0,4661
19	0,0164	0,0145	0,0275	0,1265	0,1572	0,4359	0,5106	0,4573
20	0,0134	0,0146	0,0275	0,1025	0,2569	0,4333	0,4562	0,4721
21	0,0131	0,0143	0,0297	0,1242	0,1966	0,4373	0,4725	0,4614
22	0,0138	0,0154	0,0428	0,1117	0,2054	0,4455	0,4480	0,4738
23	0,0180	0,0145	0,0277	0,1437	0,1623	0,4389	0,4475	0,4554
24	0,0132	0,0154	0,0295	0,1254	0,1713	0,4283	0,4681	0,4397
25	0,0129	0,0146	0,0342	0,0848	0,1933	0,4285	0,4963	0,4963
26	0,0135	0,0141	0,0339	0,0859	0,1691	0,4529	0,4606	0,5058
27	0,0129	0,0149	0,0309	0,1102	0,1545	0,4807	0,4840	0,5201
28	0,0146	0,0148	0,0331	0,1388	0,1575	0,4620	0,4712	0,4484
29	0,0137	0,0152	0,0290	0,0895	0,1674	0,4659	0,4991	0,4703
30	0,0137	0,0145	0,0289	0,0930	0,2093	0,4291	0,4728	0,4413
Rerata	0,0145	0,0152	0,0313	0,1050	0,1880	0,4621	0,5127	0,4834

4.2. Kualitas Model Prediktif

Tabel 2 menunjukan hasil pengukuran dari 8 parameter kualitas untuk metode k-CT, CT, RF-10, RF-50, RF-100, GB-10, GB-50 dan GB-100. Berdasarkan pada Tabel 2, terlihat bahwa metode k-CT memperoleh hasil akurasi prediktif tertinggi jika dibandingkan dengan metode lainnya. Nilai akurasi ini bersaing dengan metode seperti RF-100 dengan perbedaan sekitar 1%. Hal ini menunjukan bahwa dengan menentukan klaster yang optimal, pemilihan sampel dapat menghasilkan model yang lebih handal (*robust*) dalam memprediksikan rasa dari suatu bahan makanan.

Dari 8 jenis parameter evaluasi model, terlihat bahwa metode k-CT unggul di 5 parameter jika dibandingkan dengan metode lainnya. Dalam kasus ini, metode k-CT memiliki performa yang lebih buruk dari metode GB-10 dalam pengukuran *recall*, TPR dan FNR. Namun, perlu diperhatikan bahwa untuk metode GB-10, terjadi permasalahan salah klasifikasi: model cenderung menghasilkan prediksi kelas positif. Hal ini dapat terlihat dari rendahnya kualitas akurasi dan TNR untuk metode GB-10.

Berdasarkan pada analisis di atas, dapat disimpulkan bahwa metode k-CT memiliki kemampuan mengklasifikasikan ras dari buah jeruk Pontianak lebih baik daripada metode CT, RF dan GB.

4.3. Perbandingan Waktu Komputasi

Tabel 3 menunjukan perbandingan waktu komputasi untuk menghasilkan model prediktif dari metode k-CT, CT, RF dan GB. Perlu diperhatikan bahwa pengukuran waktu ini murni hanya memperhitungkan waktu yang dibutuhkan untuk membangun model dan mengabaikan waktu untuk pra-pemrosesan data dan prediksi data.

Berdasarkan pada hasil dari Tabel 3, terlihat bahwa metode k-CT dan CT merupakan metode dengan rerata waktu komputasi tersingkat, jika dibandingkan dengan ke 6 model lainnya. Metode GB merupakan metode yang memerlukan waktu komputasi terlama, kemudian disusul oleh metode RF. Hal ini karena metode RF merupakan kumpulan dari CT, dan metode GB adalah RF yang dioptimasi, sehingga tidak mungkin bagi kedua metode tersebut untuk memiliki waktu komputasi yang lebih cepat daripada metode CT dan k-CT.

Hasil dari Tabel 3 menyatakan bahwa metode k-CT dan CT merupakan metode dengan waktu komputasi paling singkat, dengan metode k-CT sedikit lebih cepat jika dibandingkan dengan metode CT. Hal ini sesuai dengan penjabaran matematis yang telah dijabarkan pada Bagian 2, yang menunjukan bahwa pemilihan k sesuai dengan persamaan 5 dapat mempercepat atau memiliki waktu pemrosesan yang sama dengan metode CT. Untuk membuktikan hal ini, maka dilakukan uji T dengan sampel independen (dikarena CT dan k-CT memiliki himpunan data latih yang berbeda) dengan hipotesis sebagai berikut:

H_0: Metode k-CT dan CT memiliki waktu komputasi yang sama.

H_a: Metode k-CT memiliki waktu komputasi yang lebih cepat dari metode CT.

Hasil dari uji T dengan sampel independen ditunjukkan pada Tabel 4. Berdasarkan pada hasil uji T, terlihat bahwa perbedaan waktu komputasi dari kedua metode tidak signifkan pada $\alpha = 0{,}05$ namun signifikan pada $\alpha = 0.1$. Hal ini menunjukan bahwa belum terdapat bukti kuat untuk menolak H_0. Sehingga, diperlukan pengujian lanjutan dengan data yang lebih besar untuk membuktikan klaim waktu komputasi dari metode ini.

Secara keseluruhan, terlihat bahwa metode k-CT memiliki akurasi model prediktif yang lebih baik dari ke-3 metode lainnya. Lebih lanjut, terdapat bukti (lemah) bahwa metode k-CT dapat mempersingkat waktu komputasi dari metode CT. berdasarkan hasil ini, maka dapat disimpulkan bahwa metode k-CT merupakan metode terbaik dalam mengklasifikasikan rasa buah jeruk Pontianak.

4.3. Identifikasi Rasa Buah Jeruk Pontianak

Berdasarkan pada hasil analisa pada sub bagian sebelumnya, telah dibuktikan bahwa metode k-CT merupakan metode terbaik untuk mengklasifikasikan rasa dari buah jeruk Pontianak. Oleh karena itu, hasil pada sub bagian ini akan difokuskan hanya pada metode k-CT saja. Gambar 3 menunjukan *confusion*

Tabel 5. Klasifikasi rasa buah jeruk Pontianak berdasarkan pada metode k-CT

Kondisi	Warna Kulit	Kulit	Diameter	Bentuk	Ukuran Pori	Rasa
1	Hijau	Keriput	~	~	~	Asam
2	Hijau	Mulus	$x_6 < 4{,}50$	~	~	Manis
3	Hijau	Mulus	$4{,}50 \leq x_6 < 4{,}82$	~	~	Asam
4	Hijau	Mulus	$4{,}82 \leq x_6 < 4{,}90$	~	~	Asam
5	Hijau	Mulus	$4{,}90 \leq x_6 < 5{,}57$	~	~	Manis
6	Hijau	Mulus	$x_6 \geq 5{,}57$	~	~	Manis
7	Kuning	~	$x_6 < 4{,}67$	Bulat	~	Manis
8	Kuning	~	$x_6 < 4{,}67$	Pipih	~	Asam
9	Kuning	~	$4{,}67 \leq x_6 < 4{,}97$	~	Besar	Asam
10	Kuning	~	$4{,}67 \leq x_6 < 4{,}97$	~	Kecil	Manis
11	Kuning	~	$4{,}97 \leq x_6 < 5{,}55$	Bulat	~	Manis
12	Kuning	~	$4{,}97 \leq x_6 < 5{,}55$	Pipih	~	Asam
13	Kuning	~	$5{,}55 \leq x_6 < 5{,}85$	~	~	Asam
14	Kuning	~	$x_6 \geq 5{,}85$	~	~	Manis

matrix dari metode k-CT dalam mengklasifikasikan buah jeruk Pontianak. Dari Gambar 3, terlihat bahwa metode k-CT hanya salah mengklasifikasikan 7 dari 91 sampel yang ada.

Metode k-CT dengan $k = 2$ menghasilkan 2 buah klaster, sehingga terdapat 2 buah model. Model 1 memiliki *centroid* pada kondisi buah berwarna kuning, dengan ukuran pori kecil, berbentuk elips, dengan kulit bertekstur keriput dan tidak terdapat *thrips*. Sedangkan model 2 memiliki centroid pada kondisi buah berwarna hijau, berpori kecil, berbentuk bulat, dengan tekstur kulit halus dan memiliki *thrips*. Ringkasan kondisi jika-maka atau *if-then* dari metode k-CT ditampilkan pada Tabel 5. Tanda (~) menyatakan kondisi bebas, artinya dapat independen atau terbebas dari variabel tersebut.

Berdasarkan pada Tabel 5, terlihat bahwa ukuran jeruk Pontianak mempengaruhi rasa dari buah. Secara spesifik, buah dengan ukuran besar, yakni dengan diameter di atas 5,85 cm cenderung memiliki rasa manis. Lebih lanjut, buah dengan bentuk bulat cenderung memiliki rasa manis ketimbang dengan buah berbentuk pipih.

Dalam kasus ini, terlihat bahwa warna buah tidak menjadi indikator utama untuk menentukan rasa dari buah. Terdapat probabilitas sebesar 50% untuk buah berwarna hijau untuk memiliki rasa manis, begitu pula jika warnanya kuning. Ini menandakan bahwa sulit untuk memprediksikan rasa dari jeruk Pontianak hanya berdasarkan pada warna saja.

Secara keseluruhan, dapat disimpulkan bahwa rasa jeruk Pontianak lebih bergantung pada ukurannya. Secara spesifik, buah dengan ukuran yang besar cenderung memiliki rasa manis, sedangkan buah dengan ukuran kecil memiliki rasa yang dominan asam.

5. KESIMPULAN

Penelitian ini mengagas metode baru yang disebut dengan *k-cluster classification tree* atau k-CT. metode k-CT dirancang untuk mengurangi kompleksitas waktu dari metode CT tanpa menurunkan kualitas model prediktif dari metode CT. berdasarkan pada validasi model prediktif menggunakan data yang berasal dari pengamatan ciri fisik buah jeruk Pontianak, diketahui bahwa metode k-CT memiliki akurasi model prediktif sebesar 92,31% dalam mengklasifikasikan rasa dari buah jeruk, yang dalam penelitian yang ini K-CT memiliki akurasi yang lebih tinggi jika dibandingkan dengan metode CT, RF dan GB. Lebih lanjut, pada pengujian waktu komputasi, terdapat bukti lemah ($\alpha < 0{,}1$) bahwa metode k-CT memiliki waktu komputasi yang lebih cepat dari metode CT, sesuai dengan Lema yang diturunkan.

Hasil klasifikasi dari metode k-CT menunjukan bahwa jeruk Pontianak cenderung tidak bisa dibedakan rasanya dengan hanya melihat warna kulit. Namun, buah jeruk dengan ukuran besar cenderung memiliki rasa manis, ketimbang dengan buah dengan ukuran kecil.

Penelitian lanjutan dari metode ini dapat berfokuskan untuk memvalidasi waktu komputasi dari metode. Dalam kasus ini, dapat digunakan data yang lebih besar untuk melihat apakah hasil dari metode k-CT dapat di justifikasi untuk himpunan data yang lebih besar. Lebih lanjut, metode clustering juga dapat digabungkan dengan metode lain seperti GB untuk melihat potensinya dalam mempersingkat waktu komputasi dari metode GB. Lebih lanjut, penelitian lanjutan juga dapat meninjau kemampuan dari metode k-CT terutama kehandalan (robustness) dan konsistensi hasil model pada data yang berbeda.

UCAPAN TERIMA KASIH

Penulis ingin mengungkapkan terima kasih kepada mahasiswa program studi Informatika Universitas Widya Dharma Pontianak, terutama a.n. Candra yang telah bersedia untuk menjadi responden sekaligus mengkordinasikan pengambilan sampel data buah jeruk Pontianak.

DAFTAR PUSTAKA

AGHAEI, S., GÓMEZ, A., & VAYANOS, P., 2024. Strong Optimal Classification Trees. *Operations Research*.

AHMED, M., SERAJ, R. , & ISLAM, S.M.S., 2020. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8), p.1295.

AULIA, I., WIJAYA, D.R. , & HIDAYAT, W., 2021. Rice Quality Detection Using Gradient Tree Boosting Based On Electronic Nose Dataset. In: *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. IEEE. pp.1–5.

AVERÓS, X., LAVÍN, J.L., & ESTEVEZ, I., 2024. The potential of decision trees as a tool to simplify broiler chicken welfare assessments. *Scientific Reports*, 14(1), p.22943.

BHARGAVA, A., & BANSAL, A., 2020. Automatic Detection and Grading of Multiple Fruits by Machine Learning. *Food Analytical Methods*, 13(3), pp.751–761.

BHATTACHARYYA, S.K., & PAL, S., 2024. Design and performance analysis of decision tree learning model for classification of dry and cooked rice samples. *European Food Research and Technology*, 250(10), pp.2529–2544.

CHARBUTY, B., & ABDULAZEEZ, A., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2(01), pp.20–28.

DRAGOMIR, S.S., 2021. Superadditivity of convex integral transform for positive operators in Hilbert spaces. *Revista de la Real Academia*

de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas, 115(3), p.98.

FARAH, J.S., CAVALCANTI, R.N., GUIMARÃES, J.T., BALTHAZAR, C.F., COIMBRA, P.T., PIMENTEL, T.C., ESMERINO, E.A., DUARTE, M.C.K.H., FREITAS, M.Q., GRANATO, D., NETO, R.P.C., TAVARES, M.I.B., CALADO, V., SILVA, M.C., & CRUZ, A.G., 2021. Differential scanning calorimetry coupled with machine learning technique: An effective approach to determine the milk authenticity. *Food Control*, 121, p.107585.

HAND, D.J., 2020. *Artificial Intelligence Frontiers in Statistics*. Chapman and Hall/CRC.

HASSOUN, A., JAGTAP, S., GARCIA-GARCIA, G., TROLLMAN, H., PATEIRO, M., LORENZO, J.M., TRIF, M., RUSU, A.V., AADIL, R.M., ŠIMAT, V., CROPOTOVA, J., & CÂMARA, J.S., 2023. Food quality 4.0: From traditional approaches to digitalized automated analysis. *Journal of Food Engineering*, 337, p.111216.

HEMAMALINI, V., RAJARAJESWARI, S., NACHIYAPPAN, S., SAMBATH, M., DEVI, T., SINGH, B.K., & RAGHUVANSHI, A., 2022. Food Quality Inspection and Grading Using Efficient Image Segmentation and Machine Learning-Based System. *Journal of Food Quality*, 2022, pp.1–6.

IKOTUN, A.M., EZUGWU, A.E., ABUALIGAH, L., ABUHAIJA, B., & HEMING, J., 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, pp.178–210.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., & TAYLOR, J., 2023. Tree-Based Methods. pp.331–366.

JI, H., PU, D., YAN, W., ZHANG, Q., ZUO, M., & ZHANG, Y., 2023. Recent advances and application of machine learning in food flavor prediction and regulation. *Trends in Food Science & Technology*, 138, pp.738–751.

JIMÉNEZ-CARVELO, A.M., GONZÁLEZ-CASADO, A., BAGUR-GONZÁLEZ, M.G., & CUADROS-RODRÍGUEZ, L., 2019. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, 122, pp.25–39.

KONUR, S., LAN, Y., THAKKER, D., MORKYANI, G., POLOVINA, N., & SHARP, J., 2023. Towards design and implementation of Industry 4.0 for food manufacturing. *Neural Computing and Applications*, 35(33), pp.23753–23765.

MASCELLANI, A., HOCA, G., BABISZ, M., KRSKA, P., KLOUCEK, P., HAVLIK, & J., 2021. 1H NMR chemometric models for classification of Czech wine type and variety. *Food Chemistry*, 339, p.127852.

NILASHI, M., AHMADI, H., ARJI, G., ALSALEM, K.O., SAMAD, S., GHABBAN, F., ALZAHRANI, A.O., AHANI, A., & ALAROOD, A.A., 2021. Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. *Journal of Retailing and Consumer Services*, 62, p.102630.

ROJAS, C., BALLABIO, D., CONSONNI, V., SUÁREZ-ESTRELLA, D., & TODESCHINI, R., 2023. Classification-based machine learning approaches to predict the taste of molecules: A review. *Food Research International*, 171, p.113036.

SARKER, I.H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), p.160.

SINGH, A., VAIDYA, G., JAGOTA, V., DARKO, D.A., AGARWAL, R.K., DEBNATH, S., & POTRICH, E., 2022. Recent Advancement in Postharvest Loss Mitigation and Quality Management of Fruits and Vegetables Using Machine Learning Frameworks. *Journal of Food Quality*, 2022, pp.1–9.

SMARRA, F., TJEN, J., & D'INNOCENZO, A., 2022. Learning methods for structural damage detection via entropy-based sensors selection. *International Journal of Robust and Nonlinear Control*, 32(10), pp.6035–6067.

TJEN, J., 2024. Identifikasi Parameter Kualitas Bahan Pangan dengan Metode Entropy-Based Subset Selection (E-SS) (Studi Kasus: Minuman Anggur). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 11(1), pp.47–54.

ZENG, X., CAO, R., XI, Y., LI, X., YU, M., ZHAO, J., CHENG, J., & LI, J., 2023. Food flavor analysis 4.0: A cross-domain application of machine learning. *Trends in Food Science & Technology*, 138, pp.116–125.