Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.20241116850
Vol. 11, No. 1, Februari 2024, hlm. 47-54 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

IDENTIFIKASI PARAMETER KUALITAS BAHAN PANGAN DENGAN METODE *ENTROPY-BASED SUBSET SELECTION* (E-SS) (STUDI KASUS: MINUMAN ANGGUR)

Jimmy Tjen[*1]

[1]Jurusan Informatika Universitas Widya Dharma Pontianak
Email: [1]jimmy.tjen@mathmods.eu
[*]Penulis Korespondensi

(Naskah masuk: 27 Desember 2022, diterima untuk diterbitkan: 6 Februari 2024)

Abstrak

Penelitian ini bertujuan untuk membangun sebuah algoritma yang dapat mengidentifikasi parameter dari bahan pangan yang mempengaruhi kualitas dari bahan makanan tersebut menggunakan algoritma pemilihan himpunan bagian berbasis entropi dan metode pohon klasifikasi dari pembelajaran pohon keputusan. Metode pemilihan himpunan bagian berbasis entropi secara khusus merupakan sebuah algoritma yang bertujuan untuk memilih sekumpulan dari parameter yang memiliki hubungan entropi yang baik satu sama lain, sehingga dapat menghasilkan model prediktif yang optimal. Untuk memvalidasi performa dari algoritma yang digagas, penelitian ini mengambil sampel dari minuman anggur merah dan putih yang berasal dari negara Portugal. Berdasarkan pada percobaan yang telah dilakukan, diperoleh hasil bahwa algoritma yang digagas dapat memprediksi kualitas dari anggur putih dengan akurasi hingga 97,8 % dan 96,25% untuk kualitas anggur merah. Dimana, nilai ini lebih tinggi dari metode pohon klasifikasi klasik, dan algortima yang digagas hanya membutuhkan jumlah parameter yang lebih sedikit (hanya 2 hingga 5 dari total 11 parameter input yang ada) jika dibandingkan dengan metode klasik. Lebih lanjut, berdasarkan pada percobaan yang telah dilakukan, diperoleh temuan bahwa parameter yang paling menentukan kualitas dari anggur putih adalah tingkat keasaman, kadar alkohol, pH dan kandungan klorit. Sedangkan untuk anggur merah, kualitas secara dominan ditentukan oleh kandungan sisa gula, densitas minuman dan kandungan dari sulfur oksida.

Kata kunci: *Identifikasi Parameter, Pembelajaran Mesin, Pemilihan Sub-himpunan, Pohon Klasifikasi*

FOOD QUALITY PARAMETER IDENTIFICATION WITH THE ENTROPY-BASED SUBSET SELECTION (E-SS) (CASE STUDY: WINE)

Abstract

This research aims to build an algorithm which is able to identify parameters determining the quality of food based on Entropy-based Subset Selection (E-ss) and classification tree algorithm from the decision tree learning process. E-ss is an algorithm that selects a set of parameters where each parameter contained in the set maintains a strong entropy relation with each other that allow them to produce an optimal predictive model. In this research, the dataset from the red wine and white wine produced in Portugal is used to validate the algorithm's predictive capability. Based on the experiment, the proposed algorithm is able to determine the quality of wine with an accuracy of 97.8% and 96,25% for both white and red wine, respectively. This value is higher than the classical classification tree algorithm and potentially requires fewer parameters than the classical algorithm (2 up to 5 instead of 11). Furthermore, it was discovered that there are 4 parameters that significantly affect the white wine quality: acidity, alcohol percentage, pH, and chloride concentration. On the other hand, the red wine the quality is mostly driven by the residual sugar, density, and sulfur oxide concentration.

Keywords: *Parameter Identification, Machine Learning, Subset Selection, Classification Tree*

1. PENDAHULUAN

Kebutuhan akan pangan merupakan salah satu dari kebutuhan primer dari manusia. Sebab, makanan yang dikonsumsi akan menjadi sumber energi bagi manusia untuk menjalankan aktivitasnya (Rai, et al., 2019). Oleh karena itu, penting untuk memastikan bahwa makanan yang masuk ke tubuh memiliki kualitas yang baik, agar dapat mengoptimalkan kinerja dan aktivitas manusia, serta demi menjaga kesehatan dari tubuh manusia.

Dewasa ini, setiap makanan yang diproduksi lazimnya akan menampilkan informasi gizi dan bahan baku penyusun makanan tersebut. Informasi ini penting, sebagai contoh sangat dimungkinkan untuk dibangun sebuah persamaan matematis yang dapat mengidentifikasi kualitas bahan pangan berdasarkan pada komposisi penyusunnya. Permasalahan ini secara teknis masuk ke dalam permasalahan identifikasi parameter (*parameter identification*) dan klasifikasi berdasarkan pada proses pembelajaran mesin (*machine learning*) (Saravanan & Sujatha, 2018), (Mosavi, et al., 2019), (Gomes, et al., 2019). Salah satu metode digunakan dalam permasalahan klasifikasi adalah metode pohon klasifikasi atau *classification tree* (Breiman, et al., 2017; Linero, 2018; Bertsimas, et al., 2021).

Metode pohon klasifikasi adalah sebuah metode yang berdasarkan pada pembelajaran pohon keputusan (*decision tree learning*) yang memetakan sekumpulan data berdasarkan pada kemiripan dari data satu sama lain. Serupa dengan metode pohon regresi (*regression tree*) yang berfokuskan pada output yang bersifat kontinu (berupa bilangan riil), metode pohon klasifikasi berfokuskan pada output yang bersifat diskrit, seperti informasi kualitas bahan atau pun kondisi rusak (*faulty*) atau tidak (*nominal*) dari sebuah objek. Metode ini telah digunakan digunakan diberbagai bidang, seperti pada pendeteksian kerusakan struktur, kualitas bahan bangunan, prediksi konsumsi energi listrik, dan bidang lainnya seperti pada (Smarra, et al., 2022), (Panjaitan, et al., 2022), (Ren, et al., 2021), (Supriyadi, et al., 2020), (Jiang, et al., 2021).

Keterbaharuan topik: (Smarra, et al., 2022) memperkenalkan konsep pemilihan himpunan bagian data berbasiskan pada pengukuran entropi (*Entropy-based subset selection* atau E-ss) dari teori informasi (*information theory*) yang digabungkan dengan metode pohon regresi untuk memprediksi kerusakan struktur bangunan. Smarra pada penelitiannya menunjukan bahwa himpunan bagian data sensor yang dibuat dapat meningkatkan akurasi dari pendeteksian kerusakan struktur, serta adanya potensi lokalisasi titik kerusakan dari struktur. (Panjaitan, et al., 2022) menggunakan metode hutan acak (*random forest*) yang merupakan sekumpulan dari pohon regresi yang dikombinasikan dengan metode *Kalman filter* dan model *poly-exponential* untuk memprediksi kualitas dan konsumsi energi listrik dari sebuah gedung. (Ren, et al., 2021) memodelkan kekuatan dari beton yang dibuat dari pasir sintetis dengan menggunakan metode pohon klasifikasi dan regresi. (Supriyadi, et al., 2020) membandingkan metode pohon keputusan dan *Support Vector Machine* (SVM) untuk memprediksi kualitas dari minuman anggur merah (*red wine*), dimana Supriyadi menunjukan bahwa metode pohon keputusan lebih akurat dalam mengklasifikasikan kualitas dari minuman dengan akurasi 75% jika dibandingkan dengan SVM yang hanya 65%. (Jiang, et al., 2021) menggunakan algoritma pohon keputusan yang dikombinasikan dengan *particle swarm optimization* untuk mengidentifikasi kerusakan dari sensor piezo elektrik, dimana Jiang menyatakan bahwa metode yang diusulkan mampu mendeteksi kerusakan sensor dengan tingkat akurasi mencapai 98%.

Meskipun metode pohon regresi dan klasifikasi mudah diterapkan dan memiliki presisi yang tinggi, namun tetap saja algoritma ini memiliki kelemahan. Secara umum, metode ini memiliki kompleksitas waktu (*time complexity*) yang tergolong berat, yakni pada orde $O(m \cdot n^2)$ dengan m menyatakan jumlah sampel yang terkandung dalam himpunan data dan n^2 menyatakan banyaknya parameter yang terkandung di dalam data (More & Rana, 2017). Oleh karena itu, penggunaan data yang massif akan memberatkan pekerjaan dari metode tersebut.

Kontribusi: berangkat dari permasalahan dan pemaparan ide di atas, maka penelitian ini disusun secara khusus untuk menghasilkan sebuah algoritma baru berbasiskan pada pemilihan himpunan bagian berbasis entropi (E-ss) dan metode pohon klasifikasi untuk mengidentifikasi parameter yang mempengaruhi kualitas dari bahan pangan, yang dalam kasus ini diwakilkan oleh minuman anggur merah (*red wine*) dan putih (*white wine*). Berdasarkan tujuan tersebut maka yang menjadi kontribusi utama dari penelitian ini adalah sebagai berikut:

1. Untuk merumuskan sebuah algoritma baru untuk memprediksi kualitas dari bahan makanan dengan menggunakan pohon klasifikasi dari pembelajaran pohon keputusan dan E-ss.
2. Untuk mengidentifikasi parameter apa saja yang mempengaruhi kualitas dari anggur merah dan anggur putih dengan menggunakan metode E-ss.

Artikel ini bertujuan untuk menghasilkan sebuah algoritma baru yang dapat mengidentifikasi parameter yang dapat mengidentifikasi kualitas bahan pangan melalui modifikasi pada metode E-ss. Secara spesifik, penelitian ini akan merujuk pada 2 rujukan utama yang ditulis oleh (Smarra, et al., 2022) dan (Supriyadi, et al., 2020). Terkait dengan penelitian yang telah dilakukan oleh Smarra, penelitian ini mengadopsi E-ss yang telah dipaparkan pertama kali dalam (Tjen, et al., 2020) dan disempurnakan dalam (Smarra, et al., 2022). Secara khusus, penelitian ini tidak hanya berfokuskan pada menentukan model prediktif yang dapat memprediksi kualitas dari minuman anggur saja, tetapi pada identifikasi parameter yang diduga mempengaruhi kualitas dari anggur merah dan putih. Hal ini sejalan dengan target penelitian lanjutan yang ingin dicapai oleh Smarra, dimana pada penelitiannya, telah ditunjukan potensi lokalisasi kerusakan dari sebuah struktur bangunan. Penelitian ini akan berfokuskan pada bagaimana cara menerjemahkan informasi ini sebagai informasi parameter yang mempengaruhi kualitas minuman anggur. Lebih lanjut, terhadap

penelitian yang telah dilakukan oleh (Supriyadi, et al., 2020), pada penelitian ini, akan dibahas bagaimana metode E-ss dapat meningkatkan tingkat akurasi dari metode pohon keputusan sesuai dengan data yang digunakan pada penelitian tersebut.

Artikel ini disusun dengan formulasi sebagai berikut: Bab II akan membahas metode E-ss secara sekilas dan himpunan data yang digunakan dalam penelitian ini. Bab III akan membahas bagaimana cara mengidentifikasi parameter yang mempengaruhi kualitas dari minuman anggur merah dan putih. Beserta dengan besaran yang akan digunakan untuk mengidentifikasi akurasi dari model prediktif yang telah dibangun. Bab IV akan menampilkan hasil dari simulasi, serta membahas faktor yang mempengaruhi kualitas dari minuman anggur. Terakhir, pada bab V akan dibahas kesimpulan dan saran akan penelitian lanjutan.

2. METODOLOGI

Pada Bab ini akan dibahas secara sekilas konsep dari pemilihan himpunan bagian data berdasarkan pada entropi (E-ss) dan himpunan data yang digunakan untuk memvalidasi algoritma yang digagas. Terkait dengan himpunan data yang digunakan, data diperoleh dari (Paulo, et al., 2009) yang telah melakukan penelitian serupa sebelum dilakukan pula oleh (Supriyadi, et al., 2020). Untuk mempermudah memahami bahasan pada bagian ini, disarankan kepada pembaca untuk merujuk pada (Tjen, et al., 2020) dan (Smarra, et al., 2022) terkait dengan konsep dasar dari entropi dan pembuktian formal dari pemilihan himpunan bagian berbasis entropi.

2.1 *Entropy-based Subset Selection (E-ss)*

Bahasan pada sub bagian ini merupakan versi ringkas (tanpa pembuktian) dari algoritma pemilihan himpunan bagian data berbasis entropi yang digagas pada penelitian yang dilakukan oleh (Tjen, et al., 2020) dan (Smarra, et al., 2022). Oleh karena itu, disarankan terlebih dahulu kepada pembaca, untuk dapat melihat pada 2 referensi di atas terkait pembuktian formal dari teorema yang akan digunakan pada bagian ini.

Dimisalkan terdapat sebuah himpunan data $X \in \mathbb{R}^{m\times n}$; $X = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \boldsymbol{x_3}\ \dots \boldsymbol{x_n}]$ dengan m menyatakan banyaknya sampel yang terkandung, dan n menyatakan banyaknya parameter yang terkandung di dalam X sehingga $m > n$, dengan $x_i = [\,x_i(1)\ x_i(2)\ \dots\ x_i(m)]^\intercal\ i = 1,2,3,\dots,n$ menyatakan parameter ke$-i$ dari himpunan data X. Misalkan bahwa akan dibentuk sebuah himpunan bagian data berbasiskan pada entropi untuk parameter $\boldsymbol{x_j}$ dengan $\boldsymbol{x_j} \in X$. Maka, pertama tama, didefinisikan sebuah variabel acak $Z_{ij}, \forall i \neq j$ sebagai berikut

$$Z_{ij} = \begin{cases} 1 \text{ jika } |\boldsymbol{x_i} - \beta_{ij}\boldsymbol{x_j}| \leq \theta\sigma(\boldsymbol{x_i}) \\ 0 \text{ jika lainnya} \end{cases}. \quad (1)$$

Dengan $\theta \in [1, +\infty)$, β_{ij} adalah model yang dihasilkan dari metode kuadrat terkecil antara $\boldsymbol{x_i}$ dan $\boldsymbol{x_j}$ dan $\sigma(\boldsymbol{x_i})$ menyatakan standar deviasi dari $\boldsymbol{x_i}$. Dari persamaan (1) dapat didefinisikan nilai dari *information gain* untuk pasangan $\boldsymbol{x_i}$ dan $\boldsymbol{x_j}$ sebagai berikut

$$I(Z_{ij}; \beta_{ij}) = H(Z_{ij}) - H(Z_{ij}|\beta_{ij}). \quad (2)$$

Dengan $I(Z_{ij}; \beta_{ij})$ menyatakan *information gain* dari variabel acak Z_{ij} dan β_{ij}, $H(Z_{ij})$ menyatakan entropi dari variabel Z_{ij} dan $H(Z_{ij}|\beta_{ij})$ menyatakan entropi dari Z_{ij} yang dikondisikan terhadap β_{ij}. Pada tahap ini, komponen dari i dapat dipilih dengan cara menentukan pasangan i dan j sehingga $H(Z_{ij}|\beta_{ij})$ bernilai minimum. Hal ini dapat dilakukan karena pada hakekatnya $H(Z_{ij})$ dapat menempati nilai apa saja dan tidak bisa ditentukan dengan pasti. Oleh karena itu, agar *information gain* menjadi maksimum, nilai dari $H(Z_{ij}|\beta_{ij})$ haruslah sekecil mungkin. Namun, nilai pasti dari $H(Z_{ij}|\beta_{ij})$ sulit ditentukan secara analitik, sehingga diperlukan hampiran untuk menyelesaikan permasalahan tersebut. Adapun manipulasi matematis yang dibutuhkan secara umum dapat dijelaskan dengan menggunakan teorema 1

Teorema 1
Misalkan bahwa
$\widehat{H}(Z_{ij}|\beta_{ij}) = -[p \cdot \log p + (1-p) \cdot \log(1-p)$;
dengan
$$p = 1 - \theta^2 + \theta^2 . r^2 \quad (3)$$
dan $r^2 \in [1 - \frac{\theta^2}{2}, 1]$ *menyatakan koefisien determinasi dari* $\boldsymbol{x_i}$ *dan* $\boldsymbol{x_j}$. *Maka akan berlaku hubungan berikut*
$$H(Z_{ij}|\beta_{ij}) \leq \widehat{H}(Z_{ij}|\beta_{ij}). \quad (4)$$

Teorema 1 menjelaskan secara spesifik bahwa perhitungan pada persamaan (2) dapat secara spesifik dengan menentukan estimasi dari entropi kondisional dari Z_{ij} yang hanya bergantung pada koefisien determinasi (lihat: (Zhang, 2017)), yang dimana nilai dari koefisien ini dapat ditentukan secara mudah. Keseluruhan proses dari penentuan himpunan bagian data berbasis entropi dinyatakan dalam algoritma 1.

Algoritma 1: Penentuan himpunan bagian berbasis entropi (E-ss)
Masukan: himpunan data $X \in \mathbb{R}^{m\times n}$, variabel yang ingin diprediksi $j \in n$, kardinalitas dari himpunan bagian $n^* < n$.
Keluaran: kumpulan indeks $S \subset n, |S| = n^*$
Proses:
$S \coloneqq \{j\}$
Untuk $k = 1 : n^* - 1$ kerjakan
  $j^* = \underset{j\in n\setminus S}{\operatorname{argmin}} \widehat{H}(Z_{ij}|\beta_{ij})$
  $S = S \cup j^*$
Selesai

Dari segi efisiensi, algoritma 1 memiliki kompleksitas waktu sebesar $O(mn^2)$. Namun perlu

diperhatikan bahwa n dalam kasus ini adalah ukuran dari himpunan bagian yang diingikan. Sehingga, semakin kecil himpunan bagian yang ingin dibentuk, semakin cepat himpunan bagian data dapat terbentuk.

2.2 Informasi Data Minuman Anggur

Data yang digunakan merupakan data yang diambil dari basis data gratis yang disediakan oleh University of California, Irvine (UCI) (Paulo, et al., 2009). Data yang diambil memiliki 2 himpunan data: data anggur merah (*red wine*) dan anggur putih (*white wine*) yang berasal dari anggur jenis *vinho verde* yang lazimnya tumbuh di bagian utara Portugal.

Untuk setiap varian anggur (merah dan putih), terdapat 4.898 sampel dan 11 parameter input yang terdiri atas: tingkat keasaman tetap (*fixed acidity*), tingkat keasaman yang dapat diuapkan (*volatile acidity*), kandungan asam sitrat (*citric acid*), kandungan gula tersisa (*residual sugar*), kandungan klorit (*chlorides*), kandungan sulfur oksida bebas (*free sulfur oxides*), kandungan total dari sulfur oksida (*total sulfur oxides*), kerapatan (*density*), derajat keasamaan (pH), kandungan sulfat (*sulphates*), kadar alkohol (*alcohol*) dan sebuah output yang berupa nilai kualitas dari minuman anggur tersebut (dari 0 hingga 10).

Untuk melakukan simulasi, maka sebanyak 2.449 sampel atau 50% dari data yang ada akan digunakan sebagai data latih (*train dataset*) dan 50% sisanya akan digunakan sebagai data uji (*test dataset*) untuk memvalidasi akurasi dari model yang dibangun. Untuk mempermudah kasus, sesuai dengan metodologi yang digunakan pada (Paulo, et al., 2009) maka parameter kualitas minuman anggur akan dikelompokan dalam 2 rumpun utama: minuman anggur dikatakan memiliki kualitas buruk apabila kualitasnya berada pada rentang 0 hingga 5 dan dikatakan memiliki kualitas baik apabila berada pada rentang 9 hingga 10. Untuk mempermudah penyebutan nama dari parameter input yang terkandung di dalam data, dalam penelitian ini seluruh parameter input akan diasosiasikan dengan variabel berikut:

3. IDENTIFIKASI PARAMETER MINUMAN ANGGUR BERBASIS ENTROPI

Pada bagian ini, akan dibahas bagaimana metode pemilihan himpunan bagian berbasis entropi digunakan untuk mengkatagorikan parameter yang dianggap signifikan dalam menentukan kualitas minuman anggur. Seperti halnya pada Bab II, langkah awal adalah membuat definisi matematis formal terkait dengan data masukan yang ada.

Misalkan $X = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots \boldsymbol{x_n}]$; $X \in \mathbb{R}^{m_t \times n}$ adalah himpunan data parameter input baik dari minuman anggur merah atau anggur putih, dengan m_t menyatakan sampel dari data latih untuk minuman tersebut. Maka untuk setiap $i = 1,2, \dots n = 11$ algoritma 1 akan diulangi sebanyak i kali untuk menentukan himpunan bagian yang bersesuaian dengan parameter x_i. Himpunan bagian data yang terbentuk kemudian akan digunakan untuk membentuk pohon biner terhadap parameter output yang merupakan kualitas dari minuman anggur. Secara spesifik, jika $\boldsymbol{y} = [y(1)\ y(2) \dots y(m_t)]^{\mathrm{T}}$ menyatakan kualitas dari anggur merah atau anggur putih, maka fungsi pohon biner yang akan dibentuk dinyatakan dalam persamaan (5)

$$y = f_{pk}(\{x_i | i = 1,2, \cdots n;\ i \subset S_i\}. \quad (5)$$

dengan S_i menyatakan himpunan bagian indeks yang memiliki hubungan entropi yang "baik" dengan variabel i dan F_{pk} menyatakan fungsi pohon klasifikasi.

Tabel 1. Representasi parameter sebagai variabel

Nama parameter	Representasi variabel
Fixed acidity	$\boldsymbol{x_1}$
Volatile acidity	$\boldsymbol{x_2}$
Citric acid	$\boldsymbol{x_3}$
Residual sugar	$\boldsymbol{x_4}$
Chlorides	$\boldsymbol{x_5}$
Free sulfur oxides	$\boldsymbol{x_6}$
Total sulfur oxides	$\boldsymbol{x_7}$
Density	$\boldsymbol{x_8}$
pH	$\boldsymbol{x_9}$
Sulfates	$\boldsymbol{x_{10}}$
Alcohol	$\boldsymbol{x_{11}}$

3.1 Identifikasi Parameter

Himpunan bagian yang dipaparkan di atas hanya dapat mengidentifikasi kualitas dari minuman anggur, namun tidak dapat menunjukan parameter yang memiliki hubungan signifikan terhadap parameter output. Oleh karena itu, dibutuhkan modifikasi agar himpunan bagian yang dihasilkan dapat mengkarakterisasi parameter yang ideal untuk mengukur kualitas dari minuman anggur. Adapun modifikasi yang diberikan didasarkan pada nilai dari koefisien determinasi r_{ij}^2.

Seperti yang diketahui dari pembahasan korelasi linear (lihat: (Edelmann, et al., 2021)) bahwa koefisien determinasi yang merupakan kuadrat dari koefisien korelasi bercerita mengenai seberapa banyak informasi yang dapat diperkirakan dari variabel terikat oleh variabel bebas. Secara spesifik, semakin kecil nilai dari koefisien determinasi, semakin independen hubungan dari dua buah variabel, dan begitu pula sebaliknya. Terkait dengan fakta ini, maka untuk menentukan parameter yang memiliki signifikansi tinggi, maka algoritma 1 akan diberikan kondisi tambahan, yakni koefisien determinasi minimum yang dibutuhkan agar variabel pasangan dapat dimasukan sebagai kandidat dalam menentukan himpunan bagian data. Sebagai contoh, jika terdapat 2 buah parameter x_a dan x_b yang hendak dipasangkan dengan variabel x_i, maka relatif terhadap algoritma 1 akan ditambahkan kondisi tambahan yakni jika $r_{ai}^2 < r_{min}^2$ dengan r_{min}^2 menyatakan koefisien determinasi minimum, maka

variabel a tidak bisa dijadikan kandidat untuk himpunan bagian i, dan begitupula sebaliknya.

Secara teknis, nilai dari r_{min}^2 akan bergantung pada pengguna. Namun, perlu diperhatikan adalah r_{min}^2 yang terlalu tinggi akan mengakibatkan himpunan bagian data menjadi kosong, akibat tidak ada variabel yang mampu memenuhi kriteria tersebut, sedangkan r_{min}^2 yang terlalu rendah tidak akan mengakibatkan perbedaan pada metode yang digagas pada algoritma 1. Secara spesifik, berdasarkan pada proses heuristik "*trial and error*", ditentukan bahwa nilai r_{min}^2 yang cocok untuk identifikasi parameter input dari minuman anggur adalah sebesar 0,3. Algoritma 2 menyatakan hasil modifikasi dari algoritma 1 yang secara spesifik ditujukan untuk mengidentifikasi parameter input yang sensitif dalam mendeteksi kualitas minuman anggur.

Algoritma 2: Identifikasi parameter berbasis E-ss
Masukan: himpunan data $X \in \mathbb{R}^{m\times n}$, variabel yang ingin diprediksi $j \in n$, kardinalitas dari himpunan bagian $n^* < n$.
Keluaran: kumpulan indeks $S \subset n, |S| = n^*$
Proses:
$S \coloneqq \{j\}$
$n_s = \{i \subset n | r_{ij}^2 \geq r_{min}^2\}$
Untuk $k = 1: \min(n^* - 1, n_s)$ kerjakan
  $j^* = \underset{j\in n_s \setminus S}{\operatorname{argmin}} \hat{H}(Z_{ij}|\beta_{ij})$
  $S = S \cup j^*$
Selesai

4. HASIL DAN PEMBAHASAN

Pada Bab ini, akan ditampilkan kualitas dari model prediktif dan identifikasi parameter kualitas bahan pangan yang berhubungan secara signifikan terhadap kualitas dari minuman anggur merah maupun anggur putih. Secara spesifik, terdapat 6 parameter yang digunakan, yakni *True Positive Rate* (TPR), *True Negative Rate* (TNR), *Positive Predictive Value* (PPV), *Negative Predictive Value* (NPV), akurasi (%A) dan F_1 *score* (lihat: (Hussain, 2018), (Yao, et al., 2020)) sesuai dengan persamaan (6a - 6f)

$$TPR = \frac{TP}{TP+FN} \quad (6a)$$

$$TNR = \frac{TN}{TN+FP} \quad (6b)$$

$$PPV = \frac{TP}{TP+FP} \quad (6c)$$

$$NPV = \frac{TN}{TN+FN} \quad (6d)$$

$$\%A = \frac{TP+TN}{TP+TN+FP+FN} \quad (6e)$$

$$F_1 = \frac{2TP}{2TP+FP+FN} \quad (6f)$$

Dengan TP, TN, FN dan FP secara berurutan menyatakan menyatakan jumlah dari *true positive*, *true negative*, *false positive* dan *false negative*. Sebagai pembanding, algoritma juga akan dijalankan bersama dengan metode pohon klasik, sesuai pada penelitian yang dilakukan pada (Supriyadi, et al., 2020).

Terkait dengan kardinalitas atau jumlah elemen dari setiap himpunan bagian, berdasarkan pada proses uji coba, ditentukan bahwa $n^* = 5$. Nilai ini didasarkan pada pertimbangan jumlah parameter total yang ada, dan rerata dari nilai koefisien determinasi untuk setiap pasang parameter yang ada.

4.1 Akurasi Model Prediktif

Tabel 2 dan 3 menunjukan nilai akurasi dari model prediktif untuk minuman anggur putih (*white wine*) dan anggur merah (*red wine*) dari metode pohon klasifikasi klasik dan pohon klasifikasi berbasiskan pada algoritma E-ss yang telah dimodifikasi, sesuai dengan algoritma 2.

Tabel 2. Akurasi model prediktif dari anggur putih

Metode {himpunan bagian}	TPR	TNR	PPV	NPV	%A	F_1
klasik	0,89	0,25	0,97	0,5	87,2	0,93
E-ss $\{x_1, x_9\}$	0.98	0,71	0,99	0,47	97,4	0,99
E-ss $\{x_4, x_8, x_6, x_7, x_{11}\}$	0,96	0,54	0,98	0,27	95,0	0,97
E-ss $\{x_5, x_{11}\}$	0,98	0,53	0,98	0,50	97,8	0,99
E-ss $\{x_6, x_7, x_4, x_8\}$	0,97	0,68	0,99	0,39	96,1	0,98
E-ss $\{x_4, x_8, x_5, x_7, x_{11}\}$	0,95	0,78	0,99	0,31	95,2	0,97

*Warna merah menandakan model terbaik secara keseluruhan

Berdasarkan pada Tabel 2 dan Tabel 3 terlihat bahwa metode E-ss meningkatkan akurasi model prediktif dalam memprediksi kualitas dari anggur putih dan merah (relatif terhadap metode klasik) hingga 6%. Secara spesifik, tidak terdapat kasus dimana himpunan bagian data yang terbentuk dari algoritma 2 memiliki nilai akurasi yang dibawah metode pohon klasifikasi klasik untuk jenis minuman anggur merah dan anggur putih. Hal ini menunjukan bahwa model yang dibentuk mampu mengklasifikasikan kualitas dari minuman anggur merah dan putih lebih baik jika dibandingkan dengan metode klasik.

Tabel 3. Akurasi model prediktif dari anggur merah

Metode {himpunan bagian}	TPR	TNR	PPV	NPV	%A	F_1
klasik	0,97	0,14	0,96	0,19	93,62	0,97
E-ss $\{x_1, x_3, x_8, x_9\}$	0,99	0,23	0,97	0,47	95,49	0,98
E-ss $\{x_2, x_3\}$	0,99	0,03	0,96	0,13	94,87	0,97
E-ss $\{x_3, x_1, x_2, x_8, x_9\}$	0,99	0,11	0,96	0,31	94,99	0,97
E-ss $\{x_4, x_8\}$	0,97	0,63	0,98	0,50	95,62	0,98
E-ss $\{x_5, x_{10}\}$	0,99	0,29	0,97	0,71	96,37	0,98
E-ss $\{x_6, x_7\}$	0,97	0,71	0,99	0,56	96,25	0,98

E-ss $\{x_8, x_1, x_4, x_{11}, x_9\}$	0,98	0,14	0,96	0,25	94,37	0,97
E-ss $\{x_9, x_1, x_8, x_3\}$	0,98	0,31	0,97	0,44	95,24	0,98
E-ss $\{x_8, x_{11}\}$	0,98	0,17	0,96	0,30	94,62	0,97

*Warna biru menandakan model terbaik secara keseluruhan

Akurasi maksimum yang dapat dihasilkan untuk minuman anggur merah adalah 96,25%, sedangkan untuk anggur putih adalah sebesar 97,8%. Nilai ini tergolong tinggi, menimbang jumlah data yang dibutuhkan lebih sedikit daripada metode klasik (hanya 2 sampai 5 parameter yang digunakan dari keseluruhan 11 parameter). Hal ini selaras dengan pemaparan dari (Smarra, et al., 2022) yang menjelaskan bahwa metode E-ss dapat meningkatkan akurasi dari model prediktif, karena secara tidak langsung metode E-ss dapat memilih sekelompok parameter yang memiliki linearitas yang baik satu sama lain.

Berdasarkan pada percobaan yang telah dilakukan, dapat terlihat bahwa metode E-ss yang telah dimodifikasi seperti pada algoritma 2, mampu meningkatkan kualitas model prediktif dari metode pohon klasifikasi klasik seperti yang digunakan pada penelitian yang dilakukan oleh (Supriyadi, et al., 2020). Hal ini menunjukan bahwa model E-ss dapat mengeliminasi permasalahan linearitas yang merupakan salah satu alasan kenapa metode prediktif seperti regresi dan pohon klasifikasi tidak dapat memberikan hasil prediksi yang optimal (Heinze, et al., 2018), (Wang, et al., 2019), (Rath, et al., 2020).

4.2 Identifikasi Parameter Signifikan

Pada Tabel 2 terlihat bahwa parameter yang signifikan dalam menentukan kualitas minuman (akurasi, TPR dan TNR tertinggi relatif terhadap model lain) anggur putih adalah pasangan $\boldsymbol{x_1}$ dan $\boldsymbol{x_9}$ yang merupakan parameter tingkat keasaman tetap (*fixed acidity*) dan derajat keasaman atau pH serta $\boldsymbol{x_5}$ dan $\boldsymbol{x_{11}}$ yang merupakan representasi dari kandungan klorit dan kadar alkohol. *Potential Hydrogen* atau pH atau derajat keasamaan merupakan ukuran seberapa basa atau asam suatu zat. Semakin rendah pH dari suatu zat, semakin bersifat asam zat tersebut. Dalam kasus anggur putih, parameter tingkat keasaman tetap dan pH merupakan pasangan parameter yang signifikan dalam memprediksi kualitas dari minuman anggur. Hal ini dirasa masuk akal, jika menimbang hubungan antara derajat keasaman dan rasa asam itu sendiri. Hal ini selaras dengan penelitian yang telah dilakukan oleh (Vilela, 2019) yang menjelaskan hubungan dari rasa asam itu sendiri terhadap kualitas dari minuman anggur.

Terkait dengan kandungan klorit dan alkohol, terdapat penelitian yang menjelaskan bahwa kadar klorit yang tinggi pada minuman anggur mengakibatkan adanya sensasi rasa asin akibat pembentukan senyawa NaCl di dalam minuman tersebut (Zhou, et al., 2020; Mostashari, et al., 2022). Hal ini menunjukan bahwa algoritma yang dibangun dapat secara jelas menentukan faktor yang mempengaruhi kualitas dari anggur putih.

Dari Tabel 3, terlihat bahwa 2 pasangan parameter yang memiliki akurasi tertinggi terhadap kualitas anggur merah adalah $\boldsymbol{x_4}$ dan $\boldsymbol{x_8}$ yang merupakan parameter sisa kandungan gula dan kerapatan / densitas dari anggur merah, serta $\boldsymbol{x_6}$ dan $\boldsymbol{x_7}$ yang keduanya menyatakan konsentrasi sulfur dari minuman anggur merah. Temuan ini didukung pula oleh penelitian terdahulu yang menjelaskan bahwa adanya hubungan antara densitas, kandungan gula dan kandungan sulfur dalam minuman anggur merah terhadap harga dari minuman anggur merah (Oleksy, et al., 2021; Riera & Brummer, 2022) Meskipun tidak disebutkan secara eksplisit bahwa harga dari minuman anggur merah berbanding lurus dengan kualitasnya, namun secara logika, hal ini merupakan sesuatu yang lumrah untuk diterima bahwa harga berbanding lurus dengan kualitas.

Berdasarkan pada percobaan yang telah dilakukan terlihat bahwa algoritma yang diusulkan dapat dengan presisi memodelkan dan mengidentifikasi parameter yang berhubungan dengan kualitas minuman anggur merah. Hal ini menunjukan potensi dari algoritma untuk digunakan di berbagai kebutuhan identifikasi kualitas bahan pangan.

5. PENUTUP

Penelitian ini berfokuskan kepada identifikasi parameter yang mempengaruhi kualitas bahan pangan, yang dalam kasus ini, diambil tinjauan minuman anggur merah dan putih. Algoritma ini dimodifikasi dari algoritma pemilihan himpunan bagian berbasis entropi yang pertama kali digagas oleh (Tjen, et al., 2020) dan kemudian dikembangkan dalam (Smarra, et al., 2022).

Berdasarkan pada percobaan yang telah dilakukan, diperoleh akurasi model tertinggi dengan metode E-ss yang dimodifikasi adalah 96,25% untuk anggur merah dan 97,8% untuk anggur putih, dimana kedua nilai ini lebih baik daripada akurasi model prediktif yang dihasilkan oleh algoritma pohon klasifikasi klasik yang menggunakan seluruh parameter input yang ada. Lebih lanjut, berdasarkan pada percobaan yang telah dilakukan, Temuan utama dari penelitian ini adalah secara dominan kualitas dari anggur putih dipengaruhi oleh tingkat keasaman, pH, kadar alkohol dan klorit. Sedangkan untuk anggur merah, kualitas secara dominan dipengaruhi oleh kadar sulfur, residual gula dan kerapatan dari minuman anggur merah itu sendiri.

Saran penelitian terkait: adapun saran yang dapat diberikan untuk penelitian serupa adalah penggunaan informasi parameter yang lebih banyak, dan penggunaan kelas data yang lebih luas. Lebih lanjut, dirasa akan menarik apabila solusi analitik terhadap koefisien determinasi yang pasti dapat diturunkan, sehingga untuk setiap kasus, algoritma

yang digagas akan selalu memberikan hasil yang optimal.

DAFTAR PUSTAKA

BERTSIMAS, D., DUNN, J. & WANG, Y., 2021. Near-optimal nonlinear regression trees. *Operations Research Letters,* 49(2), pp. 201-206.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J., 2017. *Classification and regression trees.* s.l.:Routledge.

EDELMANN, D., MORI, T. F. & SZEKELY, G. J., 2021. On relationships between the Pearson and the distance correlation coefficients. *Statistics & Probability Letters,* Volume 169, p. 108960.

GOMES, H. M. ET AL., 2019. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter,* 21(2), pp. 6-22.

HEINZE, G., WALLISCH, C. & DUNKLER, D., 2018. Variable selection: a review and recommendations for the practicing statistician. *Biometrical journal,* 60(3), pp. 431-449.

HUSSAIN, L., 2018. Detecting epileptic seizure with different feature extracting strategies using robust machine learning classification techniques by applying advance parameter optimization approach. *Cognitive neurodynamics,* 12(3), pp. 271-294.

JAIN, A., SMARRA, F., BEHL, M. & MANGHARAM, R., 2018. Data-driven model predictive control with regression trees—an application to building energy management. *ACM Transactions on Cyber-Physical Systems,* 2(1), pp. 1-21.

JIANG, X., ZHANG, X. & ZHANG, Y., 2021. Establishment and optimization of sensor fault identification model based on classification and regression tree and particle swarm optimization. *Materials Research Express,* 8(8), p. 085703.

LINERO, A. R., 2018. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association,* 113(552), pp. 626--636.

MORE, A. & RANA, D. P., 2017. *Review of random forest classification techniques to resolve data imbalance.* s.l., IEEE, pp. 72-78.

MOSAVI, A. ET AL., 2019. State of the art of machine learning models in energy systems, a systematic review. *Energies,* 12(7), p. 1301.

MOSTASHARI, P. ET AL., 2022. Ozone in wineries and wine processing: A review of the benefits, application, and perspectives. *Comprehensive reviews in food science and food safety,* 21(4), pp. 3129-3152.

OLEKSY, P., CZUPRYNA, M. & JAKUBCZYK, M., 2021. On fine wine pricing across different trading venues. *Journal of Wine Economics,* 16(2), pp. 189-209.

PANJAITAN, S. D. ET AL., 2022. A Forecasting Approach for IoT-Based Energy and Power Quality Monitoring in Buildings. *IEEE Transactions on Automation Science and Engineering*.

PAULO, C. ET AL., 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems,* 47(4), pp. 547-553.

RAI, P. K. ET AL., 2019. Heavy metals in food crops: Health risks, fate, mechanisms, and management. *Environment international,* Volume 125, pp. 365-385.

RATH, S., TRIPATHY, A. & TRIPATHY, A. R., 2020. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* 14(5), pp. 1467-1474.

REN, Q. ET AL., 2021. Prediction of compressive strength of concrete with manufactured sand by ensemble classification and regression tree method. *Journal of Materials in Civil Engineering,* 33(7).

RIERA, F. S., BRUMMER, B., 2022. Environmental efficiency of wine grape production in Mendoza, Argentina. *Agricultural Water Management,* Volume 262, p. 107376.

SARAVANAN, R. & SUJATHA, P., 2018. *A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification.* Madurai, India, IEEE, pp. 945-949.

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.

SUPRIYADI, R., GATA, W., MAULIDAH, N. & FAUZI, A., 2020. Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis: Jurnal Ilmiah Ekonomi Dan Bisnis,* 13(2), pp. 67-75.

TJEN, J., SMARRA, F. & D'INNOCENZO, A., 2020. *An entropy-based sensor selection algorithm for structural damage detection.* Online Virtual Meeting, s.n.

VILELA, A., 2019. Use of nonconventional yeasts for modulating wine acidity. *Fermentation,* 5(1), p. 27.

WANG, H., YANG, M. & STUFKEN, J., 2019. Information-based optimal subdata selection for big data linear regression. *Journal of the*

American Statistical Association, 114(525), pp. 393-405.

YAO, L. H. ET AL., 2020. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Toronto, IEEE.

ZHANG, D., 2017. A coefficient of determination for generalized linear models. *The American Statistician*, 71(4), pp. 310-316.

ZHOU, X. ET AL., 2020. Rapid analysis of short-and medium-chain chlorinated paraffins in wine by dispersive liquid--liquid micro-extraction coupled with high performance liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry. *Food chemistry*, Volume 319, p. 126583.