Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 5 No. 2 (2024) 136 – 143 | ISSN: 2723-1453 (Media Online)

Identifikasi Kelayakan Air Minum Dengan Metode Analisis Komponen Utama Berbasis Entropi

Thommy Willay[1], Jimmy Tjen[2*], Paskalia Kartini[3], Riyadi Jimmy Iskandar[4]
[1,3]Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Universitas Widya Dharma Pontianak
[2,4] Jurusan Informatika, Fakultas Teknologi Informasi, Universitas Widya Dharma Pontianak
[1] w.thommy@gmail.com, [2]jimmy.tjen@mathmods.eu*, [3]paskalia@widyadharma.ac.id, [4]riyadi@widyadharma.ac.id

Abstract

The need for clean water is a fundamental requirement that must be met by humans, as water constitutes 60 to 70% of the total human body weight. Therefore, it is important to be able to determine the quality of the water entering the body, as consuming unsafe water will bring various diseases, such as diarrhea, and in severe cases might lead to death. This study aimed to investigate the factors which determine the potability of drinking water. Specifically, this research aims to produce a fault detection algorithm that can detect the potability of water samples based on Principal Component Analysis (PCA) and entropy-based subset selection methods. This paper addresses the linearity problem that commonly occurred in PCA by finding a subset of data that has a good entropy relation among the parameters contained in the subset, thus maintaining linearity in the data. There were 8 parameters considered in this reseach: pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organics carbon, trihalomethanes and turbidity. The experiment was conducted with 811 water samples, where 645 samples were used to train the model and the rest for validating the model predictive accuracy. Based on experiments conducted, it is confirmed that the proposed algorithm can determine the potability of drinking water samples from synthetic data sourced from India with an accuracy of over 98% for potable water data and 100% for non-potable water data.

Keywords: Principal Component Analysis, Clean Water, Entropy, Water Potability

Abstrak

Kebutuhan akan air bersih merupakan kebutuhan mendasar yang harus terpenuhi oleh umat manusia karena air menyusun 60 hingga 70% dari keseluruhan bobot manusia. Oleh karena itu, dirasa penting untuk mengetahui kelayakan air yang masuk ke dalam tubuh, sebagaimana mengonsumsi air yang tidak layak tentunya akan mendatangkan berbagai penyakit, seperti diare, dan bahkan dapat mengakibatkan kematian. Penelitian ini digagas untuk mempelajari kriteria dan faktor yang menentukan kelayakan dari kelayakan air minum. Secara spesifik, penelitian ini bertujuan untuk menghasilkan suatu algoritma pendeteksian kerusakan (*fault detection*) yang dapat mendeteksi kelayakan minum dari sampel air berdasarkan metode Analisis Komponen Utama (*Principal Component Analysis*) atau AKU dan metode pemilihan subhimpunan berbasis entropi (*entropy-based subset selection*). Metode ini berfokuskan untuk menyelesaikan permasalahan linearitas data pada AKU, dengan menemukan sekelompok himpunan bagian data yang memiliki hubungan entropi yang baik, sehingga dapat mempertahankan faktor linearitas pada data. Pada penelitian ini terdapat 8 parameter air yang digunakan untuk menentukan kelayakan minum dari air: pH, kekerasan (*hardness*), total padatan terlarut, kandungan kloramin (*chloramines*), kandungan sulfat, konduktivitas air, kandungan karbon, kandungan trihalometana dan turbiditas. Percobaan melibatkan 811 sampel air, dengan 645 sampel digunakan untuk melatih model, dan 162 untuk tahap validasi. Berdasarkan percobaan yang telah dilakukan, diketahui bahwa algoritma yang digagas mampu menentukan kelayakan dari sampel air minum dari sebuah data sintetis yang bersumber dari India dengan akurasi di atas 98% untuk data air layak minum dan 100% untuk data air tidak layak minum.

Kata kunci: Analisis Komponen Utama, Air Bersih, Entropi, Potabilitas Air

1. Pendahuluan

Air merupakan zat yang menyusun 60 hingga 70% dari keseluruhan bobot manusia [1]. Oleh karena itu, penting bagi manusia untuk memenuhi kebutuhannya akan air layak minum. Air layak minum (*potable*) didefinisikan sebagai air yang tidak memiliki rasa, bau, warna, memiliki pH pada rentang 6,5 hingga 8,5, dan tidak memiliki kandungan logam berat pada rentang tertentu[2], [3], [4]. Di Indonesia sendiri, berdasarkan pada data yang dihimpun dari Biro Pusat Statistik (BPS) pada tahun 2022, angka ketersedian air layak minum berada pada rentang 65 hingga 98%[5].

Mengonsumsi air yang tidak layak tentunya akan mendatangkan berbagai penyakit, seperti diare dan bahkan dapat mengakibatkan kematian. Hal ini lazim terjadi di daerah terpencil dan pemukiman kumuh,

karena ketiadaan akses masyarakat untuk memperoleh air layak minum. Hal ini kemudian diperburuk dengan ketidakpahaman masyarakat terkait dengan kriteria air yang layak untuk diminum, yang menurunkan kualitas hidup masyarakat. Oleh karena itu, dirasa penting untuk diberikannya edukasi kepada masyarakat, terutama untuk dapat mengenali manakah air yang layak diminum dan mana yang tidak. Secara spesifik, untuk mengatasi permasalahan ini, dapat diterapkan metode berbasis data atau *data-driven approach* yang merupakan metode berbasis pembelajaran mesin untuk menurunkan model berdasarkan pada informasi data yang tesedia[6], [7]. Terdapat banyak algorithma yang dibangun berdasarkan metode ini. Salah satunya adalah Analisis Komponen Utama (AKU) atau *Principal Component Analysis* (PCA)[8].

AKU adalah sebuah metode yang memproyeksikan data pada bidang ortogonal yang bersesuaian sehingga varians dari data menjadi maksimal. Lewat penyederhanaan ini, pola-pola dari data akan lebih mudah teramati karena dimungkinkan untuk mengurangi dimensionalitas (kardinalitas) dari data tanpa menghilangkan informasi yang terkandung di dalam data[9]. Beberapa penelitian yang mengangkat metode AKU antara lain telah dilakukan oleh:[10], [11], [12], [13], [14].

Penulis pada [10] menggunakan metode AKU untuk mengetahui apakah seseorang terkena diabetes atau tidak. Berdasarkan pada percobaan yang telah dilakukan, disimpulkan bahwa metode AKU dapat memodelkan prevalensi penyakit diabetes dengan akurasi di atas 90%. Peneliti pada [11] menerapkan metode AKU untuk mengidentifikasi kualitas air pada sungai Gangga India. Pada penelitian [12], metode AKU diterapkan pada data air dari sungai Xin'Anjiang. Pada percobaan tersebut, diperoleh hasil bahwa metode AKU dapat mengidentifikasi keberadaan dari zat polutan dengan akurasi sebesar 87,24%. Peneliti pada [13] meneliti kelayakan dari air tanah yang diambil dari daerah Wadi El Natrun, mesir dengan menggunakan metode AKU dan *Hierarchical Clustering Analysis* (HCA). Berdasarkan pada percobaan yang dilakukan, disimpulkan bahwa kandungan dari mineral terlarut seperti sodium, kalsium, magnesium, sulfat dan strontium merupakan parameter kuat yang mempengaruhi kualitas air tanah. Terakhir, pada penelitian [14]metode AKU digunakan untuk mengidentifikasi apakah air pada sungai Hau, pada muara sungai Mekong, Vietnam telah tercemar atau tidak. Berdasarkan pada percobaan yang telah dilakukan, terlihat bahwa metode AKU dapat mengidentifikasi bahwa sungai Hau tercemar dengan zat *coliforms* sepanjang tahunnya. Secara umum, seluruh penelitian di atas telah membuktikan kemampuan dari metode AKU dalam mendeteksi ketidaknormalan dari sebuah data.

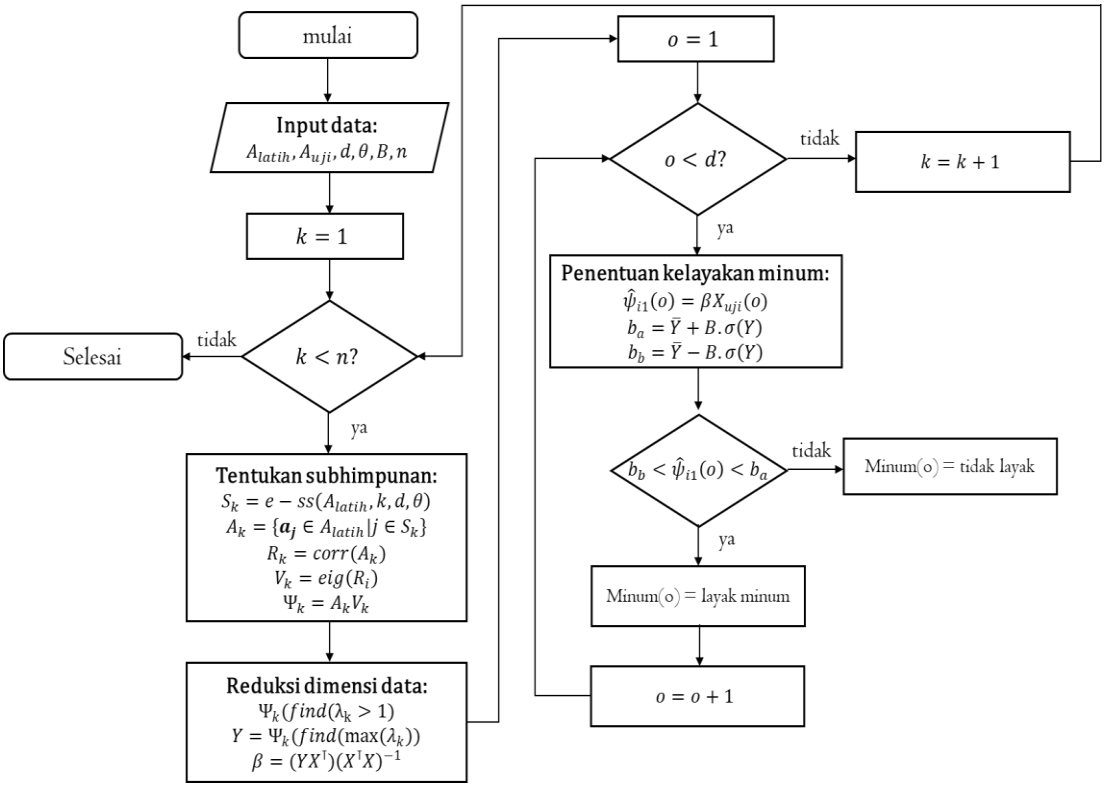
Meskipun metode AKU berfungsi dengan baik dalam mendeteksi anomali pada data, metode ini juga memiliki kelemahan. Secara spesifik, metode AKU berporos pada transformasi ortogonal dari data dengan cara melakukan kombinasi linear dari data semula. Sehingga ketika asumsi linear ini tidak terpenuhi, maka metode AKU tidak dapat bekerja sebagaimana mestinya[15]. Lebih lanjut, proses transformasi ortogonal dari AKU melibatkan proses *Eigen Value Decomposition* (EVD) yang memiliki kompleksitas komputasi sebesar $O(n^3)$ untuk matriks berukuran $n \times n$ [16].

Penelitian ini dilakukan untuk membangun algoritma AKU untuk mendeteksi kelayakan air minum dengan menyoroti permasalahan yang telah disampaikan pada paragraf sebelumnya. Secara spesifik, penelitian ini dilakukan dengan mempertimbangkan 2 buah kontribusi utama sebagai berikut: a). Penelitian ini memodifikasi alur pemilihan himpunan bagian berbasis entropi yang diperkenalkan pada [15] dan dikembangkan pada [10] untuk mendeteksi kelayakan air minum. Sehingga metode AKU dapat tetap berjalan dengan optimal tanpa membutuhkan kompleksitas yang tinggi. b). Penelitian ini bertujuan untuk mengidentifikasi parameter utama dalam menentukan kelayakan air minum dengan menggunakan metode AKU.

Penelitian ini akan dilakukan dengan mengikuti alur penelitian yang dilakukan pada [10]dan [12], [14]. Secara spesifik, penelitian ini mengambil proses AKU sesuai dengan alur yang diberikan pada [12]dan[14]. Namun, setelah proses transformasi ortogonal, akan disisipkan model linear seperti yang digunakan pada[10]. Tujuannya adalah untuk mempelajari dinamika dari data. Namun, berbeda dari model yang digunakan pada[10], pada penelitian ini, komponen utama yang digunakan akan disortir terlebih dahulu dengan menyisakan hanya komponen dengan nilai eigen di atas 1. Proses ini dikenal sebagai *kaiser criterion*[17]. Sehingga, dalam kasus ini, proses regresi akan mengambil variabel terikat sebagai komponen utama dengan nilai eigen tertinggi, sedangkan variabel bebas diambil sebagai komponen utama lain dengan nilai eigen di atas 1. Tujuan dari proses ini adalah agar model linear yang digunakan untuk menghubungkan antar parameter akan semakin sederhana, sehingga kompleksitas perhitungan juga dapat semakin ditekan. Penerapan *kaiser criterion* pada model linear merupakan pembaharuan terhadap metode AKU yang digunakan pada [10].

Karya tulis ini terdiri dari 4 bagian: pada bagian pertama telah dibahas latar belakang penelitian dan alasan yang mendasari penelitian. Pada bagian kedua akan dibahas algoritma dan alur yang digunakan pada penelitian ini. Secara spesifik akan dibahas algoritma dari metode AKU yang dikombinasikan dengan pemilihan himpunan bagian berbasis entropi. Pada bagian ketiga akan ditunjukan hasil simulasi numerik sesuai dengan algoritma yang telah ditunjukan pada bagian

DOI : https://doi.org/10.52158/jacost.v5i2.815



Gambar 1. Diagram alir metode AKU dengan pemilihan himpunan bagian berbasis entropi.

sebelumnya. Bagian terakhir akan menyimpulkan keseluruhan performa dari algoritma yang digagas, serta akan dibahas lebih lanjut terkait dengan arah penelitian dimasa mendatang, terutama terkait dengan monitor kualitas air secara berkelanjutan.

2. Metode Penelitian

Pada bagian ini akan dibahas algoritma identifikasi kualitas air minum berbasis metode AKU dan pemilihan himpunan bagian berbasis entropi. Lebih lanjut, akan dibahas data yang digunakan dalam penelitian ini. Silahkan merujuk pada [10], [15] terkait dengan metode AKU dan [18], [19] terkait dengan metode pemilihan himpunan bagian berbasis entropi.

2.1. Algoritma Identifikasi Kualitas Air Minum

Pada bagian ini, akan diperlihatkan alur dan persamaan matematis untuk algoritma penentuan kelayakan air minum atau *potability* dari air minum. Algoritma ini terdiri dari tiga bagian utama, yaitu penentuan subhimpunan berbasis entropi, pemilihan komponen utama dan penyusunan model linear.

Misalkan $A = [\boldsymbol{a_1}\ \boldsymbol{a_2} \dots \boldsymbol{a_n}]; A \in \mathbb{R}^{m \times n}$ adalah himpunan data pengukuran kualitas air minum dengan $\boldsymbol{a_i} = [a_1(1)\ a_i(2) \dots a_i(m)]^\intercal; \boldsymbol{a_i} \in \mathbb{R}^m$ menyatakan vektor dari parameter pengukuran kualitas air minum ke$-i$ dengan $i = 1,2, \dots, n$. Lebih lanjut, Misalkan $\boldsymbol{a_h} = [a_h(1)\ a_h(2) \dots\ a_h(m)]^\intercal; \boldsymbol{a_h} \in \{0,1\}; \boldsymbol{a_h} \in A$ menyatakan kondisi kelayakan air minum dari sampel air yang digunakan. Dalam kasus ini, ketika $a_h(k)$ bernilai 0 (untuk $k = 1,2, \dots, m$), maka dapat dikatakan bahwa sampel air ke-k merupakan air yang tidak layak untuk diminum dan ketika bernilai 1 maka sampel air merupakan sampel layak diminum. Dalam kasus ini, maka himpunan data A dapat dibagi menjadi 2 himpunan bagian: layak minum dan tidak layak minum. Misalkan $A_m \in \mathbb{R}^{m^* \times n}$ adalah himpunan bagian data dari air layak minum. Sebagai contoh, $a_h(k) = 1, \forall a_h(k) \in A_m; k = 1,2, \dots, m^*$ dan $A_{tm} \in \mathbb{R}^{m^{**} \times n}$ adalah himpunan bagian data dari air tidak layak minum (dengan kondisi $a_h(k) = 0, \forall a_h(k) \in A_{tm}; k = 1,2, \dots, m^{**}$). Sehingga $A_m \cup A_{tm} = A; A_m \cap A_{tm} = \emptyset$.

Langkah pertama. Berdasarkan pada definisi data di atas, maka dapat dibentuk sebuah himpunan bagian data dengan menggunakan metode pemilihan himpunan bagian berbasis entropi. Untuk setiap $\boldsymbol{a_i} \in A_m$, misalkan S_i adalah himpunan indeks ke-i yang terbentuk karena algoritma pemilihan himpunan bagian berbasis entropi seperti pada [19] dengan kardinalitas dari $S_i = u_i$. Kemudian, untuk setiap D_i, $D_i \subset A_m = \{\boldsymbol{a_k} \in A_m | k \in S_i\}$ yang merupakan data yang terkandung sesuai dengan himpunan indeks S_i, data tersebut dapat ditransformasikan pada bidang ortogonal yang bersesuaian. Sebagai contoh, misalkan $V_i \in \mathbb{R}^{u_i \times u_i}$ adalah vektor eigen dari matriks korelasi R_i yang dihasilkan dari himpunan bagian D_i. Maka untuk setiap himpunan bagian dapat didefinisikan Ψ_i sebagai

DOI : https://doi.org/10.52158/jacost.v5i2.815

$$\Psi_i = D_i V_i \,, \tag{1}$$

dengan $\Psi_i = [\boldsymbol{\psi_1}\ \boldsymbol{\psi_2}\ \dots \boldsymbol{\psi_{u_i}}]$ menyatakan himpunan dari komponen utama akibat himpunan indeks S_i.

Langkah kedua. Pada tahap ini, tujuan utama dari proses adalah untuk menentukan variabel bebas dan terikat untuk setiap himpunan bagian data yang telah ditransformasikan pada bidang ortogonal. Misalkan $\boldsymbol{\lambda_i} = [\lambda_i(1)\ \lambda_i(2)\ \dots \lambda_i(u_i)]; \boldsymbol{\lambda_i} \in \mathbb{R}^{u_i}$ merupakan sebuah vektor yang berisikan nilai eigen yang bersesuaian dengan vector eigen V_i. Misalkan

$$\lambda_{m_i} = \max \boldsymbol{\lambda_i} \tag{2}$$

Merupakan nilai eigen tertinggi untuk himpunan bagian D_i. Dalam kasus ini, $\boldsymbol{\psi_{y_i}}$ atau komponen utama yang menjadi variabel terikat untuk proses regresi pada tahap selanjutnya dapat ditentukan sebagai

$$\boldsymbol{\psi_{y_i}} = \boldsymbol{\psi_k} \in \Psi_i | \max_k \lambda_i(k), k = 1,2, \dots u_i \tag{3}$$

atau, komponen utama dengan nilai eigen tertinggi. Sedangkan untuk variabel bebas, maka himpunan variabel bebas dari himpunan bagian data D_i didefinisikan sebagai:

$$\Psi_{x_i} = \{\boldsymbol{\psi_k} \in \Psi_i | \lambda_i(k) \in \lambda_i/\lambda_{m_i} > 1, k = 1,2, \dots, u_i\} \tag{4}$$

atau dengan kata lain, Ψ_{x_i} merupakan kumpulan dari komponen utama selain dari λ_{m_i} yang memiliki nilai eigen lebih dari 1. Pembuangan komponen utama ini dapat dilakukan dengan asumsi bahwa setiap komponen utama memiliki jumlah informasi (atau varians) yang berbeda-beda. Sehingga, penyusutan dimensi dapat dilakukan dengan cara membuang komponen utama yang tidak memuat banyak informasi (nilai eigen yang kurang dari 1).

Langkah ketiga. Langkah terakhir adalah menyusun persamaan matematis yang menghubungkan ψ_{y_i} dengan Ψ_{x_i}. Untuk $\Psi_{x_i} = [\boldsymbol{\psi_{x_1}}\ \boldsymbol{\psi_{x_2}}\ \dots \boldsymbol{\psi_{x_p}}]$ maka dapat disusun sebuah persamaan linear

$$\hat{\psi}_{y_i}(k) = \sum_{j=1}^{p} \alpha_j \psi_{xi}(j) \tag{5}$$

dengan α_j adalah parameter persamaan matematis yang dapat diperoleh dengan metode regresi seperti metode kuadrat terkecil (lihat:[20]) dan $\hat{\psi}_{yi}(k)$ menyatakan prediksi dari $\boldsymbol{\psi_{yi}}$ pada saat k.

Persamaan 5 akan menghasilkan persamaan matematis pada bidang ortogonal yang sensitif terhadap data yang bersifat nominal (dalam keadaan baik). Oleh karena itu, ketika diberi masukan non-nominal, dinamik yang dihasilkan akan berubah jauh dari yang diharapkan. Prinsip ini kemudian akan digunakan untuk mendeteksi apakah sampel baru termasuk dalam kriteria air yang layak minum atau tidak layak minum. Secara khusus, dapat didefinisikan batas atas dan bawah sebagai

$$b_a = \bar{\boldsymbol{\psi}}_{\boldsymbol{y_i}} + B.\sigma(\boldsymbol{\psi_{y_i}}) \tag{6}$$

dan

$$b_b = \bar{\boldsymbol{\psi}}_{y_i} - B.\sigma(\boldsymbol{\psi_{y_i}}) \tag{7}$$

dengan b_a dan b_b berturut-turut menyatakan batas atas dan batas bawah dari suatu sampel air dapat dikatagorikan sebagai air layak minum, $B \in \mathbb{Z}^+$ menyatakan faktor pengali sembarang yang berupa bilangan bulat, serta $\bar{\boldsymbol{\psi}}_{\boldsymbol{yi}}$ dan $\sigma(\boldsymbol{\psi_{y_i}})$ secara berurutan menyatakan rerata dan standar deviasi dari $\boldsymbol{\psi_{y_i}}$. Secara spesifik, nilai dari B dapat berupa apa saja. Namun, perlu diperhatikan bahwa pemilihan B yang terlalu kecil akan cenderung meningkatkan peluang terjadinya kasus negatif palsu (*false negative*), sedangkan B yang terlalu besar akan meningkatkan peluang positif palsu (*false positive*).

Untuk titik sampel yang tidak diketahui, pengujian kelayakan dapat dilakukan dengan asumsi berikut. Misalkan $\boldsymbol{a_t} = [a_1(t)\ a_2(t) \dots a_n(t)]$ adalah titik pengujian dari air minum yang tidak diketahui nilai *potability*-nya. Dalam kasus ini, dengan sedikit penyalahgunaan notasi (*slight abuse of notation*), titik a_t dinyatakan sebagai air layak minum jika

$$b_b \leq a_t V_i \leq b_a \tag{8}$$

dan tidak layak minum apabila sebaliknya dengan $a_t V_i$ menyatakan transformasi ortogonal dari data a_t.

Algoritma 1 merupakan algoritma penentuan kelayakan air minum berbasis AKU dan pemilihan himpunan bagian berbasis entropi. Secara teknis, algoritma terdiri dari tiga tahap berbeda, yakni penentuan subhimpunan berbasis entropi, proyeksi ortogonal oleh AKU, dan penentuan parameter model dengan persamaan kuadrat terkecil. Metode AKU dan himpunan bagian berbasis entropi diketahui memiliki kompleksitas sebesar $O(mn^2)$ [15]dengan m menyatakan jumlah sampel dan n menyatakan jumlah parameter, sedangkan metode kuadrat terkecil dalam kasus ini memiliki kompleksitas $O(mn)$. Dengan demikian, keseluruhan dari algoritma memiliki kompleksitas sebesar $2 \times O(mn^2) + O(mn) = O(mn^2)$.

2.2. Data Penelitian

Untuk menguji performa dari algoritma yang digagas, maka algoritma akan diuji dengan data yang diperoleh dari [21]. Data tersebut merupakan data sintetis yang dihasilkan berdasarkan pada pengamatan air di negara India. Data yang dikumpulkan memiliki 3.276 titik sampel yang meliputi parameter: pH yang merupakan ukuran asam basa suatu zat, kekerasan (*hardness*) yang menyatakan kandungan kalsium dan garam magnesium pada air, total padatan terlarut yang meliputi kandungan potasium, kalsium, sodium, bikarbonat, klorit, magnesium, dan sulfat, kandungan kloramin (*chloramines*) yang merupakan zat disinfektan yang lazimnya ditambahkan pada air, kandungan sulfat akibat

DOI : https://doi.org/10.52158/jacost.v5i2.815

bebatuan dan tumbuhan yang tumbuh di sekitar mata air, konduktivitas listrik dari air, kandungan karbon organik yang terlarut dalam air, kandungan trihalometana yang menjadi produk sampingan dari penambahan klorin dalam air, turbiditas yang menyatakan jumlahan padatan terlarut dalam air pada kondisi tersuspensi, dan *potability* yang merupakan indikator apakah air dapat diminum atau tidak.

Dari keseluruhan 3.276 sampel diketahui bahwa terdapat 811 sampel data nominal (dapat diminum), 1200 sampel data non-nominal (tidak layak minum) dan 1.265 sampel data yang tidak dapat digunakan karena permasalahan missing data point ataupun tidak diberikan informasi apakah air tersebut layak diminum atau tidak. Pada penelitian ini, sebanyak 649 sampel atau sekitar 80% dari data nominal akan digunakan untuk melatih model. Sedangkan 162 sampel data nominal dan seluruh data non-nominal akan digunakan untuk memvalidasi performa dari algoritma yang telah digagas.

Algoritma 1. Penentuan Kelayakan Air Minum

Masukan: A_{latih}; A_{uji}; d; $\theta; B$
Keluaran: minum
Proses:
Untuk $k = 1$ sampai n, lakukan
  S_k = algoritma-ess(A_{latih}, k, d, θ)
  $u = |S_k|$
  $A_k = \{\boldsymbol{a_j} \in A_{latih} | j \in S_k\}$
  $R_k = corr(A_k)$
  $V_k = eig(R_i)$
  $\Psi_k = A_k V_k$
  $X = \Psi_k(find(\lambda_k > 1)$
  $Y = \Psi_k(find(\max(\lambda_k))$
  $\beta = (YX^\intercal)(X^\intercal X)^{-1}$
  $b_a = \bar{Y} + B.\sigma(Y)$
  $b_b = \bar{Y} - B.\sigma(Y)$
  Untuk $o = 1$ sampai panjang(X_{uji})
    $\hat{\psi}_{i1}(o) = \beta X_{uji}(o)$
    Jika $\hat{\psi}_{i1}(o) > b_b$ && $\hat{\psi}_{i1}(o) < b_a$
      minum(o) = "bisa diminum"
      lainnya
      minum(o) = "tidak bisa diminum"
    Selesai
  Selesai
Selesai

2.3 Analisis Numerik

Terdapat 4 besaran yang digunakan untuk memvalidasi performa dari algoritma yang diusulkan: akurasi, presisi *recall* dan $F1$ *score*. Akurasi digunakan untuk menyatakan persentase dari tebakan benar yang dapat dilakukan oleh algoritma. Dalam kasus ini, akurasi didefinisikan sebagai

$$A\% = \frac{TP + TN}{TP + TN + FN + FP} \times 100\%, \qquad (9)$$

dengan *TP*, *TN*, *FN* dan *FP* secara berurutan menyatakan nilai *true positive* (positive sejati), *true negative* (negatif sejati), *false negative* (negatif palsu) dan *false positive* (positif palsu). Dalam kasus ini, tebakan sampel t dikatagorikan sebagai TP apabila sampel t yang merupakan data nominal terkandung di dalam batas sesuai dengan definisi pada persamaan (6) dan (7), dan tebakan akan sampel t disebut FN apabila sebaliknya. Tebakan sampel t dikatakan sebagai TN, apabila sampel tersebut berasal dari data non-nominal dan berada di luar batas yang didefinisikan pada persamaan (6) dan (7). Apabila sampel t terkandung di dalam batas disaat t berasal dari data non-nominal, maka sampel t akan dikelompokan sebagai FP.

Presisi atau *precision* merupakan ukuran yang menyatakan proporsi dari tebakan kelas positif atau nominal yang benar (TP) terhadap semua tebakan yang diberi kelas positif (TP dan FP). Presisi dinyatakan sebagai

$$P\% = \frac{TP}{TP + FN} \times 100\%. \qquad (10)$$

Recall atau yang dikenal juga sebagai *True Positive Rate* (TPR) merupakan sebuah ukuran yang menyatakan proporsi dari tebakan kelas positif atau nominal yang benar terhadap semua data yang seharusnya memiliki kelas positif. Recall dinyatakan sebagai

$$R\% = \frac{TP}{TP + FN} \times 100\%. \qquad (11)$$

Terakhir, $F1$ *score* merupakan rerata harmonik dari presisi dan *recall*. Dalam kasus ini, $F1$ *score* dinyatakan sebagai

$$F_1 = 2 \times \frac{\%P \times \%R}{\%P + \%R}. \qquad (12)$$

$F1$ *score* digunakan ketika terdapat perbedaan panjang data antara kelas nominal dan non-nominal [22].

3. Hasil dan Pembahasan

Pada bagian ini, akan diperlihatkan bagaimana algoritma yang telah disusun mampu mendeteksi kelayakan dari air minum. Dalam kasus ini, algoritma yang digagas akan divalidasi dalam dua hal, yaitu pengukuran akurasi terhadap data nominal dan pengukuran akurasi pada data non-nominal. Lebih lanjut, berdasarkan analisis ini, akan ditentukan pula parameter yang signifikan untuk mendeteksi kelayakan air minum berdasarkan Algoritma 1. Pada penelitian ini, dipilih parameter $B = 2$ dan $\theta = 0{,}15$ yang menandakan bahwa sampel dapat terdistribusi sejauh 2 kali standar deviasinya sebelum dinyatakan sebagai tidak layak minum, sedangkan entropi yang dikehendaki antar variabel adalah sebesar 0,15.

Untuk mempermudah penyebutan variabel, maka 9 variabel yang digunakan dalam model akan direpresentasikan sebagai a_1 hingga a_9. Tabel 1 merepresentasikan parameter kualitas air dalam bentuk variabel matematis. Untuk mempermudah dan mempersingkat penyebutan, kata "model a_j" akan mengacu pada model yang dibangun dengan parameter a_j sebagai parameter utama dari himpunan bagian ke$-j$. Sebagai perbandingan, maka akan ditampilkan pula akurasi jika menggunakan metode pohon klasifikasi

DOI : https://doi.org/10.52158/jacost.v5i2.815

(dengan validasi 10 *k-folds*), metode *random forest* dengan 100 pohon biner dan metode PCA sesuai pada [13].

Tabel 1. Representasi parameter kualitas air dalam variabel matematis

Nama Parameter	Representasi Variabel
pH	a_1
kekerasan (*hardness*)	a_2
total padatan terlarut	a_3
kandungan kloramin (*chloramines*)	a_4
kandungan sulfat	a_5
konduktivitas listrik	a_6
kandungan karbon organik	a_7
kandungan trihalometana	a_8
turbiditas	a_9

3.1. Akurasi Model Prediktif

Tabel 2 menunjukan himpunan bagian yang terbentuk untuk setiap parameter.

Tabel 2. Jumlah FP, TN, TP dan FN untuk setiap metode

Model	Subhimpunan
a_1	a_1, a_4, a_5
a_2	a_2, a_5
a_3	a_3, a_5, a_4
a_4	a_4, a_1, a_3
a_5	a_5, a_3, a_8, a_2, a_1
a_6	Tidak ada
a_7	Tidak ada
a_8	x_8, x_5
a_9	Tidak ada

Berdasarkan simulasi yang telah dilakukan, beberapa parameter memiliki hubungan entropi yang "baik" satu sama lain. Sebagai contoh, parameter total padatan terlarut, kandungan kloramin, dan kandungan sulfat memiliki hubungan satu sama lain, demikian pula dengan kandungan sulfat dan kandungan trihalometana. Lebih lanjut, pada Tabel 2 terlihat bahwa model a_6, a_7 dan a_9 tidak menghasilkan himpunan bagian baru dengan parameter lain. Hal ini disebabkan karena tidak ada parameter lain yang memiliki hubungan secara entropi dengan parameter tersebut. Dalam kasus ini, metode pemilihan himpunan bagian berbasis entropi tidak dapat menghasilkan pasangan bagi ketiga parameter tersebut.

Meskipun memiliki hubungan satu sama lain, kondisi tersebut tidak serta merta mengakibatkan parameter yang bersangkutan menjadi model yang baik untuk memprediksi kelayakan air minum. Sebagai contoh, parameter kekerasan dan kandungan sulfat memiliki hubungan entropi yang baik, tetapi tidak dapat memprediksi kelayakan air minum sama sekali.

Tabel 3 menunjukan jumlah TP, TN, FP dan FN untuk setiap himpunan bagian yang dibentuk berdasarkan Algoritma 1. Sedangkan Tabel 4 menunjukan akurasi model prediktif untuk setiap himpunan bagian. Lebih lanjut, pada Tabel 4 diperlihatkan pula akurasi untuk metode AKU (PCA), metode pohon klasifikasi (CT dan K10 CT) serta metode *random forest* (RF).

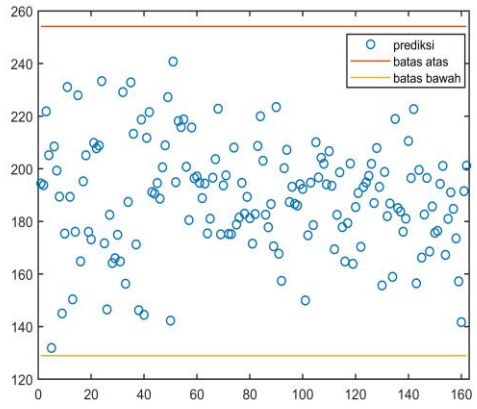
Tabel 3. Jumlah FP, TN, TP dan FN untuk setiap metode

No. Subset	FP	TN	TP	FN
1	0	1200	0	162
2	1200	0	0	162
3	833	367	116	46
4	239	961	159	3
5	132	1068	25	137
6	-	-	-	-
7	-	-	-	-
8	23	1177	162	0
PCA	185	1015	158	4
CT	64	136	102	100
K10 CT	78	159	83	82
RF	31	169	72	130

Tabel 4. Akurasi model prediktif untuk penentuan kualitas air minum

No.	A%	P%	R%	F1 score
1	88,11%	-	0	-
2	0%	0%	0	-
3	35,46%	12,22%	71,60%	20.88%
4	82,23%	39,95%	98,15%	56.79%
5	80,25%	15,92%	15,43%	15.67%
6	-	-	-	-
7	-	-	-	-
8	98,31%	87,57%	100%	93.37%
PCA	86,12%	46,06%	97,53%	62.57%
CT	59,20%	61,45%	50,50%	55.43%
K10 CT	60,20%	51,55%	50,30%	50.92%
RF	59,95%	69,90%	35,64%	47.21%

*akurasi terbaik secara keseluruhan



Gambar 2. Representasi prediksi sampel uji yang bersifat nominal menggunakan model a_8.

Berdasarkan Tabel 4, terlihat bahwa model dengan parameter kandungan trihalometana dan kandungan sulfat (model a_8) adalah model yang paling baik dalam memprediksi kelayakan air minum dengan akurasi prediksi sampel nominal sebesar 100% dan akurasi prediksi sampel non-nominal (rasio antara TN terhadap jumlahan TN dan FP) sebesar 98,31%. Ini menandakan bahwa model a_8 dapat dengan akurat menjelaskan air yang layak diminum dan tidak layak diminum. Lebih lanjut, dapat dilihat bahwa model a_3 dapat mengkategorikan air yang layak diminum dengan akurasi sebesar 71,6%, tetapi gagal dalam

DOI : https://doi.org/10.52158/jacost.v5i2.815

mengkategorikan air yang tidak dapat diminum dengan akurasi yang berkisar pada angka 30%. Hal ini menunjukkan bahwa dari 9 parameter yang memengaruhi kualitas air minum, parameter kandungan sulfat dan kandungan trihalometana adalah parameter yang paling memengaruhi kelayakan air minum, dan kemudian disusul oleh parameter kandungan sulfat, kandungan kloramin, dan total padatan terlarut. Gambar 2 menunjukan dinamika dari data untuk air yang layak minum, sedangkan Gambar 3 menyatakan dinamika dari air yang tidak dapat diminum. Dalam kasus ini terlihat bahwa sampel air layak minum terkandung dalam bidang batas, sedangkan yang tidak layak minum telah bergeser dari ambang dinamik yang diizinkan.

Pada Tabel 4 terlihat bahwa metode AKU berbasis entropi memiliki performa yang lebih baik jika dibandingkan dengan metode AKU yang digunakan pada [12]. Hal ini menunjukan bahwa metode pemilihan himpunan bagian berbasis entropi mampu menyelesaikan permasalahan linearitas data yang umumnya terjadi pada metode AKU. Lebih lanjut, terlihat pula bahwa metode AKU berbasis entropi memiliki performa *F*1 *score* yang lebih baik jika dibandingkan dengan metode pohon klasifikasi bahkan dengan metode *random forest* yang melibatkan 100 pohon biner. Hal ini menunjukan bahwa metode AKU yang diusulkan memiliki performa klasifikasi yang lebih baik jika dibandingkan dengan 3 metode yang telah disebutkan di atas. Lebih lanjut, perlu dicermati bahwa metode AKU berbasis entropi hanya menggunakan informasi 2 parameter dan bukan 9 parameter seperti pada 3 metode lainnya, namun terbukti masih memiliki akurasi, presisi, *recall* dan *F*1 *score* yang lebih baik jika dibandingkan dengan metode lainnya.

3.2. Identifikasi Kualitas Air Minum

Berdasarkan pada percobaan yang telah dilakukan, dapat diketahui bahwa model yang dibangun berdasarkan parameter kandungan trihalometana dan kandungan sulfat adalah model yang dapat dengan akurat memprediksi kelayakan air minum. Hal ini menunjukan bahwa metode AKU berbasis entropi mengidentifikasi kandungan trihalometana sebagai parameter yang memiliki informasi terbanyak jika dibandingkan dengan 7 parameter yang lain.

Kemudian, parameter lain yang memiliki akurasi baik dalam memprediksi kelayakan air minum adalah total padatan terlarut, kandungan kloramin, dan kandungan sulfat. Meskipun demikian, penelitian ini menunjukkan bahwa ketiga unsur di atas tidak memiliki akurasi model yang lebih baik daripada model yang dibangun dengan parameter trihalometana dan kandungan sulfat. Hal ini disebabkan karena metode AKU berbasis entropi tidak mampu membentuk sub himpunan yang optimal untuk memprediksi potabilitas dari air minum. Lebih lanjut, metode AKU berbasis entropi menyatakan bahwa konduktivitas listrik dan kandungan karbon merupakan parameter yang tidak dapat digunakan untuk menentukan potabilitas dari air minum. Hal ini disebabkan karena tidak adanya parameter lain yang berhubungan erat secara entropi dengan kedua parameter tersebut. Hal ini mengisyaratkan bahwa parameter tersebut dapat dibuang dari data, dan tidak akan mengakibatkan algoritma kehilangan kemampuan prediksinya.

4. Kesimpulan

Pada penelitian ini, telah digagas sebuah algoritma baru untuk memprediksi kelayakan dari air minum berdasarkan metode AKU dan pemilihan himpunan bagian berbasis entropi. Metode ini ditujukan untuk menyelesaikan permasalahan linearitas yang dapat muncul pada metode AKU tanpa meningkatkan kompleksitas dari algoritma tersebut.

Berdasarkan pada percobaan yang telah dilakukan, dapat disimpulkan bahwa metode AKU berbasis entropi memiliki performa yang lebih baik jika dibandingkan dengan metode AKU standar, metode pohon klasifikasi dan metode *random forest* dengan akurasi pada himpunan bagian terbaik mencapai 98%. Lebih lanjut, metode AKU berbasis entropi mampu mengidentifikasi kandungan sulfat dan trihalometana sebagai parameter yang menentukan potabilitas dari air minum.

Melihat performa dari algoritma yang digagas, diharapkan ke depan algoritma ini dapat dikemas dalam bentuk perangkat lunak yang disematkan pada piranti pengukuran air, sehingga kelayakan air minum dapat diketahui segera tanpa membutuhkan waktu proses yang lama. Lebih lanjut, model prediktif yang digunakan pada algoritma ini dapat dikembangkan ke dalam bentuk yang bergantung waktu (*time variant*) sehingga dapat digunakan sebagai media monitoring kelayakan air minum dari waktu ke waktu.

Daftar Pustaka

[1] F. A. Padder and A. Bashir, “Scarcity of water in the twenty-first century: Problems and potential remedies,” *MEDALION JOURNAL: Medical Research, Nursing, Health and Midwife Participation*, vol. 4, no. 1, pp. 1–5, 2023.

[2] K. de Mello *et al.*, “Multiscale land use impacts on water quality: Assessment, planning, and future perspectives in Brazil,” *J Environ Manage*, vol. 270, p. 110879, Sep. 2020, doi: 10.1016/j.jenvman.2020.110879.

[3] M. Salehi, “Global water shortage and potable water safety; Today’s concern and tomorrow’s crisis,” *Environ Int*, vol. 158, p. 106936, Jan. 2022, doi: 10.1016/j.envint.2021.106936.

[4] A. C. Johnson *et al.*, “Identification and Quantification of Microplastics in Potable Water and Their Sources within Water Treatment Works in England and Wales,” *Environmental Science & Technology*, vol. 54, no. 19, pp. 12326–12334, Aug. 2020, doi: 10.1021/acs.est.0c03211.

[5] Biro Pusat Statistik (BPS), “persentase rumah tangga menurut provinsi tipe daerah dan sumber air minum layak.” Accessed: Jan. 06, 2023. [Online]. Available: https://www.bps.go.id/indicator/29/854/1/persentase-rumah-tangga-menurut-provinsi-tipe-daerah-dan-sumber-air-minum-layak.html.

DOI : https://doi.org/10.52158/jacost.v5i2.815

[6] H. M. Gomes, J. Read, A. Bifet, J. P. Barddal, and J. Gama, "Machine learning for streaming data: state of the art, challenges, and opportunities," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 6–22, Nov. 2019, doi: 10.1145/3373464.3373470.

[7] A. Nandy, C. Duan, and H. J. Kulik, "Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery," *Curr Opin Chem Eng*, vol. 36, p. 100778, Jun. 2022, doi: 10.1016/j.coche.2021.100778.

[8] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An Overview of Principal Component Analysis," *Journal of Signal and Information Processing*, vol. 04, no. 03, pp. 173–175, 2013, doi: 10.4236/jsip.2013.43b031.

[9] B. M. Salih Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 02, no. 01, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.003.

[10] V. Pratama and J. Tjen, "Entropy-based subset selection principal component analysis for diabetes risk factor identification," *J Emerg Investig*, 2023, doi: 10.59720/23-015.

[11] M. Tripathi and S. K. Singal, "Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India," *Ecol Indic*, vol. 96, pp. 430–436, Jan. 2019, doi: 10.1016/j.ecolind.2018.09.025.

[12] W. Yang, Y. Zhao, D. Wang, H. Wu, A. Lin, and L. He, "Using Principal Components Analysis and IDW Interpolation to Determine Spatial and Temporal Changes of Surface Water Quality of Xin'anjiang River in Huangshan, China," *Int J Environ Res Public Health*, vol. 17, no. 8, p. 2942, Apr. 2020, doi: 10.3390/ijerph17082942.

[13] S. Abdelaziz, M. I. Gad, and A. H. M. H. El Tahan, "Groundwater quality index based on PCA: Wadi El-Natrun, Egypt," *Journal of African Earth Sciences*, vol. 172, p. 103964, [22] Dec. 2020, doi: 10.1016/j.jafrearsci.2020.103964.

[14] G. Thanh Nguyen, "Evaluating Current Water Quality Monitoring System on Hau River, Mekong Delta, Vietnam Using Multivariate Statistical Techniques," *Applied Environmental Research*, pp. 14–25, Jan. 2020, doi: 10.35762/aer.2020.42.1.2.

[15] J. Tjen, F. Smarra, and A. D'Innocenzo, "An entropy-based sensor selection algorithm for structural damage detection," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, IEEE, Aug. 2020, pp. 1566–1571. doi: 10.1109/case48305.2020.9216828.

[16] I. M. Johnstone and D. Paul, "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1277–1292, Aug. 2018, doi: 10.1109/jproc.2018.2846730.

[17] J. B. Schreiber, "Issues and recommendations for exploratory factor analysis and principal component analysis," *Research in Social and Administrative Pharmacy*, vol. 17, no. 5, pp. 1004–1011, May 2021, doi: 10.1016/j.sapharm.2020.07.027.

[18] J. Tjen and V. Pratama, "Penentuan Jalur Diagnostik Penyakit Berbasis Konsep Pembelajaran Mesin: Studi kasus Penyakit Hepatitis C," *Journal of Applied Computer Science and Technology*, vol. 4, no. 2, pp. 124–130, Nov. 2023, doi: 10.52158/jacost.v4i2.556.

[19] F. Smarra, J. Tjen, and A. D'Innocenzo, "Learning methods for structural damage detection via entropy-based sensors selection," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 10, pp. 6035–6067, Mar. 2022, doi: 10.1002/rnc.6124.

[20] L. Wang, "Enhanced fault detection for nonlinear processes using modified kernel partial least squares and the statistical local approach," *Can J Chem Eng*, vol. 96, no. 5, pp. 1116–1126, Nov. 2017, doi: 10.1002/cjce.23058.

[21] A. Kadiwal, "Water Quality," 2023. Accessed: Jan. 06, 2023. [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water potability.

M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," 2020.