IJSA (2024), Vol. 8, No. 2, 202-214
DOI: https://doi.org/10.29244/ijsa.v8i2p202-214

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values*

Klasifikasi Kelayakan Sumber Air Minum di Jawa Barat Menggunakan XGBoost dan Analisis Klasterisasi Berdasarkan Nilai SHAP

Annisa Permata Sari[1], Billy[2], Denanda Aufadlan Tsaqif[3‡], Bagus Sartono[4], and Aulia Rizki Firdawanti[5]

[1,2,3,4,5] Department of Statistics, IPB University, Indonesia
‡corresponding author: denandaaufadlants@gmail.com

Abstract

Water is essential for meeting the basic needs of living organisms. In Indonesia, ensuring safe and quality drinking water is crucial for public health. However, in some regions, particularly in West Java Province, people still rely on unsuitable water sources, which can negatively impact health. The classification of water source suitability can be achieved using machine learning, such as the Extreme Gradient Boosting (XGBoost) model. XGBoost with feature selection is effective in improving prediction accuracy and minimizing overfitting. This study evaluates the performance of the XGBoost model in classifying household drinking water sources in West Java and uses the K-Means algorithm for cluster SHAP values to identify key characteristics of households with safe drinking water. The results show that the XGBoost model, using only 12 selected features, with an accuracy of 77.43% and an F1-Score of 80.17%, successfully classified 4187 households, with 2349 having safe drinking water and 1838 having unsuitable sources. SHAP value analysis identified location, water collection time, and monthly per capita expenditure as significant factors influencing water source suitability. Households with water sources inside the house's fence, a short water collection time, and high monthly per capita expenditure tend to have safe drinking water sources. There are 4 clusters formed, with cluster 1 and cluster 3 needing immediate quality of drinking water sources improvement with cluster 2 as an indicator of success. Cluster 4 consists of households with high expenditure, marking it as a potential household for the government to make water quality improvements.

Keywords: Drinking Water Sources, Feature Selection, Machine Learning Classification, Water Suitability, XGBoost.

* Received: Dec 2024; Reviewed: Dec 2024; Published: Dec 2024

1. Pendahuluan

Air merupakan salah satu unsur vital bagi makhluk hidup untuk memenuhi kebutuhan dasar hidupnya. Berbagai aktivitas manusia sehari-hari tidak dapat dipisahkan dari penggunaan air. Ketersediaan air sebagai kebutuhan minum yang aman dan berkualitas merupakan faktor penting untuk menciptakan kehidupan yang sehat. Air minum yang aman mencakup beberapa aspek yang harus dipenuhi, seperti berasal dari sumber yang layak, mudah diakses, tersedia saat dibutuhkan, serta memenuhi standar kualitas fisik, kimia, dan biologis (Direktorat Jenderal Pencegahan dan Pengendalian Penyakit, 2023).

Proporsi rumah tangga dengan sumber air minum layak di Indonesia sangat beragam di tiap daerahnya. Sebagai contoh, sekitar 93,86% rumah tangga memiliki akses terhadap sumber air minum layak (BPS, 2024a). Meski rumah tangga di Jawa Barat dengan akses sumber air minum layak terbilang cukup tinggi, tetapi masih belum mencapai Visi Indonesia Emas 2045, yakni 100% rumah tangga diseluruh wilayah di Indonesia memiliki akses sumber air bersih yang layak minum, sedangkan masih terdapat rumah tangga di Jawa Barat dengan sumber air minum yang tidak layak.

Permasalahan terkait sumber air minum yang tidak layak di Provinsi Jawa Barat sangat beragam, di antaranya adalah gangguan pencernaan seperti diare. Penyakit diare yang disebabkan oleh konsumsi air yang tidak layak minum dapat membahayakan nyawa manusia. Penyakit diare yang terjadi di Jawa Barat menjadi faktor terkuat dalam peningkatan angka kematian bayi dan balita (Nursantika et al., 2023). Permasalahan terebut dapat dicegah jika masyarakat mengetahui sumber air yang digunakan dalam kegiatan sehari-hari tersebut layak atau tidak untuk diminum. Oleh karena itu, diperlukan suatu pembelajaran yang dapat mendeteksi ataupun mengklasifikasikan layak atau tidaknya sumber air minum yang digunakan oleh masyarakat di Jawa Barat.

Pengklasifikasian sumber air minum dapat dilakukan dengan pendekatan pembelajaran mesin. Pendekatan pembelajaran mesin dapat dilakukan dengan berbagai macam metode, diantaranya yaitu dengan pengklasifikasi berbasis pohon keputusan seperti *Extreme Gradient Boosting* (XGBoost). XGBoost dinilai efektif untuk mengklasifikasikan kelayakan sumber air minum (Maulana et al., 2023). Penelitian sebelumnya yang membandingkan dua pendekatan pembelajaran mesin, yaitu XGBoost dan Naive Bayes pada data penyakit diabetes menunjukkan bahwa XGBoost unggul dengan akurasi klasifikasi sebesar 90.10% (Nasution et al., 2021). Selain itu, penelitian lain yang menggunakan XGBoost dalam klasifikasi kualitas air minum menunjukkan akurasi klasifikasi yang juga cukup tinggi, yaitu sebesar 82,29% (Maulana et al., 2023). Berdasarkan kedua penelitian tersebut, metode XGBoost menunjukkan kinerja yang cukup bagus dalam melakukan klasifikasi.

Maka dari itu, penelitian ini bertujuan untuk mengevaluasi kinerja model XGBoost dalam mengklasifikasikan kelayakan sumber air minum di Jawa Barat dan menganalisis hasil klasifikasi XGBoost dengan melakukan klasterisasi nilai SHAP menggunakan algoritma K-Means untuk mengetahui informasi penting terkait karakteristik rumah tangga dengan sumber air minum layak. Penelitian ini dinilai dapat memberikan kontribusi kepada masyarakat di Jawa Barat, mengingat bahwa belum adanya penelitian yang secara eksplisit menjelaskan tentang penngklasifikasian

sumber air minum yang layak dan tidak berdasarkan beberapa faktor yang dimiliki atau melekat pada masing-masing rumah tangga di provinsi Jawa Barat.

2. Metodologi

2.1 Data

Data yang digunakan pada penelitian ini merupakan dataset SUSENAS provinsi Jawa Barat tahun 2023 yang dikumpulkan oleh Badan Pusat Statistik (BPS) Indonesia yang diambil pada *website* SILASTIK (BPS, 2024b). Data observasi penelitian yang digunakan sebanyak 20934 rumah tangga dengan indikator output, yaitu status kelayakan sumber air minum dengan kode 0 merupakan sumber tidak layak dan kode 1 merupakan sumber layak. Peubah input yang digunakan dapat dilihat pada Tabel 1 sebagai berikut.

Tabel 1: Peubah penelitian.

Peubah	Keterangan Peubah	Tipe Data
X1	Perdesaan atau Perkotaan	Nominal
X2	Lokasi Sumber Air Minum	Nominal
X3	Waktu Mengambil Air Minum	Numerik
X4	Kekurangan Air Minum dalam Waktu 24 Jam	Nominal
X5	Status Sanitasi Layak	Nominal
X6	Kondisi Fisik Sumber Air Utama	Nominal
X7	Ketersediaan Air untuk Cuci Tangan	Nominal
X8	Ketersediaan Sabun	Nominal
X9	Jumlah Anggota Rumah Tangga	Numerik
X10	Pendidikan Kepala Rumah Tangga	Skala
X11	Kekhawatiran Tidak Cukup Makanan	Nominal
X12	Kehabisan Makanan	Nominal
X13	Gizi	Numerik
X14	Pengeluaran perkapita	Numerik
X15	Status Kepemilikan Bangunan	Nominal
X16	Luas Lantai Rumah Tempat Tinggal	Numerik
X17	Status Kepemilikan Lahan	Nominal
X18	Status Kepemilikan Motor	Nominal
X19	Status Kepemilikan Mobil	Nominal

2.2 Metode Penelitian

Tahapan analisis penelitian yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan eksplorasi data menggunakan analisis korelasi untuk melihat hubungan antar peubah penelitian serta menggunakan barchart untuk melihat sebaran kelas pada peubah status kelayakan sumber air minum.
2. Membagi data menjadi data latih sebanyak 80% dan data uji sebanyak 20% secara acak menggunakan state 42 dengan data latih digunakan untuk melakukan pemodelan, sedangkan data uji digunakan dalam evaluasi hasil pemodelan.
3. Melakukan klasifikasi pemodelan *Extreme Gradient Boosting* (XGBoost) yang didasarkan pada konsep *Gradient Boosting Decision Tree* (GBDT) dengan

mengkombinasikan konsep peningkatan dan optimisasi dari pembentukan *Gradient Boosting Machine* (GBM) (Friedman, 2001). Selain itu, XGBoost termasuk ke dalam teknik *ensemble* dengan membuat pohon baru yang akan mencoba memperbaiki kesalahan prediksi yang dibuat oleh gabungan pohon-pohon sebelumnya, sehingga menghasilkan model yang lebih kuat dan akurat (Syukron et al., 2020). Model XGBoost dipilih karena memiliki fungsi regularisasi sehingga dapat membangun sistem yang cepat dan dapat mengurangi terjadinya *overfitting* (Yulianti et al., 2022). Dalam pemodelan tersebut, perlu dilakukan *tuning parameter*, salah satunya dengan menggunakan metode *Bayesian Tree structured Parzen Estimator* (TPE). Metode *tuning parameter* tersebut dapat meningkatkan peluang untuk mencapai solusi yang baik dengan jumlah iterasi yang relatif kecil (Bergstra et al., 2015) dan memberikan dampak yang substansial pada performa klasifikasi (Bergstra & Bengio, 2012).

4. Melakukan *feature selection* untuk mengidenfitikasi peubah yang perlu diseleksi dengan menggunakan *Recursive Feature Elimination* (RFE).
5. Evaluasi hasil pengklasifikasian status kelayakan sumber air minum tersebut menggunakan nilai *accuracy* dan difokuskan dengan melihat nilai *F1-score*. *F1-score* dapat memberikan penilaian yang lebih baik terhadap kinerja model pada model yang memiliki ketakseimbangan kelas tersebut (Bobbitt, 2021) dengan formula sebagai berikut.

$$F1\ Score = \frac{2\left(\frac{TP}{TP+FP}\right)\left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}+\frac{TP}{TP+FN}\right)} \tag{1}$$

dan

$$Accuracy\ = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

dengan TP merupakan *true positive*, TN merupakan *true negative*, FP merupakan *false negative*, dan FN merupakan *false negative*. Dalam evaluasi hasil pengklasifikasian, dilakukan perbandingan hasil antara model awal menggunakan semua peubah dengan model setelah dilakukan seleksi peubah. Model yang terpilih merupakan model yang memiliki nilai *F1-score* dan *accuracy* tertinggi.

6. Menganalisis *variable importance* menggunakan nilai *Shapley Additive exPlanations* (SHAP), yaitu metode untuk memperoleh wawasan mengenai pengaruh setiap peubah bebas j pada amatan i $\left(X_{ij}\right)$ terhadap hasil prediksi model (Nguyen et al., 2021). Nilai SHAP berasal dari nilai Shapley dalam teori permainan kooperatif dimana pemain diibaratkan sebagai peubah bebas dan hadiah sebagai skor dugaan. Semua kemungkinan kombinasi peubah bebas dengan dan tanpa peubah ke-j dievaluasi untuk mendapatkan nilai SHAP. Nilai SHAP untuk peubah ke-j $\left(\phi_j\right)$ dihitung menggunakan formulasi sebagai berikut (Hart, 1989).

$$\phi_j(v) = \phi_j = \Sigma_{S\subseteq M\backslash\{j\}}\left(\frac{|S|!\,(M-|S|-1)!}{M!}\right)\big(v(S\cup\{j\})-v(S)\big), j = 1,\dots,M. \tag{3}$$

dengan subset peubah $S \subseteq M\backslash\{j\}$, M jumlah peubah, $v(S\cup\{j\})$ skor prediksi yang menyertakan seluruh peubah, dan $v(S)$ adalah skor prediksi tanpa peubah ke-j. Dengan SHAP dapat memperoleh visual tentang karakteristik setiap p peubah bebas memengaruhi prediksi ke-i, dan pengaruh ini tetap konsisten pada saat

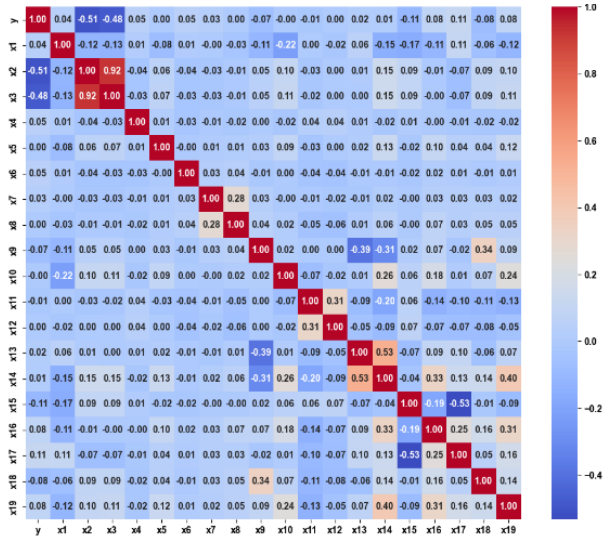
diagregat dengan semua prediksi. Hal inilah yang mendasari implementasi metode analisis SHAP untuk menggali wawasan dari hasil prediksi model XGBoost berdasarkan nilai SHAP pada seluruh peubah bebas untuk amatan ke-i.

7. Melakukan penggerombolan menggunakan *K-Means* yang merupakan metode non-hirarki yang dapat mengelompokkan peubah ke dalam *k* klaster. Setiap peubah akan dikelompokkan ke dalam suatu klaster berdasarkan titik pusat (*centroid*) klaster yang terdekat dalam peubah tersebut (Kodinariya & Makwana, 2013). Tahapan pembentukan klaster menggunakan *K-Means* pada penelitian ini adalah dengan menginisiasi jumlah klaster berdasarkan metode *elbow* dan nilai *silhoutte*, lalu setiap peubah akan dimasukkan ke dalam klaster tertentu berdasarkan nilai SHAP yang berdekatan.
8. Melakukan interpretasi dan menggali informasi penting berdasarkan klaster yang telah terbentuk.

3. Hasil dan Pembahasan

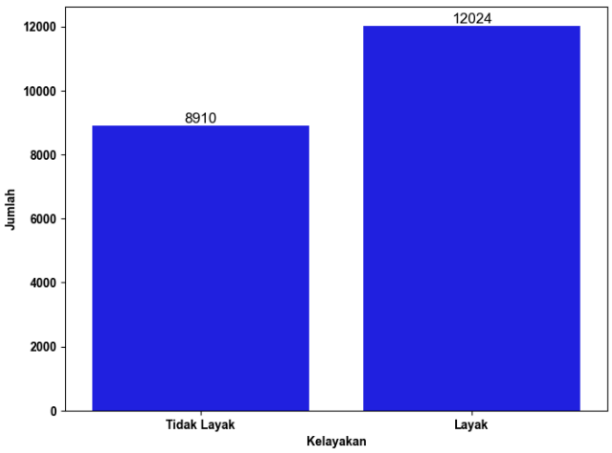
3.1 Eksplorasi Data

Sebelum melakukan analisis penelitian, pemahaman terhadap data penelitian diperlukan untuk memastikan bahwa data sudah siap untuk digunakan dengan melakukan eksplorasi data. Analisis korelasi digunakan untuk mengidentifikasi hubungan antar peubah penelitian dengan menggunakan metode spearman karena metode penelitian yang digunakan pada penelitian ini merupakan metode non-parametrik. Analisis korelasi ditampilkan pada Gambar 1 di bawah ini.



Gambar 1: Analisis Korelasi Spearman

Gambar 1 menunjukkan bahwa peubah lokasi sumber air minum dan waktu mengambil air minum merupakan peubah dengan hubungan negatif cukup kuat terhadap peubah status kelayakan sumber air minum, dengan nilai korelasi secara berurutan sebesar -0.51 dan -0.48. Namun, kedua peubah tersebut memiliki hubungan positif kuat antar satu sama lain dengan nilai korelasi sebesar 0.92. Selain itu, X15 dengan X17, X9 dengan X13, serta X9 dengan X14 juga berkorelasi cukup kuat antar satu sama lain. Selain itu, eksplorasi data digunakan untuk melihat sebaran data pada peubah status kelayakan sumber air minum agar dapat melihat keseimbangan kelas.



Gambar 2: Sebaran Status Kelayakan Sumber Air Minum

Gambar 2 menunjukkan perbedaan antara banyaknya rumah tangga yang memiliki sumber air minum layak dan tidak layak. Rumah tangga di Provinsi Jawa Barat dengan sumber air minum layak berjumlah 12024 rumah tangga, sedangkan rumah tangga dengan sumber air minum yang tidak layak berjumlah 8910 rumah tangga. Meskipun terdapat selisih sebesar 3114 rumah tangga, kedua nilai tersebut tidak berbeda secara signifikan, yang menunjukkan bahwa perbedaan status kelayakan sumber air minum rumah tangga di Provinsi Jawa Barat cenderung mendekati seimbang, sehingga tidak perlu dilakukan resampling pada status kelayakan sumber air minum.

3.2 Analisis Klasifikasi *Extreme Gradient Boosting*

Pengklasifikan dengan menggunakan model XGBoost dibentuk dengan menggunakan data latih yang perlu dilakukan *tuning parameter*. *Tuning parameter* merupakan salah satu cara untuk mencari kombinasi nilai parameter yang paling baik dalam melakukan klasifikasi. *Tuning parameter* dijalankan untuk mengurangi kesalahan prediksi klasifikasi dan meminimalkan masalah *overfit* pada model XGBoost (Wang & Ni, 2019). Tabel 2 menunjukkan beberapa parameter dengan nilai optimal yang dapat meningkatkan performa model XGBoost yang digunakan untuk pelatihan model pada data latih.

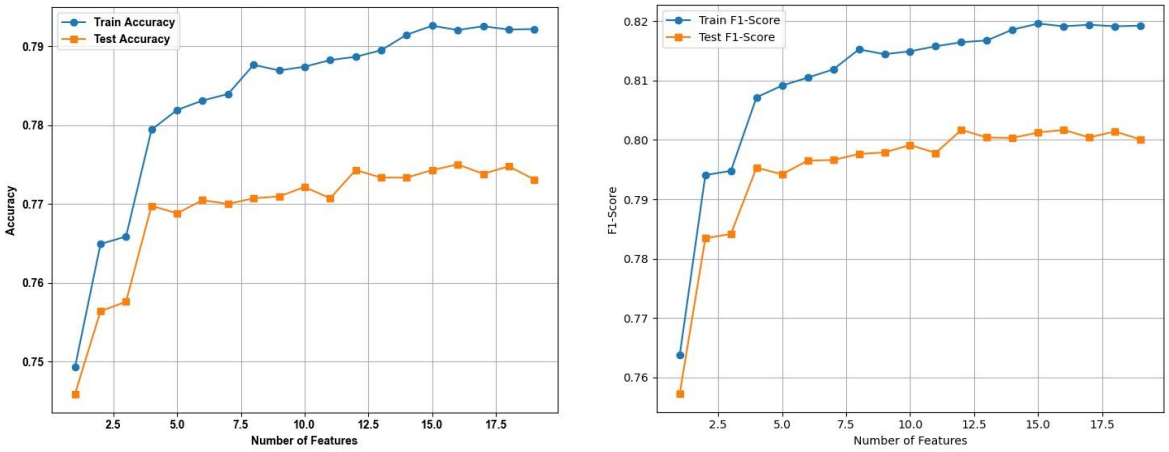
Tabel 2: Bayesian TPE *parameter tuning*.

Parameter	Nilai Bayesian	Nilai Parameter Terbaik
N estimator	100 – 500	394
Max depth	3 – 10	5
Learning rate	0.01 – 0.2	0.0174
Subsample	0.7 – 1.0	0.7075
Colsample bytree	0.7 – 1.0	0.8855

Model XGBoost yang dibentuk dari *tuning parameter* pada Tabel 2 dalam memprediksi klasifikasi kelayakan sumber air minum rumah tangga di Jawa Barat menghasilkan *confusion matrix* yang digunakan untuk melihat hasil pengklasifikasian model. *Confusion matrix* menghasilkan klasifikasi *true positive* sebanyak 1901 amatan, *true negative* sebanyak 1336 amatan, *false positive* sebanyak 435 amatan, dan *false negative* sebanyak 515 amatan.

3.3 Analisis *Feature Selection*

Feature selection perlu dilakukan untuk menghilangkan *redundant*, *noisy*, dan peubah-peubah yang tidak penting saat membangun model XGBoost (Prabha et al., 2021). *Feature selection* pada penelitian ini menggunakan *recursive feature elimination* (RFE). RFE bekerja dengan menggunakan *tuning parameter* XGBoost yang telah diperoleh sebelumnya, kemudian dilatih secara berulang dengan penambahan satu jumlah peubah dari pemodelan sebelumnya. Setelah dilakukan pelatihan berulang tersebut, diperoleh hasil sebagai berikut.



(a) (b)

Gambar 3: Perbandingan (a) nilai *accuracy* dan (b) *F1-Score* pada banyaknya sejumlah peubah

Gambar 3 menunjukkan banyaknya peubah yang menghasilkan nilai *accuracy* dan *F1-score* paling tinggi pada data uji berdasarkan *feature selection* adalah 12 peubah. Adapun peubah yang akan digunakan untuk melakukan pemodelan selanjutnya adalah peubah perdesaan atau perkotaan, lokasi sumber air minum, waktu mengambil air minum, kekurangan air minum dalam waktu 24 jam, ketersediaan sabun, jumlah anggota rumah tangga, pendidikan kepala rumah tangga, pengeluaran perkapita, status kepemilikan bangunan, status kepemilikan lahan, status kepemilikan mobil, dan status kepemilikan motor. *Confusion matrix* yang dihasilkan oleh model XGBoost dengan 12 peubah tersebut menghasilkan klasifikasi *true positive* sebanyak 1910 amatan, *true negative* sebanyak 1332 amatan, *false positive* sebanyak 439 amatan, dan *false negative* sebanyak 506 amatan.

3.4 Evaluasi Model *Extreme Gradient Boosting*

Setelah melakukan analisis pada model XGBoost dengan perbedaan banyak peubah yang digunakan, selanjutnya dilakukan evaluasi terhadap kedua model XGBoost tersebut. Hal ini bertujuan untuk menentukan model XGBoost paling baik dan banyaknya peubah yang akan digunakan untuk melakukan klasifikasi status kelayakan sumber air minum. Ukuran evaluasi model yang digunakan adalah metrik evaluasi, yaitu *accuracy*, *precision*, *sensitivity*, dan *F1-score* yang dihitung berdasarkan *confussion matrix* dari setiap model dengan hasil yang ditampilkan pada Tabel 3.

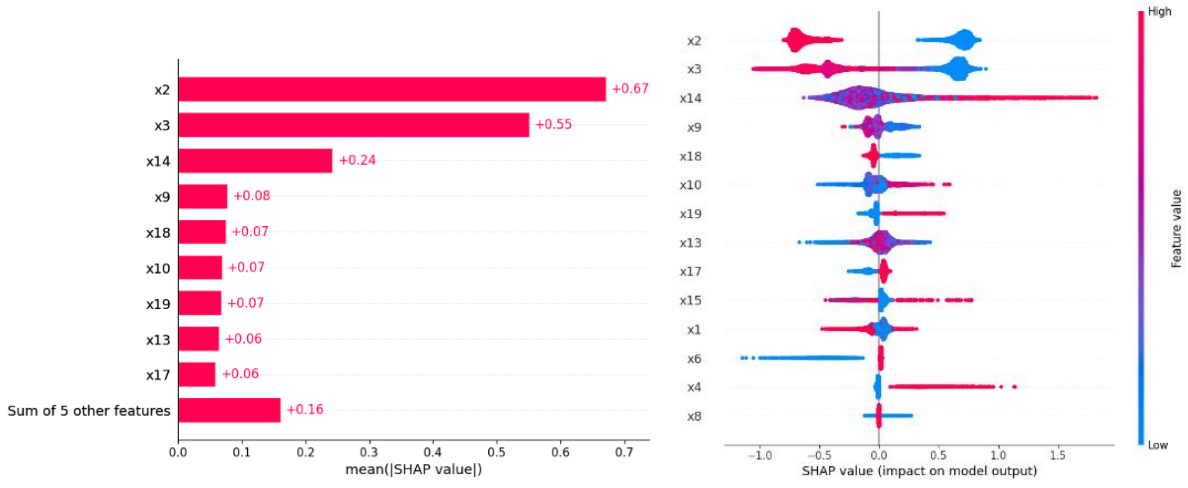
Tabel 3: Nilai metrik evaluasi XGBoost

Model	Nilai Metrik Evaluasi			
	Accuracy	*Precision*	*Sensitivity*	*F1-score*
XGBoost 19 Peubah	77.31%	81.38%	78.68%	80.01%
XGBoost 12 Peubah	77.43%	81.31%	79.06%	80.17%

Hasil metrik evaluasi XGBoost tersebut menunjukkan bahwa model XGBoost dengan 12 peubah memiliki hasil klasifikasi yang lebih baik daripada model XGBoost dengan 19 peubah. Oleh karena itu, model XGBoost dengan 12 peubah dapat digunakan untuk melihat peubah penting yang dapat berpengaruh terhadap pengklasifikasian dan untuk dilakukan analisis SHAP.

3.5 Analisis Nilai SHAP

Setelah mendapatkan model XGBoost paling baik, maka dapat dilanjutkan dengan menghitung nilai SHAP. Berdasarkan nilai SHAP, didapatkan peubah lokasi sumber air minum (X2), waktu ambil air minum (X3), dan rata-rata pengeluaran perkapita sebulan (X14) merupakan ketiga peubah paling penting berdasarkan Gambar 4(a). Berdasarkan Gambar 4(b), pada peubah X2 rumah tangga dengan lokasi sumber air di dalam pagar (ditandai dengan titik berwarna biru) mengumpul pada area dengan skor SHAP positif dan sebaliknya. Pada peubah X3, rumah tangga dengan waktu mengambil air minum yang singkat (ditandai dengan titik berwarna biru) mengumpul pada area dengan skor SHAP positif dan sebaliknya. Namun, terdapat beberapa rumah tangga dengan waktu pengambilan air minum yang singkat tetapi memiliki skor SHAP negatif, begitu pula sebaliknya. Pada peubah X14, rumah tangga dengan rata-rata pengeluaran perkapita sebulan yang rendah dan yang tinggi cenderung berada pada satu area SHAP yang sama. Tetapi, ada rumah tangga dengan pengeluaran perkapita sebulan yang tinggi yang mengumpul pada area dengan skor SHAP positif.

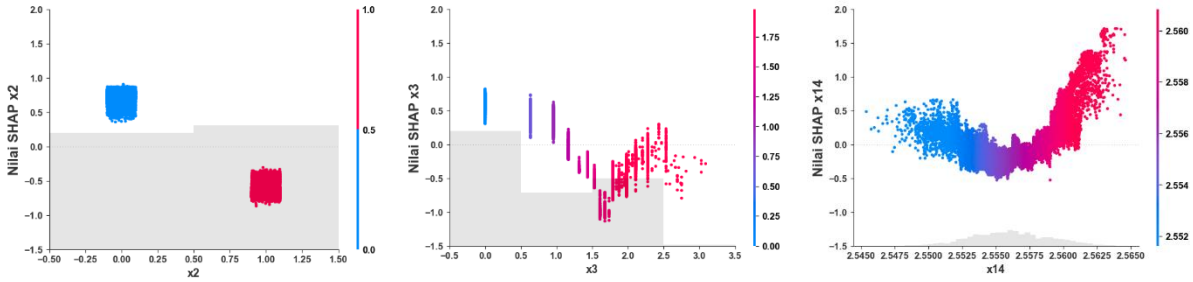


(a) (b)

Gambar 4: Urutan peubah penting berdasarkan metode SHAP (a) dan nilai SHAP setiap amatan pada seluruh peubah (b)

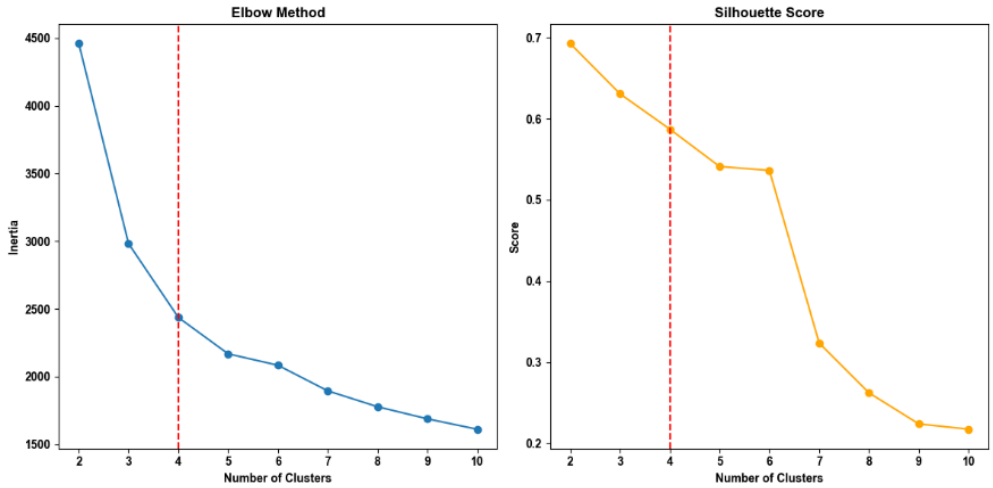
Setelah mendapatkan nilai SHAP seperti yang disajikan pada Gambar 4(b), nilai SHAP akan divisualisasikan dalam bentuk *scatterplot* untuk mempermudah dalam mengambil wawasan. Berdasarkan Gambar 5 peubah X2 memiliki pengaruh nilai SHAP yang berbeda antar kelas, sedangkan peubah X3 dan X14 memiliki pola pengaruh nilai SHAP yang *non-linear* dan cenderung berpola kuadratik. Berdasarkan Gambar 5(a), rumah tangga dengan lokasi sumber air minum di dalam pagar rumah (ditandai dengan titik berwarna biru) cenderung berkontribusi positif yang artinya rumah tangga dengan lokasi sumber air minum di dalam pagar rumah membuat model mengklasifikasikan ke rumah tangga dengan sumber air minum yang layak. Sementara itu, rumah tangga dengan lokasi sumber air minum di luar pagar rumah (ditandai dengan titik berwarna kemerahan) cenderung berkontribusi negatif yang artinya rumah tangga dengan lokasi sumber air minum di luar pagar rumah membuat model mengklasifikasikan rumah tangga dengan sumber air minum yang tidak layak.

Berdasarkan Gambar 5(b), rumah tangga dengan waktu mengambil air minum yang singkat cenderung berkontribusi positif terhadap klasifikasi kelayakan sumber air minum. Sebaliknya, rumah tangga dengan waktu mengambil air minum yang menengah dan lama umumnya berkontribusi negatif terhadap klasifikasi kelayakan sumber air minum, walaupun ada beberapa yang berkontribusi positif. Berdasarkan pada Gambar 5(c), rumah tangga dengan pengeluaran perkapita sebulan yang rendah dan tinggi berkontribusi positif terhadap klasifikasi kelayakan sumber air minum, sedangkan rumah tangga dengan pengeluaran perkapita sebulan yang menengah berkontribusi negatif terhadap klasifikasi kelayakan sumber air minum.



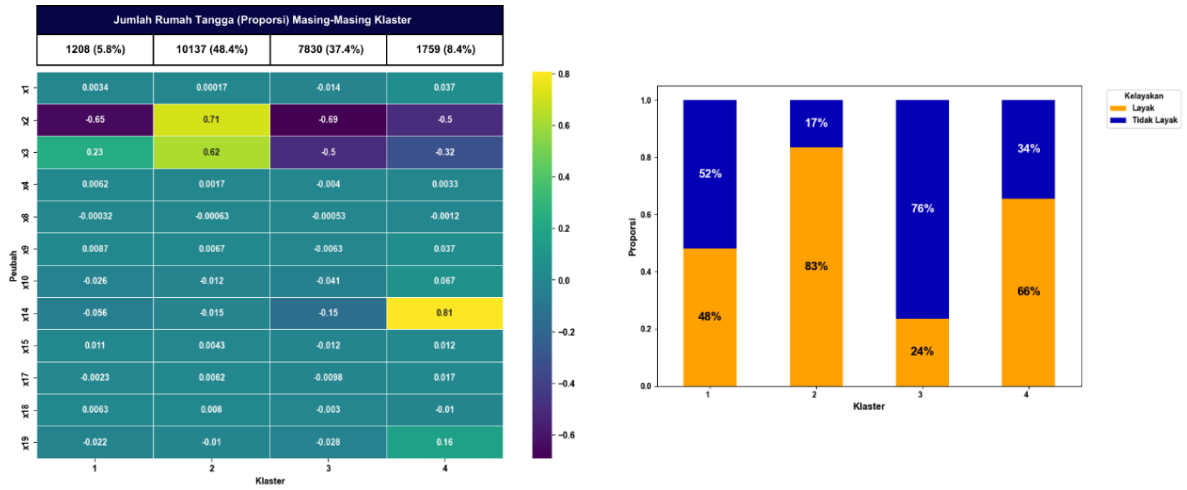
Gambar 5: Nilai SHAP pada peubah X2 (a), peubah X3 (b), dan peubah X14 (c)

Peubah X2 memiliki dua kelas dengan pengaruh nilai SHAP yang saling bertolak-belakang, yang artinya peubah ini dapat memisahkan pengaruh prediksi kelayakan sumber air minum dengan baik, seperti yang disajikan pada Gambar 5(a). Berdasarkan Gambar 5(b) dan Gambar 5(c), peubah X3 dan peubah X14 memiliki pola *non-linear* dan memiliki beberapa amatan dengan pengaruh nilai SHAP yang berbeda. Misalkan pada amatan dengan nilai X3 = 2.5, terdapat beberapa amatan dengan nilai SHAP negatif dan SHAP positif. Hal tersebut disebabkan oleh perbedaan karakteristik antar amatan, sehingga klasterisasi dapat diimplementasikan untuk mengidentifikasi karakteristik unik di setiap kelompok. Berdasarkan Gambar 6, sebanyak 4 klaster dapat dipilih, karena klaster tersebut memiliki *inertia* yang rendah dan *silhouette score* yang cukup tinggi.



Gambar 6: Pemilihan klaster optimum menggunakan metode *elbow* dan nilai *silhouet*

Klasterisasi dilakukan dengan menggunakan 4 klaster, pada Gambar 6. Banyak rumah tangga di setiap klaster disajikan pada Gambar 7(a). Sebanyak 88.8% rumah tangga berada di klaster 2 dan klaster 3, sementara klaster 1 dan klaster 4 hanya mencakup 11.2% rumah tangga. Klaster 2 dan 3 dapat disebut sebagai klaster mayoritas dan klaster 1 dan 4 sebagai klaster minioritas.



(a) (b)

Gambar 7: Karakteristik setiap klaster berdasarkan rata-rata nilai SHAP di setiap klaster dan Sebaran banyaknya rumah tangga (a) dan sebaran nilai pengeluaran perkapita dalam sebulan di setiap klaster (b)

Setiap klaster yang terbentuk memiliki peubah-peubah dengan kontribusi yang berbeda terhadap klasifikasi. Berdasarkan Gambar 7(a), peubah X2 berkontribusi negatif kuat dan X3 berkontribusi positif lemah pada klaster satu. Pada klaster kedua, peubah X2 dan X3 memiliki kontribusi positif kuat. Pada klaster ketiga peubah X2 dan X3 berkontribusi negatif kuat, serta kontribusi negatif lemah pada peubah X14. Klaster keempat menunjukkan adanya kontribusi negatif kuat pada peubah X2, negatif lemah pada Peubah X3, dan positif kuat Peubah X14. Berdasarkan Gambar 4(b) dan Gambar 7(a), klaster 4 didominasi oleh rumah tangga dengan kondisi ekonomi yang lebih sejahtera dibandingkan klaster lain. Klaster tersebut didominasi oleh Kabupaten

Bekasi (Kab. Bekasi), Kota Bandung, dan Kota Bekasi, yang memiliki laju pertumbuhan ekonomi yang tinggi (BPS, 2024c). Berdasarkan Gambar 9(b), perbaikan kualitas sumber air minum dapat difokuskan pada klaster 1 dan klaster 3 dengan klaster 2 sebagai indikator keberhasilan. Meskipun kualitas sumber air minum kurang memadai, kondisi ekonomi yang baik dapat menanggulangi permasalahan ketidaklayakan sumber air minum secara efektif, seperti pada klaster 4.

Tabel 4: Peringkat dominasi kabupaten atau kota pada setiap klaster

Peringkat Kota	Klaster			
	1	2	3	4
1	Kabupaten Cianjur	Kabupaten Bogor	Kabupaten Bandung	Kabupaten Bekasi
2	Kabupaten Karawang	Kabupaten Garut	Kabupaten Bekasi	Kota Bandung
3	Kabupaten Pangandaran	Kabupaten Sukabumi	Kabupaten Indramayu	Kota Bekasi

Kab. Bandung, Kab. Bekasi, Kab. Bogor, Kab. Cianjur, Kab. Garut, dan Kab. Sukabumi merupakan wilayah-wilayah terdampak kekeringan pada akhir tahun 2023 (Herdiana, 2023). Berdasarkan Tabel 4, kabupaten-kabupaten tersebut mendominasi klaster 1 atau 3, kecuali Kab. Bogor, Kab. Garut, dan Kab. Sukabumi yang memiliki proporsi tingkat layak sumber air minum yang tinggi dan mendominasi klaster 2, yang artinya kabupaten tersebut dapat mengatasi kekeringan dengan baik. Selain kekeringan, ketidaklayakan sumber air minum juga dapat berasal dari pencemaran lingkungan, seperti yang terjadi pada Kab. Bekasi (Sinulingga, 2023). Kab. Bekasi juga berada di klaster 4, sehingga pemerintah dapat bekerja sama dengan rumah tangga berpengeluaran tinggi untuk mengatasi masalah ketidaklayakan sumber air minum di Kab. Bekasi.

4. Simpulan dan Saran

Analisis yang telah dilakukan memperoleh model XGBoost dengan kombinasi *hyperparameter,* yaitu *N estimator* = 394, *max depth* = 5, *learning rate* = 0.0174, *subsample* = 0.7075, dan *colsample bytree* = 0.8855 serta hanya menggunakan 12 peubah merupakan model paling baik dalam melakukan pengklasifikasian sumber air minum layak di Jawa Barat dengan nilai akurasi dan *F1-score* sebesar 77.43% dan 80.17% sehingga dapat dilakukan klasterisasi berdasarkan nilai SHAP. Nilai SHAP terbagi atas 4 klaster dengan kontribusi yang berbeda terhadap klasifikasi kelayakan sumber air minum. Klaster 2 dan 3 merupakan klaster mayoritas serta klaster 1 dan 4 merupakan klaster minioritas. Pada klaster 1, peubah lokasi sumber air minum berkontribusi negatif kuat dan waktu mengambil air minum berkontribusi positif lemah. Pada klaster 2 kontribusi positif kuat terjadi pada kedua peubah tersebut. Sementara itu, pada klaster 3 menunjukkan kontribusi negatif kuat pada peubah kedua peubah tersebut dan kontribusi negatif lemah pada peubah pengeluaran perkapita. Pada klaster 4 kontribusi negatif kuat pada peubah lokasi sumber air minum, negatif lemah pada peubah waktu mengambil air minum, dan positif kuat peubah pengeluaran perkapita. Klaster 1 dan 3 didominasi oleh rumah tangga dengan sumber air minum

tidak layak, maka perlu dilakukan perbaikan lokasi sumber air minum pada kedua klaster, dengan menjadikan klaster 2 sebagai indikator keberhasilan. Kondisi ekonomi yang baik pada klaster 4 dapat menanggulangi masalah ketidaklayakan sumber air minum. Selain itu, pemerintah juga dapat bekerja sama dengan rumah tangga yang memiliki pengeluaran yang tinggi pada klaster 4 untuk meningkatkan kualitas sumber air minum.

Model XGBoost yang digunakan pada penelitian ini menghasilkan akurasi yang cukup baik, namun model tersebut masih dapat dikembangkan lebih lanjut untuk meningkatkan akurasinya. Penelitian selanjutnya dapat menggunakan parameter yang lebih banyak dengan rentang nilai yang lebih luas, dan menggunakan metode tuning parameter yang berbeda seperti GridSearch. Selain itu, dapat dilakukan metode klasterisasi yang lain, seperti Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) dalam melakukan klasterisasi nilai SHAP.

Daftar Pustaka

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2), 281-305.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 014008. https://doi.org/10.1088/1749-4699/8/1/014008

Bobbitt, Z. (2021). *F1 Score vs. Accuracy: Which Should You Use?.* Retrieved from https://www.statology.org/f1-score-vs-accuracy/

BPS. (2024a). *Persentase Rumah Tangga Menggunakan Layanan Sanitasi yang Dikelola Secara Aman Menurut Provinsi dan Tipe Daerah (Persen), 2023-2024*. Retrieved from https://www.bps.go.id/id/statistics-table/2/MjE3OSMy/persentase-rumah-tangga-menggunakan-layanan-sanitasi-yang-dikelola-secara-aman-menurut-provinsi-dan-tipe-daerah.html

BPS. (2024b). *Survei Sosial Ekonomi Nasional 2023 Maret (KOR)*. Retrieved from https://silastik.bps.go.id/v3/index.php/mikrodata/detail/ZnZSZms4aStzN2JUSVY1QklqZ08rdz09

BPS. (2024c). *Tinjauan Ekonomi Provinsi Jawa Barat 2023* (Issue ISSN : 2714-9218). BPS Provinsi Jawa Barat.

Direktorat Jenderal Pencegahan dan Pengendalian Penyakit. (2023). *Laporan Tahunan Pengamanan Kualitas Air Minum Tahun 2022*. Jakarta: Kementerian Kesehatan RI.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hart, S. (1989). Shapley value. In *Game theory* (pp. 210-216). London: Palgrave Macmillan UK. Retrieved from https://link.springer.com/book/10.1007/978-1-349-20181-5

Herdiana, I. (2023). *Bencana Kekeringan Melanda 23 Kabupaten dan Kota di Jawa Barat, Mengancam Sawah di Kabupaten Bandung | BandungBergerak.id*. Retrieved from https://bandungbergerak.id/article/detail/158922/bencana-kekeringan-melanda-23-kabupaten-dan-kota-di-jawa-barat-mengancam-sawah-di-kabupaten-bandung

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90-95.

Maulana, M. D., Hadiana, A. I., & Umbara, F. R. (2023). Algoritma Xgboost untuk klasifikasi kualitas air minum. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(5), 3251-3256. https://doi.org/10.36040/jati.v7i5.7308

Nasution, M. K., Saedudin, R. R., & Widartha, V. P. (2021). Perbandingan akurasi algoritma naïve bayes dan algoritma Xgboost pada klasifikasi penyakit diabetes. *eProceedings of Engineering*, *8*(5), 9765-9772.

Nguyen, H. T. T., Cao, H. Q., Nguyen, K. V. T., & Pham, N. D. K. (2021). Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference* (pp. 1-6).

Nursantika, M., Faridhan, Y. E., & Kamila, I. (2023). Analisis pengaruh faktor risiko penyakit pneumonia terhadap angka mortalitas bayi dan balita menggunakan regresi poisson dan regresi binomial negatif (studi kasus: Provinsi Jawa Barat). *Interval: Jurnal Ilmiah Matematika*, *3*(2), 102-111. https://doi.org/10.33751/interval.v3i2.9093

Prabha, A., Yadav, J., Rani, A., & Singh, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Computers in Biology and Medicine*, *136*, 104664. https://doi.org/10.1016/j.compbiomed.2021.104664

Sinulingga, B. (2023). *Kali Bekasi tercemar parah, ribuan pelanggan PDAM krisis air bersih*. Retrieved from https://www.liputan6.com/news/read/5402885/kali-bekasi-tercemar-parah-ribuan-pelanggan-pdam-krisis-air-bersih

Syukron, M., Santoso, R., & Widiharih, T. (2020). Perbandingan metode smote random forest dan smote xgboost untuk klasifikasi tingkat penyakit hepatitis C pada imbalance class data. *Jurnal Gaussian, 9*(3), 227-236. https://doi.org/10.14710/j.gauss.9.3.227-236

Wang, Y., & Ni, X. S. (2019). *A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization*. *arXiv preprint arXiv:1901.08433*.

Yulianti, S. E. H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan metode extreme gradient boosting (Xgboost) pada klasifikasi nasabah kartu kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21-26. https://doi.org/10.31605/jomta.v4i1.1792