Eksplorasi dan Klasifikasi K-NN Terhadap Kejadian Luar Biasa Diare di Jawa Barat

Exploration and Classification KNN for Diarrheal Epidemic Events in West Java

Tahira Fulazzaky[1], Yully Sofyah Waode[2], Anwar Fitrianto[3], Erfiani[4], Alfa Nugraha Pradana[5]
[1,2,3,4,5]Statistika dan Sains Data, IPB University
E-mail: [1]tahirafulazzaky@apps.ipb.ac.id, [2]yullysofyah.waode@apps.ipb.ac.id,
[3]anwarstat@gmail.com, [4]erfiani@apps.ipb.ac.id, [5]alfanugraha@apps.ipb.ac.id

Abstrak

Tujuan dari penelitian ini adalah untuk mengkaji bagaimana kualitas air dan sanitasi mempengaruhi Kejadian Luar Biasa (KLB) Diare di Provinsi Jawa Barat, Indonesia, menggunakan data Pendataan Potensi Desa (PODES) tahun 2021. Diare merupakan permasalahan serius dalam kesehatan masyarakat Indonesia, terutama pada kelompok anak balita, dan salah satu faktor penyebab utamanya adalah rendahnya kualitas air dan sanitasi. Dalam konteks penelitian ini, kami menerapkan metode algoritma *K-Nearest Neighbors* (K-NN) untuk mengklasifikasikan wilayah-wilayah yang mengalami KLB Diare. Hasil eksplorasi data menunjukkan variasi yang signifikan dalam jumlah kasus diare di sejumlah kabupaten dan kota yang tersebar di wilayah Jawa Barat. Untuk menangani masalah ketidakseimbangan data, kami menerapkan teknik Pengurangan Acak (*Random Under Sampling*), Penambahan Acak (*Random Over Sampling*), dan *Synthetic Minority Oversampling Technique* (SMOTE). Hasil analisis menunjukkan bahwa model K-NN dengan penggunaan metode SMOTE menghasilkan tingkat akurasi tertinggi, yaitu sebesar 71.28%. Meskipun demikian, nilai F1 score untuk semua model cenderung rendah, yang mengindikasikan adanya tantangan dalam mengklasifikasikan wilayah-wilayah dengan KLB Diare. Penelitian ini memberikan wawasan yang penting mengenai korelasi antara kualitas air, sanitasi, dan KLB Diare di Jawa Barat, serta mengidentifikasi wilayah-wilayah yang memerlukan perhatian lebih dalam upaya pencegahan dan pengendalian penyakit diare. Hasil ini dapat digunakan sebagai dasar untuk merancang program-program kesehatan yang lebih efektif di daerah-daerah dengan tingkat insiden diare yang tinggi.

Kata kunci: Algoritma K-Nearest Neighbors (K-NN), SMOTE, Ketidakseimbangan data dan teknik pengambilan sampel ulang, Kualitas air dan sanitasi, Program pencegahan dan pengendalian diare

Abstract

The objective of this research is to scrutinize the impact of water quality and sanitation on Acute Diarrheal Disease Outbreaks (ADDO) in West Java Province, Indonesia, utilizing data from the Village Potential Census (PODES) of the year 2021. Diarrhea is a serious public health issue in Indonesia, especially among young children, and poor water quality and sanitation are major contributing factors. In the context of this research, the K-Nearest Neighbors (K-NN) algorithm is employed to classify regions with ADDO.

The data exploration reveals significant variations in the number of diarrhea cases across different regencies and municipalities in West Java. To address the data imbalance issue, we apply three techniques, namely Random Under Sampling, Random Over Sampling, and Synthetic Minority Oversampling Technique (SMOTE). The findings indicate that the K-NN model with SMOTE achieves the greatest level of accuracy at 71.28%. However, F1 scores for all models tend to be low, indicating the challenge of classifying regions with ADDO. This study provides critical key observations regarding the correlation between water quality, sanitation, and ADDO in West Java, identifying areas that require more attention for diarrhea prevention and control programs. These findings serve as a foundation for designing more effective health programs in regions with high diarrhea incidence rates.

Keywords: K-Nearest Neighbors (K-NN) algorithm, SMOTE, Data imbalances and resampling techniques, Water quality and sanitation, Diarrhea prevention and control programs

1. PENDAHULUAN

Perkembangan teknologi yang pesat saat ini telah membawa manusia ke zaman dimana pemanfaatan teknologi pada hampir semua bidang kehidupan. Salah satunya pemanfaatan penting teknologi ialah kemampuan untuk memudahkan proses pengolahan dan komputasi data sesuai dengan tipe data yang diberikan [1]. Klasifikasi merupakan salah satu metode pengolahan data yang paling sering digunakan. Metode ini digunakan untuk meramalkan kategori dari suatu objek yang belum memiliki label kelasnya berdasarkan pola atau informasi yang ditemukan dalam data [2],[3]. Pada konteks klasifikasi, data yang digunakan dibagi menjadi dua bagian utama yaitu data latih untuk melatih algoritma klasifikasi yang digunakan dan data uji untuk mengetahui kinerja algoritma yang telah dilatih menggunakan data baru yang tidak terdapat dalam data latih [4]. *K-Nearest Neighbors* (K-NN) merupakan salah satu algoritma klasifikasi yang sederhana dan akurat.

Air adalah elemen penting bagi semua makhluk hidup, dan kualitas air bersih yang sesuai standar kesehatan sangat dibutuhkan [5],[6]. Hubungan erat antara kesehatan manusia dan air menjadi nyata, terutama saat air terkontaminasi dapat menjadi sarana penularan berbagai penyakit [6]. Diare adalah masalah kesehatan serius di Indonesia, khususnya pada anak balita, dengan angka kematian mencapai 7,7% dan tingkat prevalensi mencapai 25,2% [7],[5]. Provinsi Jawa Barat mencatat tingkat insiden diare tertinggi di Indonesia pada tahun 2021, dengan Kabupaten Bogor sebagai yang paling tinggi (91.434 kasus) [8]. Faktor utama dalam risiko diare adalah kualitas sumber air minum, dan hanya sekitar 91,83% rumah tangga di Jawa Barat memiliki akses ke air minum yang sesuai standar kesehatan. Oleh karena itu, penting untuk meningkatkan kualitas dan ketersediaan air minum yang aman, terutama di daerah dengan tingkat insiden diare yang tinggi.

Berbagai penelitian telah dilakukan sebelumnya mengenai hubungan antara kualitas air dan sanitasi dengan kejadian diare. Misalnya, penelitian oleh Woldemariam Merid et al. tahun 2023 menunjukkan bahwa akses ke air minum yang aman dan sanitasi dapat mengurangi beban penyakit diare [9]. Penelitian lain oleh Mebrahtom et al. tahun 2022 juga menemukan bahwa faktor-faktor seperti usia ibu, penyimpanan air minum yang tidak aman, tidak adanya praktek pengolahan air di rumah, sanitasi yang tidak memadai, penanganan feses anak yang tidak aman, dan manajemen limbah padat yang tidak tepat berkontribusi terhadap kematian bayi akibat diare [10].

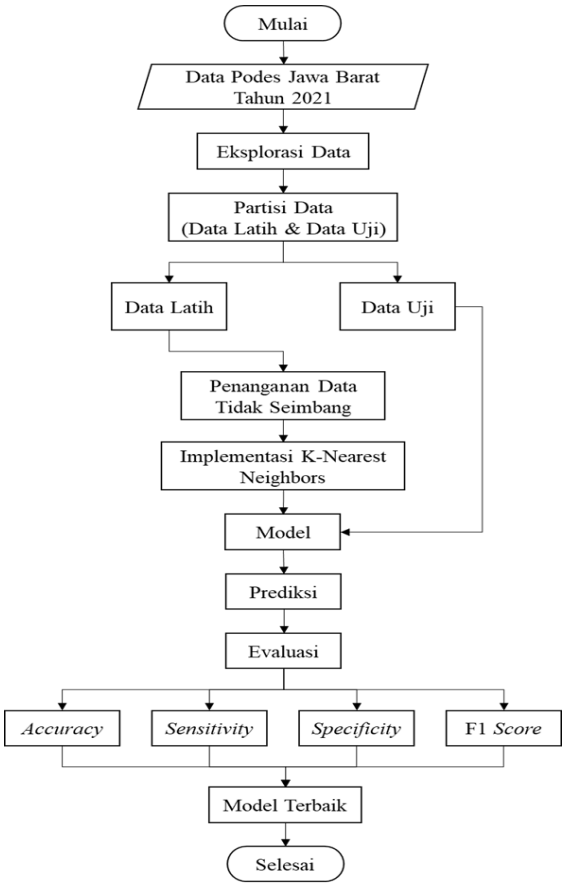
Namun, penelitian ini memiliki beberapa keunikan dibandingkan dengan penelitian sebelumnya. Pertama, penelitian ini menggunakan data Pendataan Potensi Desa (PODES) tahun 2021, yang merupakan sumber data yang sangat kaya dan belum banyak digunakan dalam penelitian sejenis. Kedua, penelitian ini menerapkan algoritma *K-Nearest Neighbors* (K-NN) dalam konteks klasifikasi Kejadian Luar Biasa (KLB) Diare, yang belum banyak dilakukan dalam penelitian sebelumnya. Ketiga, penelitian ini juga mencoba mengatasi masalah ketidakseimbangan data dengan menerapkan teknik Pengurangan Acak (*Random Under Sampling*), Penambahan Acak (*Random Over Sampling*), dan *Synthetic Minority Oversampling Technique* (SMOTE), yang merupakan pendekatan yang inovatif dalam konteks ini.

Penelitian ini bertujuan untuk melakukan eksplorasi dan visualisasi data Kejadian Luar Biasa (KLB) diare di Jawa Barat untuk memberikan penjelasan dan pemahaman yang lebih mendalam mengenai distribusi dan tren kejadian diare di wilayah tersebut. Selanjutnya dilakukan klasifikasi Kasus Luar Biasa (KLB) Diare menggunakan algoritma *K-Nearest Neighbors* (K-NN) dan mengidentifikasi faktor-faktor yang berkontribusi pada KLB tersebut. Dengan pemahaman yang lebih tentang hubungan antara kualitas air minum, sanitasi, dan faktor-faktor lain dengan insiden diare, diharapkan penelitian ini dapat membantu mendeteksi keberadaan KLB diare dan mengurangi jumlah insiden diare di Jawa Barat. Selain itu, hasil penelitian ini dapat memberikan

dasar bagi pengambilan keputusan dalam pengembangan program-program kesehatan yang lebih efektif dan terfokus, terutama di daerah-daerah yang membutuhkan intervensi kesehatan yang lebih intensif.

2. METODE PENELITIAN

Tahapan-tahapan pada penelitian ini dijelaskan pada diagram alir yang ditunjukkan dalam Gambar 1 berikut :



Gambar 1 Diagram Alir Penelitian

Proses klasifikasi kejadian luar biasa diare di Provinsi Jawa Barat diawali dengan eksplorasi univariat dan bivariat dari variabel-variabel data Podes yang digunakan. Selanjutnya data di partisi menjadi 2 subset yaitu data latih sebesar 80% dan data uji sebesar 20%. Kemudian dilakukan penanganan data tidak seimbang pada data latih. Tahapan berikutnya mengklasifikasikan data menggunakan metode *K-Nearest Neighbors* (K-NN) yang kemudian dilanjutkan menggunakan data uji untuk menguji dan memvalidasi keakuratan metode yang digunakan. Penelitian ini menggunakan akurasi, sensitivitas, spesifisitas dan *F1 score* untuk membandingkan performa K-NN dengan beberapa penanganan data tidak seimbang. Penelitian ini menggunakan *software RStudio* dalam melakukan eksplorasi dan klasifikasi data.

2.1 Data Penelitian

Data dalam penelitian ini merupakan data sekunder yang berasal dari kegiatan Pendataan Potensi Desa (PODES) 2021 Provinsi Jawa Barat. Data ini terdiri dari 5957 observasi dengan 884 variabel. Dalam studi ini, variabel-variabel yang diterapkan hanya yang berkaitan dengan ada atau tidaknya kejadian luar biasa diare di suatu wilayah, sumber dan penggunaan air bersih serta sanitasi lingkungan yaitu sejumlah 7 variabel. Variabel tersebut dapat dilihat dalam tabel berikut:

Tabel 1 Variabel Yang Digunakan Untuk Klasifikasi Kejadian Luar Biasa (KLB) Diare di Jawa Barat

Nama Variabel	Keterangan	Referensi
Y (Keberadaan KLB diare)	1 : Ada; 2 : Tidak Ada	[11] , [12]
X1 (Sumber air minum sebagian besar keluarga)	1 : Air kemasan bermerek; 2 : Air isi ulang; 3 : Ledeng dengan meteran (PAM/PDAM); 4 : Ledeng tanpa meteran; 5 : Sumur bor atau pompa; 6 : Sumur; 7 : Mata air; 8 : Sungai / danau / kolam / waduk / situ / embung / bendungan; 9 : Air hujan	[13] , [14]
X2 (Sumber air mandi/cuci sebagian besar keluarga)	1 : Ledeng dengan meteran (PAM/PDAM); 2 : Ledeng tanpa meteran; 3 : Sumur bor atau pompa; 4 : Sumur; 5 : Mata air; 6 : Sungai / danau / kolam / waduk / situ / embung / bendungan; 7 : Air hujan; 8 : Lainnya	[15]
X3 (Tempat pembuangan sampah sebagian besar keluarga)	1 : Tempat sampah yang kemudian diangkut; 2 : Dalam lubang atau dibakar; 3 : Sungai / saluran irigasi / danau / laut; 4 : Drainase (got / selokan); 5 : Lainnya	[16]
X4 (Tempat/saluran untuk membuang limbah cair yang bersumber dari air mandi/cuci sebagian besar keluarga)	1 : Lubang resapan; 2 : Drainase (got / selokan); 3 : Sungai / saluran irigasi / danau / laut; 4 : Dalam lubang atau tanah terbuka; 5 : Lainnya	[16]
X5 (Fasilitas buang air besar yang digunakan sebagian besar keluarga)	1 : Jamban sendiri; 2 : Jamban bersama; 3 : Jamban umum	[17] , [15]
X6 (Tempat untuk pembuangan akhir tinja yang digunakan sebagian besar keluarga)	1 : Tangki septik; 2 : IPAL; 3 : Kolam / sawah / sungai / danau / laut; 4 : Lubang tanah; 5 : Pantai / tanah lapang / kebun; 6 : Lainnya	[16] , [18]

2.2 K-Nearest Neighbors

Metode *K-Nearest Neighbors* (K-NN) merupakan metode yang melakukan klasifikasi berdasarkan pada jarak antara data baru dengan k data yang sudah ada. Jarak antara kedua data ini dapat ditentukan dengan menggunakan fungsi jarak seperti jarak Euclidean, Manhattan dan Monkowski [19]. Metode K-NN mengasumsikan bahwa data yang memiliki jarak dekat memiliki kemiripan atau keterkaitan yang tinggi. Metode ini juga termasuk dalam metode pembelajaran berbasis contoh (*instance-based learning*) yang tidak memerlukan proses pembelajaran sebelumnya, tetapi langsung menggunakan data yang ada sebagai dasar klasifikasi [20]. Cara kerja metode *K-Nearest Neighbors* (K-NN) adalah sebagai berikut [21]:

1. Menentukan parameter K, yaitu jumlah tetangga terdekat yang akan digunakan untuk klasifikasi.
2. Menghitung jarak antara data baru dengan setiap data yang ada.
3. Mengurutkan data yang ada berdasarkan jarak terkecil dengan data baru.
4. Memilih K data teratas sebagai tetangga terdekat dari data baru.

5. Menentukan kelas dari data baru sesuai dengan kelas mayoritas dari tetangga terdekat.

2.3 Performa Klasifikasi

Pada penelitian ini, performa klasifikasi diukur dengan menghitung akurasi, sensitivitas, spesifisitas dan *F1 score* menggunakan *confusion matrix*. *Confusion Matrix* menampilkan perbandingan antara nilai aktual dan nilai prediksi yang menghasilkan empat kemungkinan skenario, yakni *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Performa klasifikasi dihitung menggunakan rumus berikut [21],[22]:

1. Akurasi

$$\text{Akurasi} = (TP+TN)/(TP+FN+FP+TN) \quad (1)$$

2. Sensitivitas

$$\text{Sensitivitas} = TP/(TP+FN) \quad (2)$$

3. Spesifisitas

$$\text{Spesifisitas} = TN/(FP+TN) \quad (3)$$

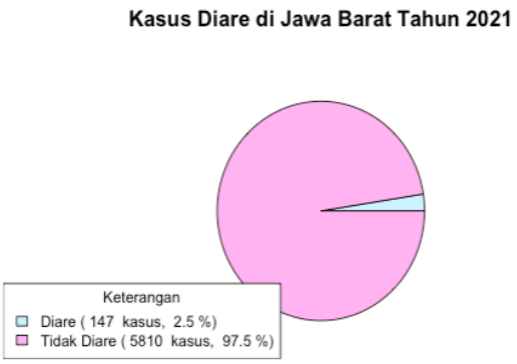
4. *F1 Score*

$$\text{Presisi} = TP/(TP+FP)$$
$$\text{Recall} = TP/(TP+FN) \quad (4)$$
$$\textit{F1 Score} = (2(\text{Presisi} \times \text{Recall}))/(\text{Presisi} + \text{Recall})$$

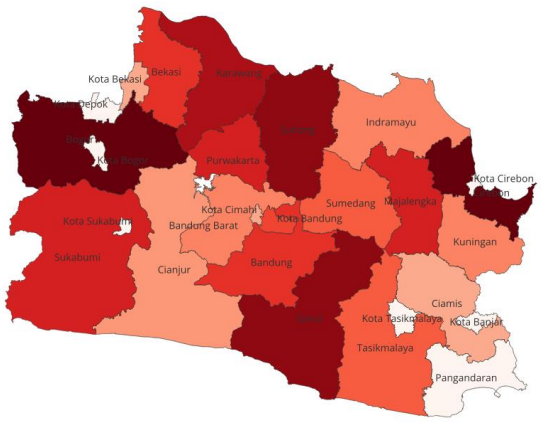
3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Data

Hasil dari eksplorasi univariat dari kasus Kejadian Luar Biasa (KLB) Diare di Jawa Barat Tahun 2021 dapat dilihat sebagai berikut:



(a)



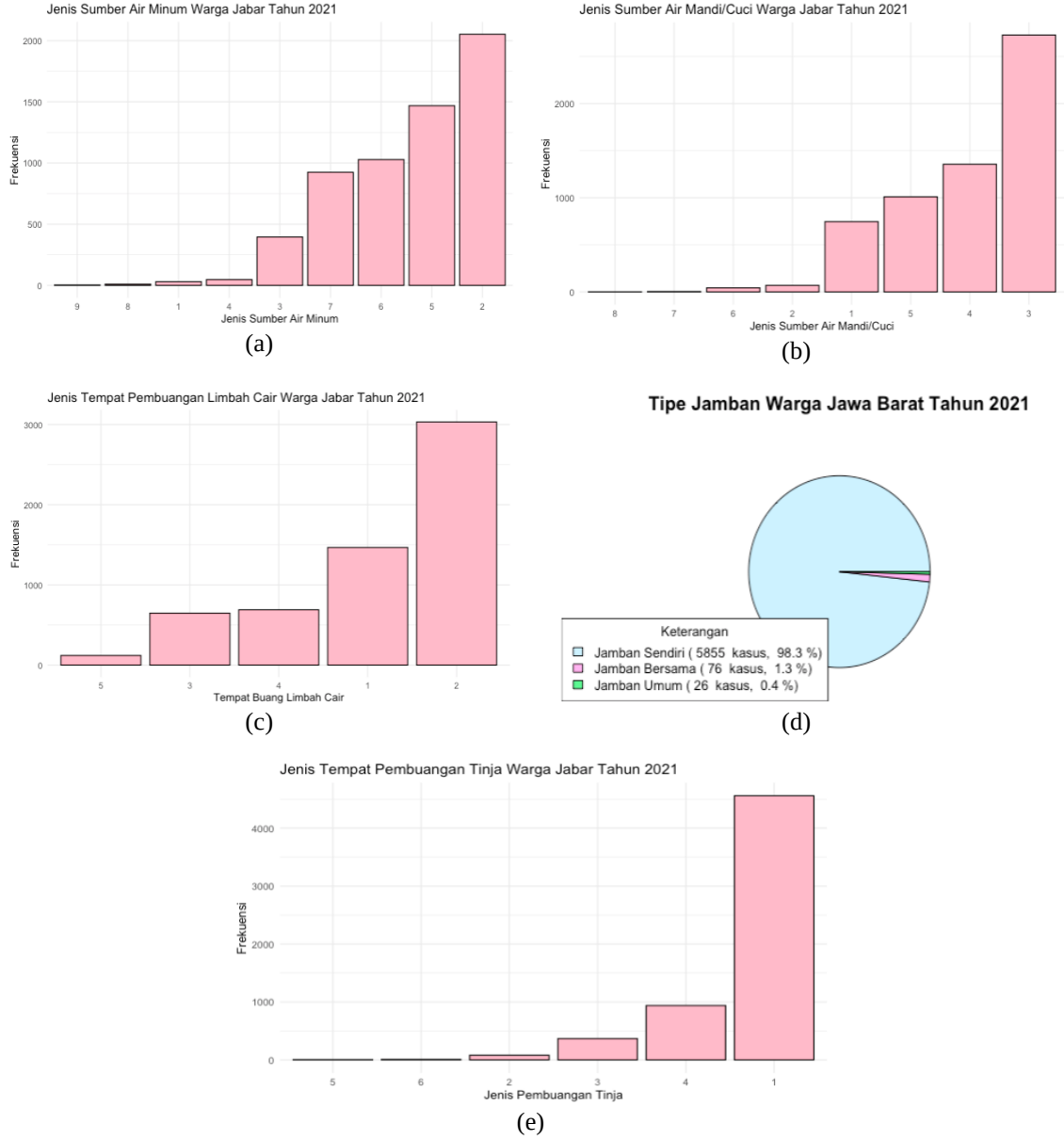
(b)

Gambar 2 (a) Pie Chart Distribusi Frekuensi Kejadian KLB Diare di Jawa Barat Tahun 2021; (b) Distribusi Peta Kejadian Diare di Jawa Barat Tahun 2021

Dari hasil eksplorasi univariat pada Gambar 2 (a) didapatkan jumlah kasus diare di Jawa Barat berdasarkan data PODES tahun 2021 sebesar 2.5% atau 147 kasus dan untuk jumlah warga yang tidak menderita diare sebesar 97.5% atau 5810 kasus.

Data kasus diare di Jawa Barat pada tahun 2021 menggambarkan variasi yang signifikan dalam jumlah kasus diare di berbagai kabupaten dan kota di provinsi ini seperti yang dapat dilihat pada Gambar 2 (b). Kabupaten/kota dengan jumlah kasus diare tertinggi adalah Kabupaten Cirebon dengan 17 kasus, Kabupaten Bogor dengan 16 kasus, serta Kabupaten Garut dan Kabupaten Subang masing-masing dengan 14 dan 13 kasus. Kabupaten lain seperti Bekasi,

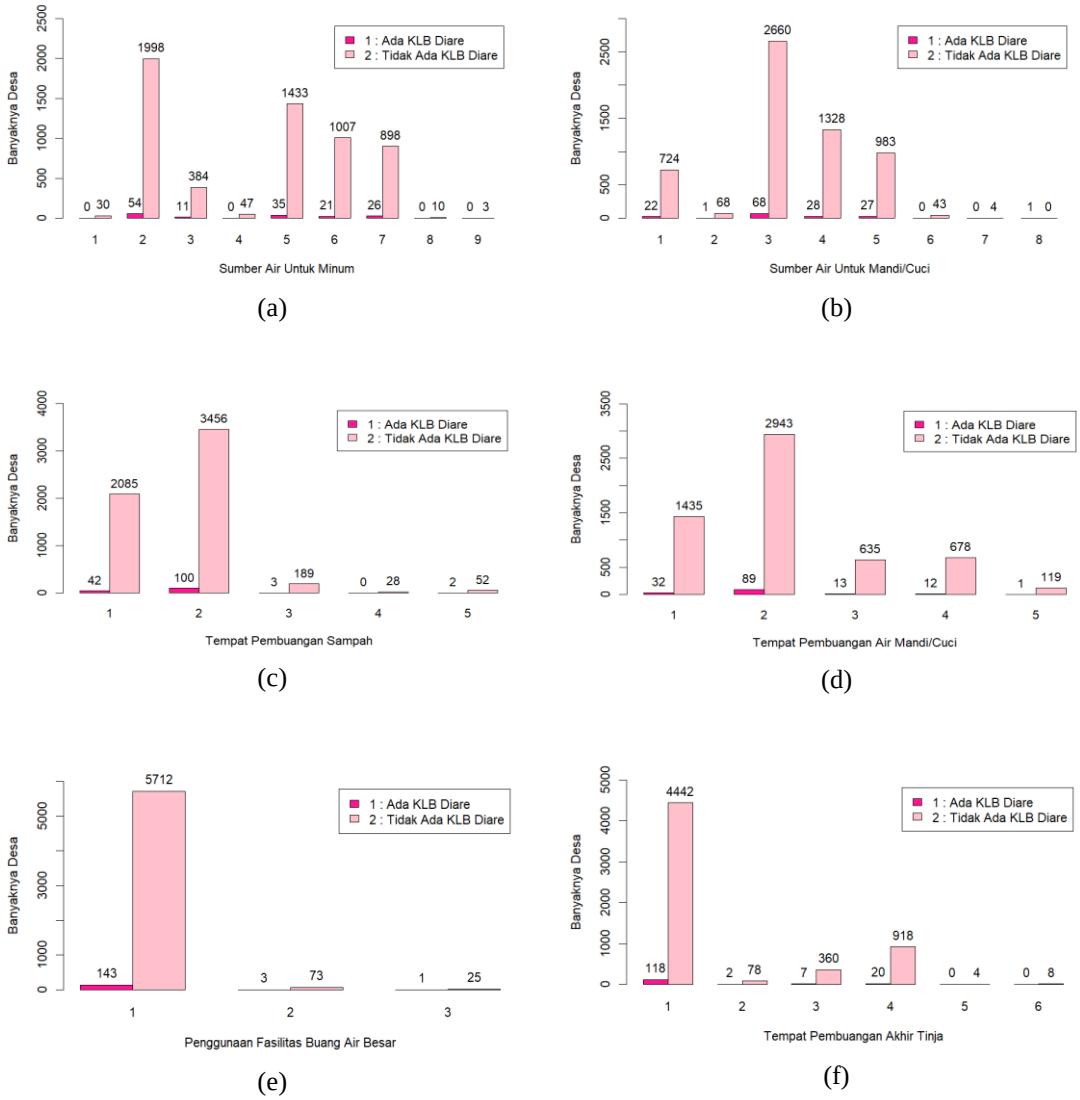
Karawang, Majalengka, dan Sukabumi juga melaporkan jumlah kasus yang cukup tinggi, berkisar antara 8 hingga 11 kasus. Di sisi lain, beberapa Kota seperti Kota Banjar, Kota Bogor, Kota Cirebon, Kota Depok, Kota Pangandaran, Kota Sukabumi, dan Kota Tasikmalaya tidak melaporkan adanya kasus diare pada tahun tersebut.



Gambar 3 (a) Bar Chart Distribusi Sumber Air Minum yang Digunakan Warga Jawa Barat Tahun 2021; (b) Bar Chart Distribusi Sumber Air Mandi/Cuci yang Digunakan Warga Jawa Barat Tahun 2021; (c) Bar Chart Distribusi Tempat Pembuangan Limbah Cair yang Digunakan Warga Jawa Barat Tahun 2021; (d) Pie Chart Penggunaan Fasilitas BAB yang Digunakan Warga Jawa Barat Tahun 2021; (e) Bar Chart Distribusi Tempat Pembuangan Akhir Tinja yang Digunakan Warga Jawa Barat Tahun 2021

Penggunaan fasilitas jamban yang biasa digunakan oleh warga Jawa Barat tahun 2021 yaitu sebagian besar menggunakan jamban milik sendiri sebanyak 98.3%. Lalu masih ada 76 warga atau sebesar 1.3% yang menggunakan jamban bersama dan ada 26 warga atau 0.4% yang menggunakan jamban milik umum.

Selanjutnya disajikan eksplorasi bivariat antara variabel terikat yaitu keberadaan KLB diare di seluruh desa di Jawa Barat dengan variabel bebas menggunakan barplot.



Gambar 4 Barplot eksplorasi bivariat antara variabel terikat yaitu keberadaan KLB diare di seluruh desa di Jawa Barat dengan variabel bebas

Berdasarkan data Podes 2021, terdapat Kejadian Luar Biasa Diare pada 147 desa di Provinsi Jawa Barat. Gambar 4 menunjukkan diantara desa desa tersebut, 36.73% (54 desa) diantaranya menggunakan air isi ulang sebagai sumber air untuk minum, 46.25% (68 desa) menggunakan sumur bor atau pompa sebagai sumber air untuk mandi/cuci, 68.02% (100 desa) menggunakan lubang atau dibakar sebagai tempat pembuangan sampah, 60.54% (89 desa) menggunakan drainase (got / selokan) sebagai tempat/saluran pembuangan limbah cair dari air mandi/cuci, 97.27% (143 desa) menggunakan jamban sendiri sebagai fasilitas buang air besar dan 80.27% (118 desa) menggunakan tangki septik sebagai tempat pembuangan akhir tinja.

3.2 Pengujian dan Analisis

Gambar 2 (a) menunjukkan bahwa variabel respon (*Y*) tidak seimbang sehingga dapat menyebabkan bias pada kelas mayoritas dan berpengaruh terhadap proses klasifikasi [23]. Oleh sebab itu setelah dilakukan pembagian data menjadi data latih dan data uji, perlu dilakukan penyeimbangan pada data latih. Teknik penanganan yang digunakan pada penelitian ini adalah

Random Under Sampling, *Random Over Sampling* dan *Synthetic Minority Oversampling Technique* (SMOTE).

Teknik *Random Under Sampling* merupakan teknik penanganan data tidak seimbang dengan mengambil beberapa data mayoritas sehingga jumlahnya sama dengan jumlah data minoritas. Teknik *Random Over Sampling* merupakan teknik penanganan data tidak seimbang dengan membuat data buatan dari data minoritas sebanyak data mayoritas [24]. Teknik SMOTE merupakan teknik penanganan data tidak seimbang yang mirip dengan teknik oversampling namun teknik ini akan membuat kelas minoritas lebih beragam dengan sampel baru yang mirip dengan data asli kelas minoritas [23].

Tabel 2 Perbandingan Performa K-NN dengan Beberapa Penanganan Data Tidak Seimbang

Metode	**k**	***Accuracy***	***Sensitivity***	***Specificity***	***F1 Score***
K-NN dengan *Under Sampling*	25	0.4055	0.79310	0.39587	0.06101
K-NN dengan *Over Sampling*	5	0.4752	0.72414	0.46902	0.06297
K-NN dengan SMOTE	5	0.7128	0.44828	0.71945	0.07065

Kriteria yang akan dibandingkan pada K-NN dengan beberapa penanganan data tidak seimbang adalah akurasi, sensitivitas, spesifisitas dan *F1 score*. Tabel 2 menampilkan ringkasan perbandingan performa K-NN dengan penanganan data tidak seimbang yang kemudian diterapkan pada data uji. Model K-NN dengan SMOTE menghasilkan nilai *accuracy*, *specificity* dan *F1 score* yang tertinggi sedangkan model K-NN dengan *Under Sampling* menghasilkan nilai *sensitivity* tertinggi.

Berdasarkan hasil Tabel 2 diperoleh informasi bahwa model K-NN dengan penanganan *Under Sampling* mampu memprediksi keberadaan KLB diare (*Y*) sebesar 40.55%, memprediksi wilayah yang ada KLB diare (Kelas 1) sebesar 79.31% dan memprediksi wilayah yang tidak ada KLB diare (Kelas 2) sebesar 39.58%. Selanjutnya model K-NN dengan penanganan *Over Sampling* mampu memprediksi keberadaan KLB diare (*Y*) sebesar 47.52%, memprediksi wilayah yang ada KLB diare (Kelas 1) sebesar 72.41% dan memprediksi wilayah yang tidak ada KLB diare (Kelas 2) sebesar 46.9%. Kemudian untuk model K-NN dengan penanganan SMOTE mampu memprediksi keberadaan KLB diare (*Y*) sebesar 71.28%, memprediksi wilayah yang ada KLB diare (Kelas 1) sebesar 44.82% dan memprediksi wilayah yang tidak ada KLB diare (Kelas 2) sebesar 71.94%.

Berdasarkan *F1 score*, ketiga model tersebut memiliki nilai mendekati 0 yang artinya ketiga model tersebut mungkin memiliki masalah serius dalam mengklasifikasikan Kelas 1 (ada KLB diare). Hal ini diakibatkan karena masalah ketidakseimbangan data yang sangat ekstrim dimana Kelas 1 (ada KLB diare) jauh lebih sedikit daripada Kelas 2 (tidak ada KLB diare).

4. KESIMPULAN DAN SARAN

Dalam penelitian ini, dilakukan analisis terhadap dampak kualitas air dan sanitasi lingkungan terhadap Kejadian Luar Biasa (KLB) Diare di Provinsi Jawa Barat, Indonesia, dengan menggunakan data Pendataan Potensi Desa (PODES) tahun 2021. Diare tetap menjadi masalah kesehatan masyarakat yang serius di Indonesia, terutama di kalangan balita, dengan kualitas air dan sanitasi yang buruk menjadi faktor kontributor signifikan. Penelitian ini menggunakan algoritma *K-Nearest Neighbors* (K-NN) untuk mengklasifikasikan wilayah-wilayah yang mengalami KLB Diare.

Hasil eksplorasi data mengungkapkan variasi yang signifikan dalam jumlah kasus diare di berbagai kabupaten dan kota di Jawa Barat. Terjadi ketidakseimbangan pada data keberadaan KLB diare di Jawa Barat. Untuk mengatasi ketidakseimbangan tersebut, berbagai teknik termasuk *Random Under Sampling*, *Random Over Sampling*, dan *Synthetic Minority*

Oversampling Technique (SMOTE) diterapkan. Analisis menunjukkan bahwa model K-NN dengan SMOTE mencapai tingkat akurasi tertinggi sebesar 71,28%. Meskipun begitu, nilai *F1 score* untuk semua model relatif rendah, menunjukkan adanya tantangan dalam mengklasifikasikan wilayah dengan KLB Diare.

DAFTAR PUSTAKA

[1] A. P. Wibawa, M. Guntur, A. Purnama, M. F. Akbar, dan F. A. Dwiyanto, "Metode-Metode Klasifikasi," *In Prosiding Seminar Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 1, 2018.
[2] P. A. Rahayuningsih, "Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Tingkat Kematian Dini Kanker dengan Dataset Early Death Cancer," *JTIK (Jurnal Teknik Informatika Kaputama)*, vol. 3, no. 2, hal. 1-10, 2019.
[3] Y. Azhar, A. K. Firdausy, dan P. J. Amelia, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke," *SINTECH (Science and Information Technology) Journal*, vol. 5, no. 2, hal. 191-197, 2022.
[4] E. Retnoningsih, dan R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insani Ict Journal*, vol. 7, no. 2, hal. 156-165, 2020.
[5] A. C. Lobo, "Tinjauan Yuridis Mengenai Dampak Pencemaran Air Terhadap Kesehatan Masyarakat Di Wilayah Poponcol Kabupaten Karawang," *Jurnal Justitia: Jurnal Ilmu Hukum dan Humaniora*, vol. 9, no. 3, hal. 1386-1394, 2022.
[6] Majelis Ulama Indonesia, "Air, Kebersihan, Sanitasi dan Kesehatan Lingkungan menurut Agama Islam (Vol. 2015)," Jakarta: Sekolah Pascasarjana Universitas Nasional.
[7] Kementerian Kesehatan Republik Indonesia, "Profil Kesehatan Indonesia Tahun 2018," 2018. [Online]. Tersedia: https://www.kemkes.go.id/downloads/resources/download/pusdatin/profil-kesehatan-indonesia/profil-kesehatan-indonesia-2018.pdf. [Diakses: 7 September 2023].
[8] Badan Pusat Statistik Provinsi Jawa Barat, "Persentase Rumah Tangga yang Memiliki Akses Terhadap Sumber Air Minum Layak di Jawa Barat," 2021. [Online]. Tersedia: https://jabar.bps.go.id/indicator/29/729/1/persentase-rumah-tangga-yang-memiliki-akses-terhadap-sumber-air-minum-layak-.html. [Diakses: 7 September 2023].
[9] W. M. Merid, A. Z. Alem, D. Chilot, D. G. Belay, A. A. Kibret, M. H. Asratie, Y. Y. Shibabaw, dan F. M. Aragaw, "Impact of Access to Improved Water and Sanitation on Diarrhea Reduction Among Rural Under-Five Children in Low and Middle-Income Countries: A Propensity Score Matched Analysis," *Tropical Medicine and Health*, vol.51, no. 1, hal. 1-10, 2023.
[10] S. Mebrahtom, A. Worku, dan D. J. Gage, "The risk of water, sanitation and hygiene on diarrhea-related infant mortality in eastern Ethiopia: a population-based nested case-control," *BMC Public Health*, vol. 22, no. 1, hal. 1-14, 2022.
[11] M. Dewi dan A. Hidayatullah, "Hubungan Faktor Lingkungan Dengan Kejadian Diare Pada Anak Balita Di Wilayah Kerja Puskesmas Landasan Ulin Kota Banjarbaru Tahun 2018," *Jurnal Ilmiah Kesehatan Diagnosis*, vol. 9, no. 2, hal. 137-142, 2018.
[12] A. Wulandari dan B. Setiawan, "Hubungan Antara Sumber Air Dengan Kejadian Diare Pada Warga Desa Kedungrejo Kecamatan Tanggul Kabupaten Jember," *Jurnal Administrasi Manajemen dan Kesehatan Masyarakat (JAMS)*, vol. 1, no. 1, hal. 1-8, 2019.
[13] R. P. Sari dan N. Suryani, "Hubungan Kualitas Sumber Air Minum dan Pemanfaatan Jamban Keluarga Dengan Kejadian Diare Pada Balita Di Desa Cikadu Kecamatan Cisarua Kabupaten Bandung Barat," *Jurnal Kesehatan Masyarakat*, vol. 6, no. 1, hal. 1-7, 2018.
[14] A. Wulandari dan N. Sari, "Hubungan Antara Sumber Air Dengan Kejadian Diare Pada Warga Desa Kedungrejo Kecamatan Tanggul Kabupaten Jember," *Jurnal Kesehatan Masyarakat*, vol. 4, no. 4, hal. 1-10, 2016.
[15] UNICEF Indonesia, "Hampir 70 persen sumber air minum rumah tangga di Indonesia tercemar limbah tinja," [Webpage]. Diakses tanggal 7 September 2023 dari https://www.unicef.org/indonesia/id/press-releases/hampir-70-persen-sumber-air-minum-rumah-tangga-di-indonesia-tercemar-limbah-tinja, 2022.

[16] Y. Hasneli, D. Karim, dan R. Woferst, "Identifikasi Dan Analisis Sarana Sanitasi Dasar Terhadap Kejadian Penyakit Diare Di Daerah Pesisir Provinsi Riau," *Repository Universitas Riau*, 2013. [Online]. Tersedia: https://repository.unri.ac.id/xmlui/handle/123456789/1234. [Diakses: 7 September 2023].
[17] B. Hamzah dan S. Hamzah, "Hubungan Penggunaan Air Bersih Dan Jamban Keluarga Dengan Kejadian Diare Pada Balita," *PREPOTIF: Jurnal Kesehatan Masyarakat*, vol. 5, no. 2, hal. 761-769, 2021.
[18] F. WOR dan M. Agusta, "Hubungan Air Bersih Dan Sanitasi Lingkungan Terhadap Kejadian Luar Biasa Diare," *Jurnal Endurance: Kajian Ilmiah Problema Kesehatan*, vol. 7, no. 3, hal. 24-291, 2022.
[19] M. Çakir, M. Yilmaz, M. A. Oral, H. Ö. Kazanci, dan O. Oral, "Accuracy Assessment of RFerns, NB, SVM, and KNN Machine Learning Classifiers in Aquaculture," *Journal of King Saud University-Science*, vol. 35, no. 6, hal. 1-6, 2023.
[20] M. F. Ardiansyah, "Implementasi Algoritma K-Nearest Neighbor Untuk Klasifikasi Data Penyakit Jantung," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 4, no. 2, hal. 237-248, 2018.
[21] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, hal. 29-33, 2020.
[22] R. J. Alfirdausy, dan S. Bahri, "Implementasi Algoritma K-Nearest Neighbor untuk Klasifikasi Diagnosis Penyakit Alzheimer," *Techno. Com*, vol. 22, no. 3, hal. 635-642, 2023.
[23] L. Sari, A. Romadloni, dan R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, vol. 14, no. 1, hal. 155-162, 2023.
[24] R. Mohammed, J. Rawashdeh, dan M. Abdullah, "Machine Learning With Oversampling And Undersampling Techniques: Overview Study And Experimental Results," dalam *2020 11th International Conference On Information And Communication Systems (ICICS)*, hal. 243-248.