JOINED JOURNAL
JOURNAL OF INFORMATICS EDUCATION

Optimalisasi Prediksi Kasus Demam Berdarah Dengue (DBD) Menggunakan Regresi Logistik Biner dengan Pendekatan SMOTE dan Tuning Hyperparameter

Rosa Ratri Kusuma Hariningsih, Diwahana Mutiara Candrasari, Endang Setyawati

Prodi Teknik Multimedia dan Jaringan, Sekolah Tinggi Ilmu Komputer Yos Sudarso Purwokerto, Indonesia

Info Articles

Keywords:

Demam Berdarah Dengue; Regresi Logistik; SMOTE; Hyperparameter Tuning; Data Tidak Seimbang.

Abstrak

Demam Berdarah Dengue (DBD) merupakan penyakit endemis yang masih menjadi tantangan kesehatan masyarakat di Indonesia. Deteksi dini terhadap potensi munculnya kasus DBD sangat krusial untuk penanggulangan yang cepat dan tepat. Penelitian ini bertujuan untuk mengembangkan model prediksi kasus DBD menggunakan Regresi Logistik Biner dengan penanganan data tidak seimbang melalui teknik Synthetic Minority Over-sampling Technique (SMOTE) dan optimasi model menggunakan hyperparameter tuning. Dataset yang digunakan mencakup data cuaca dan kasus DBD di wilayah Purwokerto tahun 2022–2024. Hasil penelitian menunjukkan bahwa setelah dilakukan penyeimbangan data dan tuning parameter, model mampu mencapai akurasi validasi silang sebesar 84,12%, meskipun akurasi pada data uji menurun menjadi 64%. Meskipun demikian, pendekatan ini menunjukkan potensi dalam pemodelan prediktif kasus DBD yang lebih akurat dan inklusif.

Abstract

Dengue Hemorrhagic Fever (DHF) remains an endemic disease and a significant public health challenge in Indonesia. Early detection of potential DHF outbreaks is crucial for timely and effective intervention. This study aims to develop a predictive model for DHF cases using Binary Logistic Regression, addressing data imbalance through the Synthetic Minority Over-sampling Technique (SMOTE) and optimizing model performance via hyperparameter tuning. The dataset comprises weather variables and DHF incidence data from the Purwokerto region spanning 2022 to 2024. The

results demonstrate that after data balancing and parameter tuning, the model achieved a cross-validation accuracy of 84.12%, although performance declined on the test set to 64%. Despite this decrease, the approach shows promise in enhancing predictive modeling for DHF cases by improving accuracy and inclusiveness.

Alamat Korespondensi: Jln. SMP 5 Karang Klesem, Purwokerto 53144
E-mail: rosaratri23@gmail.com

p-ISSN 2621-9484
e-ISSN 2620-8415

PENDAHULUAN

Demam Berdarah Dengue (DBD) merupakan penyakit tropis yang disebabkan oleh virus dengue dan ditularkan melalui gigitan nyamuk *Aedes aegypti*. Penyakit ini bersifat endemis dan masih menjadi tantangan serius bagi sistem kesehatan di Indonesia, terutama karena peningkatan kasus yang cenderung terjadi secara musiman selama musim penghujan. Upaya pengendalian DBD selama ini masih cenderung bersifat reaktif, seperti fogging dan pemberantasan sarang nyamuk, yang hanya dilakukan setelah munculnya kasus. Padahal, pendekatan prediktif berbasis data dapat menjadi solusi yang lebih efektif dalam upaya pencegahan dini dan pengambilan keputusan oleh pemangku kebijakan kesehatan masyarakat.

Sejumlah penelitian telah mengkaji penggunaan model prediktif berbasis pembelajaran mesin (*machine learning*) untuk mendeteksi potensi kasus DBD. Sari, Permana, dan Lestari (2022) menggunakan algoritma *Random Forest* untuk prediksi kasus DBD berdasarkan data iklim, sementara Wijaya dan Nugroho (2018) memanfaatkan model XGBoost untuk prediksi kejadian DBD di wilayah tropis. Namun, sebagian besar studi tersebut belum secara spesifik menangani permasalahan ketidakseimbangan kelas dalam data, yaitu jumlah data kasus DBD yang jauh lebih sedikit dibandingkan data tanpa kasus. Ketidakseimbangan ini menyebabkan model menjadi bias terhadap kelas mayoritas dan menurunkan kemampuan deteksi terhadap kelas minoritas. Selain itu, belum banyak studi yang secara khusus menyoroti penerapan model prediktif ini dalam konteks lokal seperti Purwokerto, yang memiliki dinamika iklim dan pola penyebaran penyakit tersendiri.

Penelitian ini bertujuan untuk mengembangkan model prediksi kasus DBD berbasis Regresi Logistik Biner yang dikombinasikan dengan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan data, serta optimasi performa melalui tuning hyperparameter. Studi dilakukan dengan menggunakan data kejadian DBD dan data iklim (suhu, kelembaban, curah hujan, dan bulan) di empat kecamatan di Purwokerto selama periode 2022–2024. Kontribusi kebaruan dari penelitian ini adalah pada integrasi pendekatan penyeimbangan data dan optimasi model dalam konteks lokal, yang belum banyak dibahas dalam literatur sebelumnya. Diharapkan model ini dapat menjadi dasar sistem peringatan dini yang lebih adaptif dan responsif dalam mendeteksi potensi kasus DBD di wilayah tropis Indonesia.

METODE

Lokasi dan Data Penelitian

Penelitian ini dilakukan dengan menggunakan data sekunder dari wilayah administratif Purwokerto, yang mencakup empat kecamatan: Purwokerto Utara, Purwokerto Selatan, Purwokerto Barat, dan Purwokerto Timur. Dataset mencakup data cuaca dan data kasus Demam Berdarah Dengue (DBD) bulanan dari Januari 2022 hingga Maret 2024. Data cuaca diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dan terdiri atas variabel suhu udara (°C), kelembaban udara (%), curah hujan (mm), dan bulan pencatatan. Sementara itu, data jumlah kasus DBD diperoleh dari Dinas

Kesehatan setempat dan dinyatakan dalam satuan jumlah kasus per bulan di setiap kecamatan.

Pra-Pemrosesan Data

Data awal melalui tahap pembersihan dengan mengidentifikasi nilai hilang pada seluruh variabel. Nilai hilang pada data cuaca diimputasi menggunakan nilai rata-rata berdasarkan bulan dan kecamatan terkait. Jika data kasus DBD pada bulan tertentu tidak tersedia dan proporsi hilangnya melebihi 30%, maka data tersebut dihapus dari analisis. Target variabel berupa status risiko DBD dibentuk melalui binarisasi, dengan nilai 1 jika jumlah kasus DBD lebih dari 5 dan nilai 0 jika sama dengan atau kurang dari 5. Variabel bulan dikodekan ke dalam format one-hot encoding untuk menangkap pola musiman secara eksplisit dan menghindari interpretasi ordinal yang keliru.

Pembagian Data dan Penanganan Ketidakseimbangan

Dataset yang telah dibersihkan dibagi menjadi data pelatihan (70%) dan data pengujian (30%) menggunakan metode stratified sampling agar distribusi kelas pada target tetap seimbang di kedua subset. Ketidakseimbangan kelas pada data pelatihan diatasi dengan menerapkan metode Synthetic Minority Over-sampling Technique (SMOTE), sesuai dengan prosedur yang dikembangkan oleh Chawla et al. (2002). SMOTE menghasilkan sampel sintetis dari kelas minoritas berdasarkan interpolasi data dengan k-nearest neighbors (k=5), yang meningkatkan representasi kelas tanpa melakukan duplikasi.

Transformasi dan Standarisasi Fitur

Seluruh fitur numerik (suhu, kelembaban, dan curah hujan) distandarisasi menggunakan *StandardScaler* dari pustaka *scikit-learn*, sehingga masing-masing fitur memiliki rata-rata nol dan deviasi standar satu. Transformasi ini bertujuan untuk mengoptimalkan kinerja algoritma Regresi Logistik, khususnya ketika regularisasi diterapkan.

Pemodelan dan Tuning Hyperparameter

Algoritma klasifikasi yang digunakan adalah Regresi Logistik Biner dengan solver saga, yang mendukung regularisasi L1 dan L2 serta efisien untuk dataset yang telah melalui one-hot encoding. Tuning parameter dilakukan menggunakan metode *GridSearchCV* dengan validasi silang sebanyak 5 lipatan (*k*-fold cross-validation, *k*=5). Hyperparameter yang dievaluasi mencakup nilai *C* (inverse dari kekuatan regularisasi) dalam rentang {0.01, 0.1, 1, 10, 100}, dan jenis regularisasi (‘l1’, ‘l2’). Pemilihan konfigurasi optimal didasarkan pada nilai rata-rata tertinggi dari skor F1 pada data validasi silang.

Evaluasi Kinerja Model

Kinerja model dievaluasi menggunakan data pengujian yang tidak terlibat dalam pelatihan model. Dua skenario threshold klasifikasi digunakan: default threshold 0.5 dan custom threshold 0.3 untuk meningkatkan sensitivitas deteksi. Metrik evaluasi yang digunakan meliputi:

a. Akurasi (Accuracy)
b. Presisi (Precision)
c. Sensitivitas (Recall)
d. F1-*Score*

HASIL DAN PEMBAHASAN

Deskripsi Data

Data yang digunakan terdiri dari 36 bulan pengamatan (Januari 2022 sampai Desember 2024) dengan variabel suhu rata-rata, kelembaban relatif, curah hujan, dan

bulan sebagai fitur, serta jumlah kasus DBD sebagai target. Distribusi kelas target setelah binarisasi adalah:

- Kelas 0 (DBD ≤ 5 kasus): 70%
- Kelas 1 (DBD > 5 kasus): 30%

Hal ini menunjukkan adanya ketidakseimbangan kelas yang dapat mempengaruhi performa model.

Pra-pemrosesan dan Penanganan Data Imbalance

Setelah melakukan one-hot encoding pada variabel bulan, fitur numerik distandarisasi menggunakan StandardScaler. SMOTE berhasil menyeimbangkan data pelatihan dengan menambah sampel sintetis untuk kelas minoritas, sehingga jumlah kelas 0 dan 1 menjadi seimbang. Hal ini penting untuk menghindari bias model terhadap kelas mayoritas.

```
============================================================
Distribusi kelas sebelum SMOTE: {0: 85, 1: 15}
Distribusi kelas setelah SMOTE: {0: 85, 1: 85}
============================================================
```

Gambar 1. Pra Pemrosesan data imbalance

Hasil Tuning Hyperparameter

GridSearchCV menemukan kombinasi hyperparameter terbaik pada model Regresi Logistik:

- c = 1 (regularisasi sedang)
- penalty = 'l2'

Model ini memberikan keseimbangan optimal antara bias dan varians.

```
==================== Hasil Tuning Hyperparameter ====================
Best parameters: {'C': 1, 'penalty': 'l2'}
Best CV accuracy: 0.8412
============================================================
```

Gambar 2. Hasil tuning

Evaluasi Model pada Data Pengujian

Model klasifikasi diuji pada dataset berukuran 44 sampel dengan dua kelas:

- Kelas 0 (tidak berisiko / DBD ≤ 5 kasus) sebanyak 37 sampel
- Kelas 1 (berisiko / DBD > 5 kasus) sebanyak 7 sampel

```
==================== Evaluasi Model (Threshold 0.5) =========
Akurasi pada test set: 0.64

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.68      0.76        37
           1       0.20      0.43      0.27         7

    accuracy                           0.64        44
   macro avg       0.53      0.55      0.52        44
weighted avg       0.76      0.64      0.68        44

============================================================
...
   macro avg       0.55      0.60      0.53        44
weighted avg       0.78      0.61      0.66        44

============================================================
```

Gambar 3. Hasil evaluasi model

Evaluasi terhadap model klasifikasi dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score pada masing-masing kelas, serta nilai rata-rata makro dan tertimbang (macro average dan weighted average). Berdasarkan hasil pengujian terhadap data uji, model memperoleh nilai akurasi sebesar 0,64, yang menunjukkan bahwa sebanyak 64% prediksi model sesuai dengan label aktual. Namun demikian, dalam konteks klasifikasi dengan distribusi kelas yang tidak seimbang, akurasi bukanlah indikator yang sepenuhnya representatif karena dapat terdistorsi oleh dominasi kelas mayoritas.

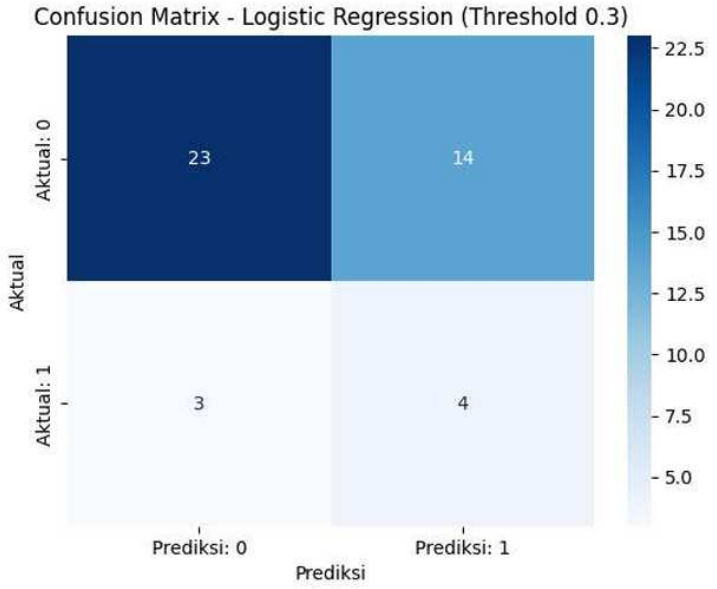
Analisis lebih lanjut pada masing-masing kelas menunjukkan bahwa untuk kelas 0 (kasus DBD tidak berisiko), model mampu mencapai precision sebesar 0,86 dan recall sebesar 0,68, menghasilkan nilai F1-score sebesar 0,76. Hal ini menunjukkan bahwa model cukup andal dalam mengenali dan mengklasifikasikan kasus-kasus non-berisiko. Sebaliknya, pada kelas 1 (kasus DBD berisiko), performa model menunjukkan kelemahan yang signifikan. Precision yang diperoleh hanya sebesar 0,20, dengan recall sebesar 0,43 dan F1-score sebesar 0,27. Nilai-nilai tersebut mengindikasikan bahwa model memiliki tingkat kesalahan prediksi yang tinggi dalam mengidentifikasi kasus berisiko, baik berupa prediksi positif palsu (false positives) maupun negatif palsu (false negatives).

Nilai macro average untuk precision, recall, dan F1-score masing-masing sebesar 0,53, 0,55, dan 0,52. Nilai ini mencerminkan rata-rata performa model tanpa mempertimbangkan distribusi kelas, dan menegaskan ketidakseimbangan performa antara kedua kelas. Sementara itu, nilai weighted average untuk metrik-metrik tersebut, yang memperhitungkan proporsi jumlah sampel di masing-masing kelas, menunjukkan precision sebesar 0,76, recall sebesar 0,64, dan F1-score sebesar 0,68. Meskipun lebih tinggi, nilai ini cenderung dipengaruhi oleh performa pada kelas mayoritas dan tidak

mencerminkan kemampuan model dalam mengklasifikasikan kasus berisiko secara spesifik.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa model memiliki kinerja yang cukup baik dalam mengidentifikasi kasus DBD tidak berisiko, namun belum mampu mendeteksi kasus berisiko secara optimal. Hal ini menjadi perhatian penting, mengingat tujuan utama dari model prediktif ini adalah untuk memberikan peringatan dini terhadap potensi lonjakan kasus DBD. Oleh karena itu, diperlukan pendekatan lanjutan untuk meningkatkan sensitivitas terhadap kelas minoritas, seperti penyesuaian ambang batas klasifikasi, penerapan teknik penyeimbangan data tambahan, atau eksplorasi algoritma klasifikasi alternatif yang lebih adaptif terhadap ketidakseimbangan kelas.

Hasil confusion matrix menunjukkan bahwa dari total 44 sampel, model menghasilkan:



Gambar 4. Hasil akurasi

- True Negative (TN) sebanyak 23: Kasus tidak berisiko yang berhasil diprediksi dengan benar.
- False Positive (FP) sebanyak 14: Kasus tidak berisiko yang salah diprediksi sebagai berisiko.
- False Negative (FN) sebanyak 3: Kasus berisiko yang salah diklasifikasikan sebagai tidak berisiko.
- True Positive (TP) sebanyak 4: Kasus berisiko yang berhasil dikenali dengan tepat oleh model.

Berdasarkan gambar tersebut, untuk memperoleh gambaran yang lebih komprehensif mengenai kinerja model klasifikasi, dilakukan analisis terhadap confusion matrix dengan ambang batas probabilitas (threshold) sebesar 0,3. Penurunan threshold

dari nilai default 0,5 ke 0,3 bertujuan untuk meningkatkan sensitivitas model terhadap kelas minoritas (kasus DBD berisiko).

Berdasarkan distribusi ini, dapat disimpulkan bahwa penurunan threshold berhasil meningkatkan jumlah True Positive dari sebelumnya, yang berdampak positif terhadap recall kelas 1. Recall meningkat menjadi 0,57 (4 dari total 7 kasus aktual berisiko), dibandingkan hanya 0,43 pada threshold 0,5. Peningkatan recall ini penting dalam konteks deteksi dini penyakit, karena kegagalan dalam mengidentifikasi kasus berisiko dapat berdampak langsung terhadap upaya pencegahan dan pengendalian penyakit di lapangan.

Namun demikian, peningkatan sensitivitas terhadap kelas berisiko juga disertai dengan peningkatan jumlah False Positive (14 kasus), yang menyebabkan penurunan precision. Hal ini menandakan adanya kompromi antara kemampuan model mendeteksi lebih banyak kasus berisiko dan potensi kesalahan dalam memberikan peringatan palsu. Dalam konteks kesehatan masyarakat, kompromi ini sering kali dianggap wajar dan dapat diterima, selama recall yang tinggi dapat dicapai untuk meminimalisir risiko luputnya kasus aktual.

Dengan demikian, penyesuaian threshold memberikan kontribusi positif dalam meningkatkan kemampuan model untuk mendeteksi kasus DBD berisiko. Namun, peningkatan ini perlu disertai dengan pertimbangan praktis terkait implementasi sistem peringatan dini, seperti kesiapan sumber daya untuk menindaklanjuti prediksi positif palsu.

SIMPULAN

Penelitian ini berhasil merancang model prediksi kejadian Demam Berdarah Dengue (DBD) di wilayah Purwokerto menggunakan algoritma Regresi Logistik Biner yang dioptimalkan melalui penerapan teknik Synthetic Minority Over-sampling Technique (SMOTE) untuk penanganan ketidakseimbangan data, serta tuning hyperparameter guna memperoleh konfigurasi model terbaik. Model dikembangkan berdasarkan variabel lingkungan seperti suhu udara, kelembaban, curah hujan, dan faktor musiman, yang semuanya diketahui berkontribusi terhadap dinamika penyebaran nyamuk Aedes aegypti. Hasil tuning menunjukkan bahwa konfigurasi optimal diperoleh pada regularisasi L2 dengan nilai parameter C sebesar 1, yang menghasilkan akurasi validasi silang sebesar 84,12%. Namun demikian, akurasi model pada data pengujian turun menjadi 64%, menunjukkan adanya perbedaan performa antara proses pelatihan dan generalisasi terhadap data baru. Ketidakseimbangan performa juga teridentifikasi dari rendahnya F1-score pada kelas minoritas (risiko tinggi), meskipun precision dan recall untuk kelas mayoritas tetap tinggi. Upaya penyesuaian threshold prediksi dari 0,5 ke 0,3 terbukti meningkatkan recall pada kelas minoritas dari 0,43 menjadi 0,57, walaupun disertai peningkatan prediksi positif palsu. Temuan ini menunjukkan bahwa kombinasi pendekatan Regresi Logistik, SMOTE, dan tuning hyperparameter dapat menjadi strategi

yang efektif untuk membangun sistem peringatan dini berbasis data yang lebih sensitif terhadap potensi kasus DBD, meskipun penguatan lanjutan tetap diperlukan untuk meningkatkan akurasi klasifikasi pada kasus berisiko secara lebih presisi.

DAFTAR PUSTAKA

Badan Pusat Statistik Kabupaten Banyumas. (2023). Kabupaten Banyumas dalam angka 2023. BPS Kabupaten Banyumas. https://banyumaskab.bps.go.id/publication

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Kementerian Kesehatan Republik Indonesia. (2022). Profil kesehatan Indonesia tahun 2021. Pusat Data dan Informasi Kesehatan, Kemenkes RI. https://pusdatin.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Profil-Kesehatan-Indonesia-2021.pdf

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305. http://jmlr.org/papers/v13/bergstra12a.html

Zhang, C., & Ma, Y. (2012). Ensemble machine learning: Methods and applications. Springer. https://doi.org/10.1007/978-1-4419-9326-7

Barros, R. C., Basgalupp, M. P., de Carvalho, A. C. P. L. F., & Freitas, A. A. (2014). A survey of evolutionary algorithms for decision-tree induction. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(3), 445–463. https://doi.org/10.1109/TSMC.2013.226

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer. https://doi.org/10.1007/978-3-319-98074-4

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), Multiple classifier systems (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7