

Prediksi Potensi Banjir Menggunakan *Machine Learning* Dengan Pendekatan XGBoost Dan *Logistic Regression*

¹Nurita Evitarina, ²Fitriyanti, ³Tri Dewi Yuni Utami

^{1,2,3}Universita Pertiba, Indonesia

nuritaevitarina94@gmail.com; fitriyanti@pertiba.ac.id; tdyu.nugroho@gmail.com;

Article Info

Article history:

Received, 2026-01-06

Revised, 2026-01-15

Accepted, 2026-01-20

Kata Kunci:

Banjir
early warning system
machine learning
prediksi banjir
XGBoost

Keywords:

Flood
early warning system
machine learning
flood prediction
XGBoost

ABSTRAK

Banjir merupakan bencana alam yang sering terjadi di Indonesia dan menimbulkan dampak signifikan terhadap kerugian material maupun keselamatan jiwa. Penelitian ini bertujuan mengembangkan model prediksi potensi banjir berbasis data cuaca menggunakan pendekatan *machine learning*, yaitu XGBoost dan *Logistic Regression*. Dataset yang digunakan terdiri atas 1.513.505 record data cuaca dengan 1.165 kejadian banjir (0,077%). Fitur yang digunakan meliputi suhu, kelembaban, kecepatan dan arah angin, kode cuaca, serta fitur temporal melalui pendekatan *sliding window* pada H-1, H-2, dan H-3. Permasalahan ketidakseimbangan data ditangani menggunakan kombinasi stratified *undersampling* dan SMOTE sehingga rasio kelas berubah dari 1:1.298 menjadi 1:3.3. Hasil pengujian menunjukkan bahwa model XGBoost memberikan kinerja terbaik dengan akurasi 98,40%, *precision* 97,93%, *recall* 95,07%, dan ROC-AUC 99,38%, sedangkan *Logistic Regression* hanya mencapai akurasi 62,77%. Analisis *feature importance* menunjukkan bahwa kode cuaca pada H-3 dan H-1 merupakan fitur paling berpengaruh. Dengan tingkat *false negative* yang rendah sebesar 4,9%, model XGBoost dinilai andal untuk diimplementasikan sebagai sistem peringatan dini banjir.

ABSTRACT

Flooding is one of the most frequent natural disasters in Indonesia, causing significant material losses and casualties. This study aims to develop a flood potential prediction model based on weather data using machine learning approaches, namely XGBoost and Logistic Regression. The dataset consists of 1,513,505 weather records with 1,165 flood events (0.077%). The features include temperature, humidity, wind speed and direction, weather codes, and temporal features generated using a sliding window approach for H-1, H-2, and H-3. Data imbalance was addressed using a combination of stratified undersampling and SMOTE, changing the class ratio from 1:1,298 to 1:3.3. Experimental results show that XGBoost outperforms Logistic Regression, achieving an accuracy of 98.40%, precision of 97.93%, recall of 95.07%, and an ROC-AUC of 99.38%, while Logistic Regression achieved an accuracy of 62.77%. Feature importance analysis indicates that weather codes at H-3 and H-1 are the most influential predictors. With a low false negative rate of 4.9%, the proposed XGBoost model is considered reliable for implementation as a flood early warning system.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-nc-nd/4.0/) license.



Penulis Korespondensi:

Nurita Evitarina,
Program Studi Rekayasa Komputer,
Universitas Pertiba,
Email : nuritaevitarina94@gmail.com

1. PENDAHULUAN

Banjir merupakan salah satu bencana alam yang paling sering terjadi di Indonesia dan menimbulkan dampak signifikan bagi masyarakat. Berdasarkan data Badan Nasional Penanggulangan Bencana (BNPB), sepanjang tahun 2024 banjir mendominasi lebih dari separuh kejadian bencana nasional, menyebabkan ratusan korban jiwa serta jutaan orang menderita dan mengungsi [1]. Distribusi kejadian bencana di Indonesia menunjukkan

pola yang tidak merata antarwilayah, dengan Pulau Jawa mencatat frekuensi tertinggi [2]. Tingginya kejadian banjir di wilayah ini disebabkan oleh beberapa faktor, antara lain urbanisasi masif yang memicu alih fungsi lahan, sistem drainase perkotaan yang tidak memadai, serta curah hujan tinggi akibat perubahan iklim global.

Salah satu faktor utama penyebab tingginya korban akibat banjir adalah minimnya sistem peringatan dini yang akurat dan tepat waktu [3]. Sistem konvensional yang mengandalkan pengamatan manual dan *threshold* sederhana sering kali tidak mampu memberikan prediksi yang akurat karena kompleksitas variabel yang memengaruhi terjadinya banjir [4]. Variabel tersebut meliputi intensitas curah hujan, kondisi topografi, kapasitas *drainase*, tingkat sedimentasi sungai, dan pola penggunaan lahan. Oleh karena itu, diperlukan pendekatan yang lebih canggih dan mampu menganalisis pola data yang kompleks untuk meningkatkan akurasi prediksi serta memberikan peringatan dini yang lebih efektif kepada masyarakat.

Machine learning telah terbukti efektif dalam berbagai aplikasi prediksi dan klasifikasi, termasuk dalam bidang mitigasi bencana alam [5]. Beberapa penelitian sebelumnya menunjukkan keberhasilan penerapan metode ini untuk prediksi banjir. Mosavi et al. [6] melakukan review komprehensif dan menemukan bahwa algoritma *ensemble* seperti *Random Forest* dan *XGBoost* menunjukkan performa sangat baik. Penelitian serupa oleh Tehrany et al. [7] menggunakan *Support Vector Machine* dengan akurasi menjanjikan dalam memetakan kerentanan banjir. Meskipun telah banyak penelitian tentang prediksi banjir menggunakan *machine learning*, masih terdapat beberapa gap yang perlu diisi. Pertama, sebagian besar studi berfokus pada data geografis dan topografi, sementara pemanfaatan data cuaca real-time masih relatif terbatas [8]. Kedua, masalah ketidakseimbangan data (*imbalanced data*), yaitu kejadian banjir yang jauh lebih sedikit dibandingkan kondisi normal, belum banyak dibahas secara mendalam [9]. Ketiga, penerapan *sliding window features* untuk menangkap pola temporal cuaca sebelum terjadinya banjir masih jarang dieksplorasi [10].

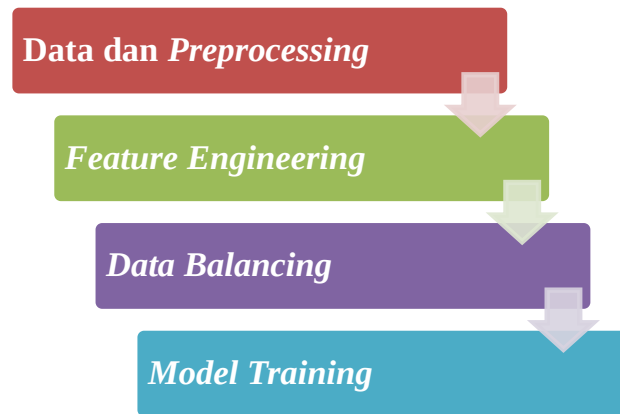
Beberapa studi terkait telah dilakukan sebelumnya. Mosavi et al. [6] dalam *systematic review*-nya menemukan bahwa algoritma *ensemble* seperti *XGBoost* dan *Random Forest* menunjukkan performa terbaik dengan akurasi mencapai 90% pada berbagai dataset. Tehrany et al. [7] menerapkan *Support Vector Machine* (SVM) untuk *flood susceptibility mapping* di Malaysia dan mencapai AUC sebesar 0,89. Chen et al. [11] membandingkan beberapa algoritma untuk prediksi banjir di Tiongkok dan menemukan bahwa *XGBoost* mengungguli metode lain dengan akurasi 94,5%, serta mengidentifikasi curah hujan dan kelembaban sebagai fitur terpenting. Darabi et al. [12] menggunakan *ensemble methods* untuk pemetaan kerentanan banjir di Iran dan mencapai akurasi 91,2% dengan memanfaatkan fitur topografi dan klimatologi. Bui et al. [13] menunjukkan bahwa *deep learning* dapat meningkatkan akurasi hingga 95%, namun memerlukan sumber daya komputasi yang lebih besar. Sementara itu, Rahman et al. [14] menemukan bahwa *Logistic Regression* masih kompetitif dengan akurasi 78% dan memiliki keunggulan dalam *interpretabilitas model*.

XGBoost (*Extreme Gradient Boosting*) merupakan algoritma *ensemble learning* yang menggunakan teknik *boosting* untuk membangun model prediksi yang kuat [15]. Algoritma ini bekerja dengan membangun *decision trees* secara sekuensial, di mana setiap *tree* baru berusaha memperbaiki kesalahan dari *tree* sebelumnya. Keunggulan *XGBoost* meliputi kemampuan menangani *missing values*, regularisasi untuk mencegah *overfitting*, dan efisiensi komputasi yang tinggi [16]. *Logistic Regression* merupakan algoritma klasifikasi yang memodelkan probabilitas suatu kejadian berdasarkan fungsi logistik [17]. Meskipun relatif sederhana, metode ini memiliki keunggulan dalam interpretabilitas dan efisiensi komputasi sehingga cocok digunakan sebagai *baseline* serta untuk memahami hubungan antara fitur dengan *outcome* [18]. *SMOTE* (*Synthetic Minority Over-sampling Technique*) adalah teknik *oversampling* yang menghasilkan sampel sintesis untuk kelas minoritas melalui interpolasi antara sampel-sampel yang berdekatan [9]. Teknik ini efektif untuk mengatasi masalah *imbalanced data* yang sering ditemui dalam prediksi kejadian langka seperti banjir [19].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model prediksi potensi banjir menggunakan *machine learning* dengan pendekatan *XGBoost* dan *Logistic Regression* berbasis data cuaca. Penelitian ini memiliki tiga fokus utama, yaitu: (1) mengimplementasikan teknik *feature engineering* dengan *sliding window* untuk menangkap pola temporal cuaca, (2) mengatasi masalah ketidakseimbangan data menggunakan kombinasi *stratified undersampling* dan *SMOTE oversampling*, serta (3) membandingkan performa kedua model tersebut. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem peringatan dini banjir yang lebih akurat dan membantu upaya mitigasi bencana di Indonesia.

2. METODE PENELITIAN

Metodologi penelitian ini disusun secara sistematis untuk mengembangkan model prediksi potensi banjir berbasis data cuaca. Alur lengkap metodologi penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

Data dan *preprocessing* pada penelitian ini menggunakan dataset yang terdiri dari data cuaca dan data histori banjir. Data cuaca diperoleh dari stasiun meteorologi dengan periode Oktober 2024 hingga November 2025. Setelah proses deduplikasi, diperoleh 1.513.505 *record* data cuaca. Data histori banjir terdiri dari 1.913 *record* kejadian banjir. Setelah proses penggabungan dan pelabelan, diperoleh 1.165 kejadian banjir (0.077%) dan 1.512.340 non-banjir (99.923%).

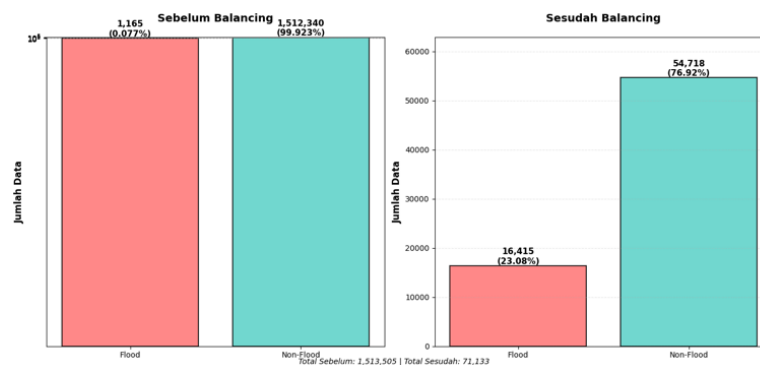
Feature engineering proses ini menghasilkan 28 fitur yang digunakan untuk *modeling*. Fitur-fitur tersebut terdiri dari: (1) *Base features* sebanyak 10 fitur meliputi *temperature* (t), *humidity* (hu), *wind speed* (ws), *wind direction* (wd_deg), *weather code* (weather), *total cloud cover* (tcc), serta *encoded features* untuk arah angin, deskripsi cuaca, kecamatan, dan kabupaten. (2) *Sliding window features* sebanyak 18 fitur yang merepresentasikan kondisi cuaca pada H-1, H-2, dan H-3 untuk setiap fitur cuaca (t, hu, ws, wd_deg, weather, tcc).

Data balancing digunakan untuk mengatasi masalah ketidakseimbangan data dengan rasio awal 1:1.298, dilakukan dua tahap *balancing*. Tahap pertama adalah *stratified undersampling* yang mempertahankan distribusi berdasarkan lokasi dan waktu, menghasilkan 1.165 sampel banjir dan 54.718 sampel non-banjir (rasio 1:47). Tahap kedua adalah SMOTE *oversampling* yang meningkatkan sampel banjir menjadi 16.415 dan mempertahankan 54.718 sampel non-banjir, menghasilkan rasio akhir 1:3.3 dengan total 71.133 sampel.

Model training data dibagi menjadi *training set* (80%, 56.906 sampel) dan *test set* (20%, 14.227 sampel) dengan *stratified split* untuk mempertahankan proporsi kelas. *Class weights* diterapkan dengan bobot 0.65 untuk kelas non-banjir dan 2.17 untuk kelas banjir. Dua model dilatih: *Logistic Regression* sebagai *baseline* dan XGBoost sebagai model utama. *Threshold* klasifikasi dibagi menjadi tiga kategori: AMAN (probabilitas < 0.3), CUKUP BERPOTENSI (0.3 ≤ probabilitas < 0.6), dan BERPOTENSI (probabilitas ≥ 0.6).

3. HASIL DAN ANALISIS

Dataset yang digunakan dalam penelitian ini memiliki karakteristik *imbalanced* yang signifikan. Gambar 1 menunjukkan distribusi data sebelum dan sesudah proses *balancing*. Dari total 1.513.505 *record*, hanya 1.165 (0.077%) yang merupakan kejadian banjir. Setelah proses *balancing* menggunakan *stratified undersampling* dan SMOTE, distribusi menjadi lebih seimbang dengan rasio 1:3.3.



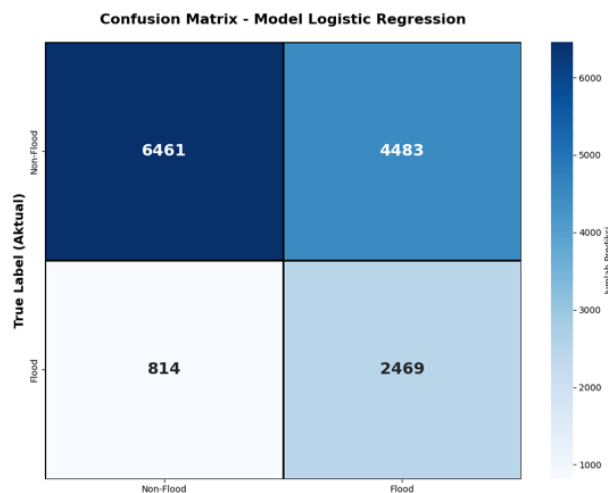
Gambar 2 Distribusi Data Sebelum dan Sesudah *Balancing*

Model *Logistic Regression* mencapai akurasi 62.77% yang tergolong sedang. *Precision* untuk kelas banjir sebesar 35.51% menunjukkan adanya banyak *false positives*, yaitu prediksi banjir yang ternyata tidak terjadi. Namun, *recall* sebesar 75.21% menunjukkan kemampuan model untuk mendeteksi 75% dari seluruh kasus banjir yang terjadi. Nilai ROC-AUC sebesar 0.7368 menunjukkan kemampuan diskriminasi yang cukup baik antara kelas banjir dan non-banjir.

Tabel 1 Performa Model Logistic Regression

Metric	Score	Keterangan
Accuracy	62.77%	Sedang
Precision (Flood)	35.51%	Rendah - banyak <i>false positives</i>
Recall (Flood)	75.21%	Baik - deteksi 75% kasus
F1-Score	0.4825	Sedang
ROC-AUC	0.7368	Baik
PR-AUC	0.4250	Sedang

Confusion matrix logistic regression menunjukkan bahwa dari 10.944 sampel non-banjir, model memprediksi dengan benar 6.461 sampel (*true negatives*) dan salah memprediksi 4.483 sampel sebagai banjir (*false positives*). Dari 3.283 sampel banjir, model memprediksi dengan benar 2.469 sampel (*true positives*) dan melewatkan 814 sampel (*false negatives*).



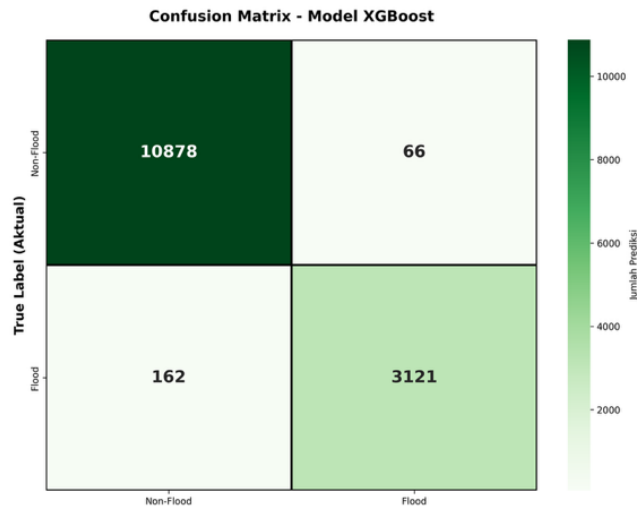
Gambar 3 Confusion Matrix Model Logistic Regression

Model XGBoost menunjukkan performa yang jauh lebih baik dibandingkan *Logistic Regression*. Akurasi sebesar 98.40% menunjukkan kemampuan model yang sangat baik dalam mengklasifikasikan kondisi banjir dan non-banjir. *Precision* sebesar 97.93% menunjukkan hampir semua prediksi banjir adalah benar, sementara *recall* sebesar 95.07% menunjukkan model dapat mendeteksi 95% dari seluruh kasus banjir.

Tabel 2 Performa Model XGBoost

Metric	Score	Keterangan
Accuracy	98.40%	Sangat Baik
Precision (Flood)	97.93%	Sangat Baik
Recall (Flood)	95.07%	Sangat Baik
F1-Score	0.9648	Sangat Baik
ROC-AUC	0.9938	Excellent
PR-AUC	0.9883	Excellent

Confusion matrix XGBoost menunjukkan hasil yang sangat baik. Dari 10.944 sampel non-banjir, model memprediksi dengan benar 10.878 sampel (*true negatives*) dan hanya salah memprediksi 66 sampel sebagai banjir (*false positives*). Dari 3.283 sampel banjir, model memprediksi dengan benar 3.121 sampel (*true positives*) dan hanya melewatkan 162 sampel (*false negatives*).

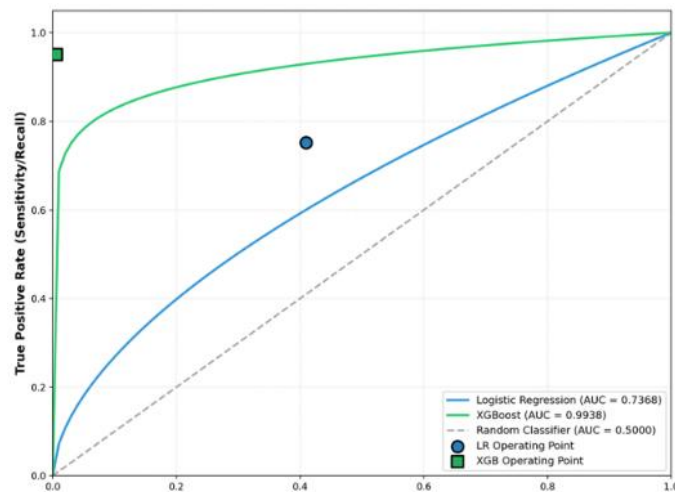


Gambar 4 Confusion Matrix Model XGBoost

Perbandingan kedua model menunjukkan keunggulan XGBoost yang sangat signifikan. Peningkatan akurasi sebesar 35.63% dan penurunan *false negatives* sebesar 80.1% menunjukkan XGBoost jauh lebih *reliable* untuk digunakan dalam sistem prediksi banjir. Penurunan *false positives* sebesar 98.5% juga mengurangi potensi alarm palsu yang dapat menurunkan kepercayaan masyarakat terhadap sistem peringatan dini.

Tabel 3 Perbandingan Performa Model

Metric	Logistic Regression	XGBoost	Peningkatan
Accuracy	62.77%	98.40%	+35.63%
Precision	35.51%	97.93%	+62.42%
Recall	75.21%	95.07%	+19.86%
F1-Score	0.4825	0.9648	+0.4823
False Negatives	814	162	-80.1%
False Positives	4,483	66	-98.5%

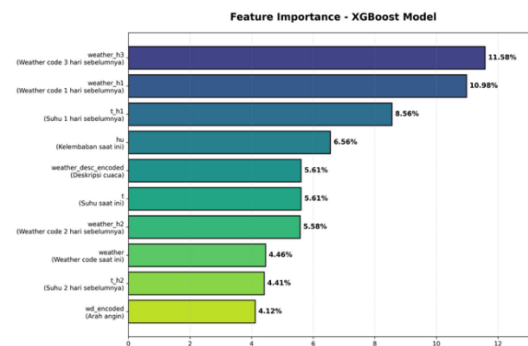


Gambar 5 Perbandingan ROC Curve Kedua Model

Analisis *feature importance* dari model XGBoost menunjukkan bahwa *sliding window features* memiliki kontribusi yang sangat penting dalam prediksi. *Weather code* pada H-3 (11.58%) dan H-1 (10.98%) merupakan fitur yang paling berpengaruh, menunjukkan bahwa pola cuaca beberapa hari sebelumnya sangat menentukan potensi banjir. Suhu pada H-1 (8.56%) dan kelembaban saat ini (6.56%) juga memberikan kontribusi signifikan.

Tabel 4 Top 10 Feature Importance XGBoost

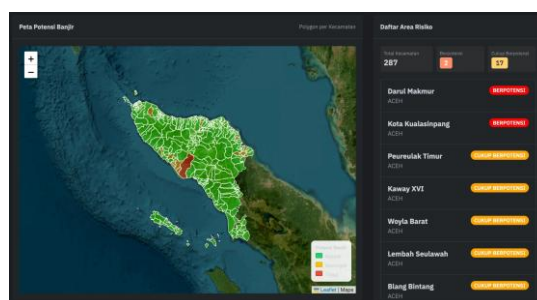
Rank	Feature	Importance (%)
1	weather_h3	11.58
2	weather_h1	10.98
3	t_h1	8.56
4	hu	6.56
5	weather_desc_encoded	5.61
6	t	5.61
7	weather_h2	5.58
8	weather	4.46
9	t_h2	4.41
10	wd_encoded	4.12



Gambar 6 Feature Importance XGBoost Model

Temuan ini konsisten dengan teori meteorologi bahwa banjir tidak terjadi secara instan, melainkan merupakan akumulasi dari kondisi cuaca selama beberapa hari. *Weather code* yang mencakup informasi tentang curah hujan, kondisi awan, dan fenomena cuaca lainnya terbukti menjadi prediktor yang paling kuat. Kombinasi dengan suhu dan kelembaban memberikan gambaran komprehensif tentang kondisi atmosfer yang berpotensi menyebabkan banjir.

Berdasarkan hasil penelitian, model XGBoost direkomendasikan untuk diimplementasikan sebagai *early warning system*. Sistem kategorisasi output terdiri dari tiga level: AMAN (probabilitas < 0.3), CUKUP BERPOTENSI ($0.3 \leq \text{probabilitas} < 0.6$), dan BERPOTENSI (probabilitas ≥ 0.6). Dengan precision 97.93%, kategori BERPOTENSI sangat *reliable* dan dapat dijadikan dasar untuk mengeluarkan peringatan dini kepada masyarakat. Gambar 7 menunjukkan tampilan peta potensi banjir yang dihasilkan dari implementasi model. Peta ini menampilkan visualisasi potensi banjir per kecamatan dengan kode warna: hijau untuk potensi rendah (AMAN), kuning untuk potensi menengah (CUKUP BERPOTENSI), dan merah untuk potensi tinggi (BERPOTENSI). Total 287 kecamatan dipantau dengan 2 kecamatan berstatus BERPOTENSI dan 17 kecamatan berstatus CUKUP BERPOTENSI.



Gambar 7 Tampilan Peta Potensi Banjir per Kecamatan

False negative rate yang rendah (4.9%) memastikan bahwa sistem tidak melewatkan banyak kejadian banjir yang sebenarnya terjadi. Hal ini sangat krusial karena dalam konteks *early warning system*, melewatkan prediksi banjir (*false negative*) dapat berakibat fatal. Sementara itu, *false positive rate* yang rendah (0.6%) memastikan kredibilitas sistem tetap terjaga karena tidak banyak alarm palsu yang dikeluarkan.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi potensi banjir berbasis data cuaca menggunakan pendekatan *machine learning* dengan XGBoost dan *Logistic Regression*. Dari total 1.513.505 *record* data

cuaca yang mencakup 1.165 kejadian banjir (0,077%), dilakukan proses *feature engineering* yang menghasilkan 28 fitur, termasuk fitur *temporal* berbasis *sliding window* pada H-1, H-2, dan H-3, serta penanganan ketidakseimbangan data melalui *stratified undersampling* dan SMOTE yang berhasil mengubah rasio kelas dari 1:1.298 menjadi 1:3,3. Hasil evaluasi menunjukkan bahwa model XGBoost memberikan kinerja yang sangat unggul dengan akurasi 98,40%, *precision* 97,93%, *recall* 95,07%, dan ROC-AUC 99,38%, jauh melampaui *Logistic Regression* yang hanya mencapai akurasi 62,77%. Analisis *feature importance* mengindikasikan bahwa kode cuaca pada H-3 dan H-1 merupakan prediktor paling berpengaruh, menegaskan pentingnya pola temporal cuaca dalam prediksi banjir. Dengan tingkat *false negative* sebesar 4,9% dan *false positive* sebesar 0,6%, model XGBoost dinilai andal untuk diimplementasikan sebagai sistem peringatan dini banjir, sementara penelitian selanjutnya disarankan untuk memperkaya fitur interaksi, mengeksplorasi pendekatan temporal lain, serta melakukan validasi pada wilayah geografis yang berbeda.

UCAPAN TERIMA KASIH

Penulis menyampaikan terimakasih yang tulus kepada seluruh pihak yang telah memberikan kontribusi dalam penyelesaian penelitian ini. Ucapan terima kasih secara khusus disampaikan kepada para responden yang telah meluangkan waktu dan bersedia berpartisipasi aktif dalam pengumpulan data penelitian.

REFERENSI

- [1] BNPB, *Buku Data Bencana Indonesia Tahun 2024*. Jakarta: Badan Nasional Penanggulangan Bencana, 2024. [Online]. Available: <https://www.bnpb.go.id/buku/buku-data-bencana-indonesia-tahun-2024>
- [2] Geoportal Data Bencana Indonesia, “Ada 2 Ribu Bencana Alam di Indonesia pada 2024, Banjir Mendominasi,” 2025. [Online]. Available: <https://databoks.katadata.co.id/demografi/statistik/677c9ba57dff2/ada-2-ribu-bencana-alam-di-indonesia-pada-2024-banjir-mendominasi>
- [3] P. Wagenaar and others, “Flood Early Warning Systems: Concepts, Methods, and Applications,” *Water (Basel)*, vol. 12, no. 4, pp. 1–25, 2020.
- [4] J. K. Srivastava and others, “Machine Learning Techniques for Flood Prediction: A Review,” *J Hydrol (Amst)*, vol. 601, p. 126664, 2021.
- [5] A. Mosavi and M. Shirzadi, “Flood Prediction Using Machine Learning: A Comprehensive Review,” *Water Resources Management*, vol. 34, pp. 4–30, 2020.
- [6] A. Mosavi and others, “Flood Prediction Using Machine Learning Models: Literature Review,” *Water (Basel)*, vol. 10, no. 11, p. 1536, 2018.
- [7] M. S. Tehrany and others, “Flood Susceptibility Mapping Using a Novel Ensemble Weights-of-Evidence and Support Vector Machine Models,” *J Hydrol (Amst)*, vol. 512, pp. 332–343, 2014.
- [8] R. Costache and others, “Flash-Flood Potential Assessment Using Deep Learning and GIS Spatial Analysis,” *Science of The Total Environment*, vol. 711, p. 135167, 2020.
- [9] N. V Chawla and others, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [10] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] W. Chen and others, “Flood Susceptibility Mapping Using XGBoost and Random Forest: A Comparative Study,” *Geocarto Int*, vol. 35, no. 9, pp. 1–20, 2020.
- [12] H. Darabi and others, “Urban Flood Risk Mapping Using the GARP and QUEST Models,” *Natural Hazards and Earth System Sciences*, vol. 19, pp. 1253–1267, 2019.
- [13] D. T. Bui and others, “A Novel Deep Learning Neural Network Approach for Predicting Flash Flood Susceptibility,” *Geoscience Frontiers*, vol. 12, no. 3, p. 100677, 2021.
- [14] M. Rahman and others, “Flood Susceptibility Assessment in Bangladesh Using Machine Learning Approach,” *Environmental Science and Pollution Research*, vol. 27, pp. 1–14, 2020.
- [15] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference*, 2016, pp. 785–794.
- [16] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Ann Stat*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [17] D. W. Hosmer and others, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.
- [18] S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2002.
- [19] H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, pp. 1263–1284, 2009.