Analisis Prediksi Risiko Banjir Menggunakan Algoritma Random Forest

Tri Prasetyo*[1], Ardianto[2]
[1,2]Universitas Pamulang, Jl. Raya Puspitek, Tangerang Selatan 15310, Indonesia
e-mail: *dosen02669@unpam.ac.id

Abstrak

Banjir merupakan salah satu bencana alam yang sering terjadi dan menimbulkan dampak signifikan terhadap kehidupan masyarakat, sehingga diperlukan upaya prediksi yang akurat sebagai dasar mitigasi dan sistem peringatan dini. Penelitian ini bertujuan untuk menganalisis dan memprediksi risiko banjir menggunakan algoritma Random Forest berbasis data banjir per jam. Dataset yang digunakan terdiri dari 105.408 data observasi dengan sembilan variabel, yang mencakup faktor hidrologi, temporal, dan administratif, serta variabel target berupa status banjir. Metode penelitian meliputi tahapan pemahaman data, preprocessing, pembentukan model Random Forest, prediksi, dan evaluasi model menggunakan confusion matrix dan classification report. Hasil penelitian menunjukkan bahwa model Random Forest menghasilkan tingkat akurasi sebesar 98%, dengan performa yang sangat baik dalam mengklasifikasikan kondisi tidak banjir, serta cukup baik dalam mendeteksi kejadian banjir meskipun masih terdapat keterbatasan akibat ketidakseimbangan data. Analisis feature importance menunjukkan bahwa curah hujan dan debit air merupakan faktor paling dominan dalam menentukan risiko banjir. Hasil penelitian ini menegaskan bahwa pendekatan machine learning menggunakan Random Forest efektif untuk prediksi risiko banjir dan berpotensi mendukung pengembangan sistem peringatan dini banjir berbasis data.

Kata kunci *banjir, prediksi risiko banjir, Random Forest, machine learning, feature importance*

Abstract

Flooding is one of the most frequent natural disasters and causes significant impacts on society, making accurate prediction essential to support mitigation efforts and early warning systems. This study aims to analyze and predict flood risk using the Random Forest algorithm based on hourly flood data. The dataset consists of 105,408 observation records with nine variables, including hydrological, temporal, and administrative factors, as well as a target variable representing flood status. The research methodology includes data understanding, data preprocessing, Random Forest model development, prediction, and model evaluation using a confusion matrix and classification report. The results show that the Random Forest model achieves an accuracy of 98%, demonstrating excellent performance in classifying non-flood conditions and fairly good performance in detecting flood events, despite some limitations caused by data imbalance. Feature importance analysis indicates that rainfall and water discharge are the most dominant factors in determining flood risk. These findings confirm that a machine learning approach using the Random Forest algorithm is effective for flood risk prediction and has strong potential to support the development of data-driven flood early warning systems.

Keywords: *flood, flood risk prediction, Random Forest, machine learning, feature importance*

1. PENDAHULUAN

Banjir merupakan salah satu bencana alam yang sering terjadi dan menimbulkan dampak signifikan terhadap kehidupan masyarakat, terutama di wilayah yang dipengaruhi oleh curah hujan tinggi dan kondisi aliran sungai yang fluktuatif [1]. Dalam upaya mitigasi banjir, pemanfaatan data historis lingkungan menjadi sangat penting untuk memprediksi risiko banjir secara dini dan akurat. Dataset yang digunakan dalam penelitian ini terdiri dari 105.408 data per jam dengan delapan variabel prediktor, yaitu desa, tanggal, jam, curah hujan (mm), luas desa (ha), debit air (m^3/jam), aliran sungai, dan debit maksimum (m^3/jam), serta satu variabel target berupa status banjir (true/false). Kelengkapan data yang baik tanpa nilai hilang menjadikan dataset ini representatif untuk analisis prediksi risiko banjir berbasis data.[2]

Permasalahan yang diselesaikan dalam penelitian ini adalah bagaimana membangun model prediksi risiko banjir yang mampu menangkap pola kompleks dan tidak linear dari data lingkungan per jam. Analisis

awal menunjukkan adanya ketidakseimbangan kelas antara kondisi tidak banjir dan banjir, yang tercermin dari hasil confusion matrix dan classification report. Meskipun model Random Forest menghasilkan akurasi yang tinggi, tantangan utama terletak pada peningkatan kemampuan model dalam mendeteksi kejadian banjir secara tepat, mengingat kejadian banjir relatif lebih sedikit dibandingkan kondisi normal[3]. Oleh karena itu, diperlukan algoritma yang robust, stabil, dan mampu bekerja secara efektif pada data dengan karakteristik tersebut.

Isu lain yang terkait dengan permasalahan ini adalah kebutuhan akan interpretabilitas model. Dalam konteks mitigasi bencana, tidak cukup hanya mengetahui hasil prediksi, tetapi juga penting untuk memahami faktor-faktor yang paling berpengaruh terhadap risiko banjir. Hasil analisis *feature importance* pada penelitian ini menunjukkan bahwa curah hujan dan debit air merupakan variabel paling dominan dalam menentukan risiko banjir, diikuti oleh debit maksimum, waktu kejadian (jam), dan aliran sungai. Temuan ini menegaskan bahwa variabel hidrologi memiliki peranan penting dalam pembentukan risiko banjir dan perlu menjadi perhatian utama dalam sistem peringatan dini.[4]

Penelitian terdahulu telah banyak membahas penggunaan metode *machine learning* dalam prediksi banjir. Breiman (2001) memperkenalkan Random Forest sebagai metode *ensemble learning* yang mampu meningkatkan akurasi dan mengurangi risiko *overfitting*. Beberapa penelitian selanjutnya, seperti yang dilakukan oleh Tehrany et al. (2015), Khosravi et al. (2018), dan Mosavi et al. (2018), menunjukkan bahwa Random Forest memiliki performa yang unggul dibandingkan algoritma lain dalam prediksi banjir, khususnya pada data lingkungan yang kompleks[5]. Selain itu, penelitian oleh Ahmadlou et al. (2019) menegaskan bahwa penggunaan data dengan resolusi waktu tinggi, seperti data per jam, dapat meningkatkan ketepatan prediksi risiko banjir. Berdasarkan hal tersebut, penelitian ini menerapkan algoritma Random Forest pada data banjir per jam untuk memprediksi risiko banjir sekaligus menganalisis faktor-faktor utama yang memengaruhi terjadinya banjir.[6]

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *machine learning* untuk memprediksi risiko banjir berdasarkan data historis lingkungan per jam[7]. Dataset yang digunakan terdiri dari 105.408 data observasi dengan sembilan atribut, yaitu delapan variabel independen berupa desa, tanggal, jam, curah hujan (mm), luas desa (ha), debit air (m^3/jam), aliran sungai, dan debit maksimum (m^3/jam), serta satu variabel dependen berupa status banjir yang bersifat biner (banjir dan tidak banjir). Data yang digunakan tidak memiliki nilai kosong sehingga dapat langsung digunakan dalam proses pemodelan.

Tahapan penelitian diawali dengan eksplorasi dan pemahaman data untuk mengetahui karakteristik serta distribusi setiap variabel. Selanjutnya dilakukan proses pemisahan data menjadi variabel input dan variabel target, kemudian dataset dibagi menjadi data latih dan data uji dengan perbandingan 80% data latih dan 20% data uji. Pembagian data ini bertujuan untuk menguji kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya.

Model prediksi risiko banjir dibangun menggunakan algoritma Random Forest Classifier. Algoritma ini bekerja dengan membentuk sejumlah pohon keputusan secara acak dan menggabungkan hasil prediksi dari setiap pohon untuk menghasilkan keputusan akhir[8]. Proses pelatihan model dilakukan menggunakan data latih, sedangkan data uji digunakan untuk mengevaluasi performa model. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *f1-score*, serta confusion matrix untuk mengetahui kemampuan model dalam mengklasifikasikan kondisi banjir dan tidak banjir secara tepat, khususnya pada data dengan ketidakseimbangan kelas.

Selain evaluasi kinerja, penelitian ini juga melakukan analisis *feature importance* untuk mengetahui tingkat pengaruh masing-masing variabel terhadap prediksi risiko banjir. Analisis ini digunakan untuk mengidentifikasi faktor-faktor utama yang memengaruhi terjadinya banjir, sehingga hasil penelitian tidak hanya memberikan prediksi, tetapi juga informasi yang dapat dimanfaatkan dalam pengambilan keputusan dan pengembangan sistem peringatan dini banjir.[9]

Berikut Flowchart Metodologi Penelitian Analisis Prediksi Risiko Banjir Menggunakan Algoritma Random Forest

Gambar 1. Flowchart Metodologi Penelitian

Flowchart penelitian ini menunjukkan alur proses prediksi risiko banjir yang dimulai dari pengumpulan dan pemrosesan data, dilanjutkan dengan pembangunan dan pengujian model Random Forest, kemudian dievaluasi untuk menilai kinerja model, hingga menghasilkan kesimpulan sebagai hasil akhir penelitian.

3. HASIL DAN PEMBAHASAN

3.1 Import Library

Tahap awal dalam penelitian ini adalah melakukan *import library* yang diperlukan untuk mendukung seluruh proses analisis prediksi risiko banjir. Pada tahap ini, pustaka pandas dan numpy digunakan untuk membaca, mengelola, dan memanipulasi dataset banjir per jam. Selanjutnya, pustaka scikit-learn digunakan untuk membangun model *machine learning*, khususnya algoritma Random Forest, serta untuk melakukan evaluasi kinerja model. Selain itu, library visualisasi digunakan untuk membantu menampilkan hasil analisis secara grafis. Tahap *import library* ini menjadi fondasi penting dalam penelitian karena memastikan seluruh fungsi dan alat yang dibutuhkan telah tersedia sebelum dilakukan pengolahan data dan pemodelan lebih lanjut.

3.2. Load Dataset

Pada tahap load dataset, data banjir per jam berhasil dimuat ke dalam bentuk *DataFrame*. Dataset terdiri dari 105.408 baris data dan 9 kolom, yang mencakup informasi desa, waktu pengamatan, variabel hidrologi, serta status banjir sebagai variabel target. Data dapat dimuat tanpa kendala, menunjukkan bahwa format dan struktur dataset telah sesuai untuk proses analisis lebih lanjut

3.3 Data Understanding

Tahap data understanding dilakukan untuk memahami karakteristik dataset. Hasil analisis menunjukkan bahwa seluruh kolom memiliki jumlah data yang lengkap tanpa adanya nilai kosong (*missing value*). Tipe data pada masing-masing kolom juga telah sesuai, terdiri dari data numerik dan boolean. Kondisi ini menunjukkan bahwa dataset memiliki kualitas yang baik dan representatif untuk digunakan dalam pemodelan prediksi risiko banjir.

3.4 Data Preprocessing

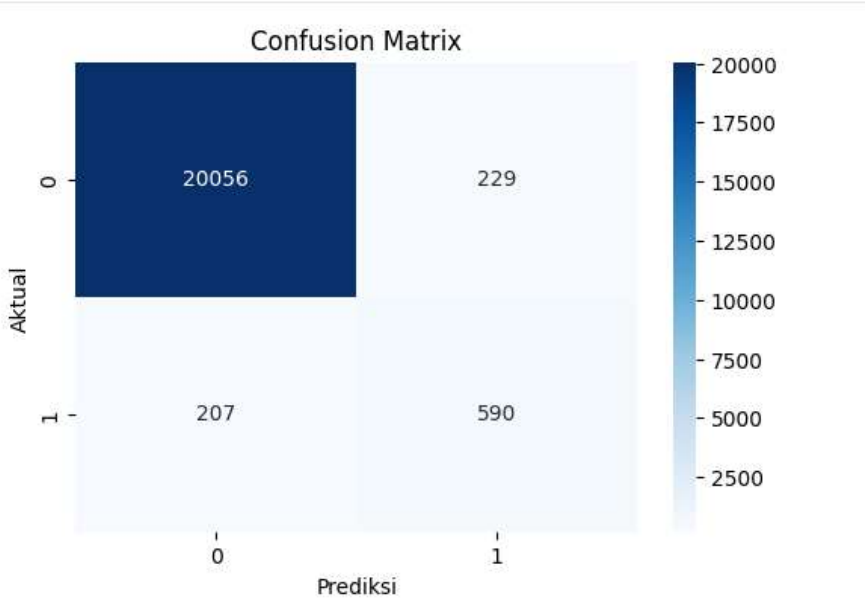
Pada tahap data preprocessing, dilakukan pemisahan antara variabel input dan variabel target. Variabel input terdiri dari delapan atribut lingkungan dan temporal, sedangkan variabel target adalah status banjir. Dataset kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80% data latih dan 20% data uji. Tahap ini bertujuan untuk menyiapkan data agar model dapat dilatih dan diuji secara objektif.

3.5 Pembuatan Model Random Forest

Tahap berikutnya adalah pembuatan model menggunakan RandomForestClassifier, sebagaimana ditunjukkan pada gambar model Random Forest. Model dilatih menggunakan data latih dengan parameter *random_state* untuk menjaga konsistensi hasil. Algoritma Random Forest bekerja dengan membangun banyak pohon keputusan dan menggabungkan hasilnya, sehingga mampu menangani data berukuran besar dan hubungan non-linear antar variabel.

3.6 Prediksi dan Evaluasi Model

Setelah model terbentuk, dilakukan prediksi pada data uji dan evaluasi kinerja model. Berdasarkan *classification report* pada gambar, model menghasilkan tingkat akurasi sebesar 98%. Kelas tidak banjir (False) memiliki nilai *precision*, *recall*, dan *f1-score* yang sangat tinggi, yaitu 0,99, yang menunjukkan bahwa model sangat andal dalam mengklasifikasikan kondisi normal. Sementara itu, kelas banjir (True) memiliki nilai *precision* sebesar 0,72, *recall* 0,74, dan *f1-score* 0,73, yang menunjukkan bahwa model cukup baik dalam mendeteksi kejadian banjir meskipun masih terdapat kesalahan prediksi.

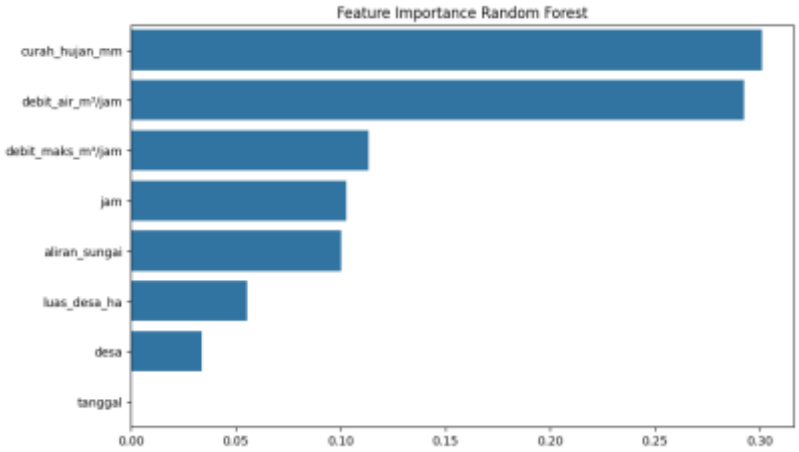


Gambar 2. Confusion Matrix

Gambar 2. Berdasarkan hasil confusion matrix, model berhasil mengklasifikasikan 20.056 data tidak banjir secara benar (true negative) dan 590 data banjir secara benar (true positive). Namun demikian, masih terdapat 207 data banjir yang tidak terdeteksi (false negative), yaitu kondisi banjir yang diprediksi sebagai tidak banjir, serta 229 data tidak banjir yang salah diprediksi sebagai banjir (false positive). Kesalahan *false negative* menjadi perhatian penting karena berpotensi menyebabkan kejadian banjir tidak terantisipasi, sedangkan *false positive* dapat menimbulkan peringatan yang tidak diperlukan.

Hasil tersebut menunjukkan bahwa model Random Forest bekerja sangat baik pada kelas mayoritas (tidak banjir), tetapi performanya pada kelas minoritas (banjir) masih perlu ditingkatkan. Kondisi ini disebabkan oleh ketidakseimbangan jumlah data, di mana data tidak banjir jauh lebih banyak dibandingkan data banjir. Meskipun demikian, nilai *recall* pada kelas banjir sebesar 0,74 menunjukkan bahwa sebagian besar kejadian banjir masih dapat terdeteksi oleh model. Oleh karena itu, secara keseluruhan model sudah memiliki performa yang baik untuk prediksi risiko banjir, namun pengembangan lebih lanjut seperti penanganan data tidak seimbang atau penyesuaian parameter model dapat dilakukan untuk meningkatkan sensitivitas terhadap kejadian banjir.

3.7 Feature Importance



Gambar 3. Feature Importance

Gambar 3. Berdasarkan grafik *feature importance*, variabel curah_hujan_mm (0,301) dan debit_air_m³/jam (0,293) merupakan faktor paling dominan dalam prediksi risiko banjir. Hal ini menunjukkan bahwa peningkatan curah hujan dan debit air memiliki pengaruh langsung terhadap terjadinya banjir. Variabel debit_maks_m³/jam, jam, dan aliran_sungai juga memberikan kontribusi yang cukup signifikan, yang mengindikasikan adanya pengaruh kapasitas aliran air dan pola waktu terhadap kejadian banjir. Sebaliknya, variabel luas_desa_ha, desa, dan tanggal memiliki pengaruh yang relatif kecil, dengan variabel tanggal hampir tidak berkontribusi terhadap prediksi. Hasil ini menegaskan bahwa faktor hidrologi merupakan penentu utama risiko banjir, sementara faktor administratif dan kalender memiliki pengaruh yang terbatas.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan algoritma Random Forest pada data banjir per jam yang terdiri dari 105.408 data observasi dengan sembilan variabel mampu menghasilkan model prediksi risiko banjir dengan performa yang sangat baik. Hasil evaluasi menunjukkan tingkat akurasi sebesar 98%, di mana model sangat andal dalam mengklasifikasikan kondisi tidak banjir, meskipun masih terdapat keterbatasan dalam mendeteksi kejadian banjir akibat ketidakseimbangan jumlah data. Berdasarkan hasil confusion matrix, sebagian besar data berhasil diklasifikasikan dengan benar, namun kesalahan prediksi pada kejadian banjir menjadi perhatian penting dalam konteks mitigasi bencana. Analisis feature importance menunjukkan bahwa curah hujan dan debit air merupakan faktor paling dominan dalam menentukan risiko banjir, diikuti oleh debit maksimum, waktu kejadian, dan aliran sungai, sedangkan faktor administratif dan kalender memiliki pengaruh yang relatif kecil. Secara keseluruhan, hasil penelitian ini membuktikan bahwa Random Forest efektif digunakan untuk memprediksi risiko banjir serta mengidentifikasi faktor utama penyebab banjir, sehingga berpotensi mendukung pengembangan sistem peringatan dini banjir berbasis data.

DAFTAR PUSTAKA

[1] N. Haque, T. Islam, and M. Erfan, “An exploration of machine learning approaches for early Autism Spectrum Disorder detection,” *Healthc. Anal.*, vol. 7, no. January, p. 100379, 2025.

[2] L. S. Qamarani and M. Riasetiawan, “Klasifikasi Level Banjir Menggunakan Random Forest dan Support Vector Machine,” *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 14, no. 2, p. 199, 2024.

[3] D. F. Zahra and C. -, “Studi Literatur Pemanfaatan Artificial Intelligence untuk Prediksi Bencana Banjir,” *J. Teknol. Inf. dan Komun.*, vol. 18, no. 1, pp. 15-26, 2025.

[4] I. Hapsari and S. Pandya Wisesa, “Evaluasi Model Prediksi Curah Hujan Berbasis Machine Learning di Kota Bandung,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 11, no. 2, pp. 136-143, 2025.

[5] S. M. Natzir, “Prediksi Banjir menggunakan Naive Bayes Di Sleman,” *J. Teknol. Inf.*, vol. 14, no. 2, pp. 59-64, 2023.

[6] I. L. Rahmah, A. Nugroho, Y. Manahan, and T. Siregar, “Prediksi Curah Hujan Menggunakan Random Forest dan VAR di Kediri Raya,” *J. Teknol. Dan Sist. Inf. Bisnis-JTEKSIS*, vol. 7, no. 4, p. 539, 2025.

[7] S. E. Purwati and Y. Pristyanto, “Model Random Forest and Support Vector Machine for Flood Classification in Indonesia,” *Sinkron*, vol. 8, no. 4, pp. 2261-2268, 2024.

[8] Jupron and Sutrisno, "Analysis of Heart Disease Using the Random Forest Method," *J. Inotera*, vol. 10, no. 1, pp. 167-174, 2025.

[9] A. Suaif and E. Sylvianti Rahayu, "Analisis Faktor Dan Pola Kejadian Banjir Di Bandar Lampung Menggunakan Arima, Random Forest, Dan Xgboost," *J. Teknol. Komput. dan Inform.*, vol. 3, no. 2, 2025.