

BAB I

PENDAHULUAN

1.1 Latar Belakang

Curah hujan memiliki peran penting dalam berbagai aspek kehidupan terutama di negara beriklim tropis seperti Indonesia. Indonesia memiliki pola curah hujan tinggi dengan variasi musiman yang signifikan. Daerah pegunungan atau berbukit dengan variasi perbedaan topografi akan mempengaruhi besarnya curah hujan yang berbeda pada masing-masing perbedaan elevasi tersebut (Yuono *et al.*, 2014).

Variabilitas curah hujan memberikan dampak signifikan terhadap berbagai sektor termasuk sektor pertanian, ketersediaan air, dan bencana hidrometeorologi seperti banjir dan tanah longsor. Beberapa wilayah mengalami peningkatan curah hujan ekstrem dengan intensitas lebih dari 300 mm per bulan pada musim hujan. Pada musim kemarau terdapat wilayah tertentu yang mengalami penurunan curah hujan yang signifikan hingga berpotensi menyebabkan kekeringan. Kondisi ini merupakan dampak dari perubahan iklim yang menyebabkan pergeseran pola curah hujan secara tidak menentu. Hal tersebut menjadi salah satu dampak signifikan dari perubahan iklim yang dihadapi Indonesia (Surmaini *et al.*, 2011).

Data historis curah hujan yang dikumpulkan secara berkala memungkinkan diterapkannya teknik *machine learning* untuk mengidentifikasi dan mengklasifikasikan pola curah hujan secara lebih akurat. Salah satu metode yang banyak digunakan dalam klasifikasi data cuaca adalah algoritma *random forest*. *Random forest* merupakan algoritma *ensemble learning* berbasis pohon keputusan yang dikembangkan dari metode *Classification and Regression Tree* (CART). Algoritma ini membangun sejumlah pohon keputusan dari sampel berbeda

menggunakan teknik *bootstrap aggregating (bagging)* serta hanya mempertimbangkan sebagian variabel pada setiap pembentukan pohon keputusan (*random feature selection*). Pendekatan ini membantu mengatasi permasalahan *overfitting* dan sangat efektif dalam menangani *dataset* yang kompleks (Breiman, 2001). *Random forest* mampu menghasilkan model klasifikasi yang lebih stabil, akurat, dan konsisten dalam mengenali pola data yang beragam dengan menggabungkan banyak pohon keputusan (Hastie *et al.*, 2009).

Random forest telah terbukti unggul dalam berbagai studi terkait klasifikasi curah hujan. Penelitian dilakukan oleh Maheswara dan Al'aziz (2025) menunjukkan bahwa klasifikasi curah hujan di Bogor menggunakan metode *random forest* mencapai nilai *F1-score* sebesar 65,5% melampaui performa *naïve bayes* dan *decision tree*. Indaryono *et al.* (2025) meneliti klasifikasi curah hujan berdasarkan iklim di Indonesia dengan algoritma yang sama pada data curah hujan memperoleh akurasi sebesar 85,55%. Hasil-hasil tersebut menunjukkan bahwa *random forest* merupakan metode yang andal dalam menganalisis pola curah hujan di berbagai wilayah dengan karakteristik iklim yang berbeda.

Data curah hujan sering kali memiliki distribusi kelas yang tidak seimbang (*imbalanced*) dalam praktiknya. Ketidakseimbangan ini disebabkan oleh perbedaan kondisi cuaca ekstrem antarwilayah dan antarwaktu. Ketidakseimbangan ini dapat berdampak buruk terhadap kinerja model klasifikasi karena model cenderung lebih akurat dalam mengenali kelas mayoritas tetapi mengabaikan kelas minoritas yang justru sering menjadi fokus penting dalam analisis. Beberapa pendekatan umum digunakan diantaranya adalah teknik *undersampling* dan *oversampling* untuk mengatasi permasalahan ini. *Undersampling* mengurangi banyaknya data pada kelas

mayoritas secara acak namun berisiko menghilangkan informasi penting. *Oversampling* memperbanyak data kelas minoritas dengan menduplikasi secara acak yang berpotensi menimbulkan *overfitting* akibat kemunculan data identik (Napierala, 2012).

Synthetic Minority Oversampling Technique (SMOTE) dan *Adaptive Synthetic Sampling Approach* (ADASYN) merupakan teknik *oversampling* berbasis sintetik yang dikembangkan sebagai solusi yang lebih canggih dalam menangani permasalahan ketidakseimbangan kelas pada data. SMOTE menghasilkan data sintetis dengan menginterpolasi antara sampel minoritas dan tetangganya sehingga menciptakan variasi baru yang tidak identik dengan data asli. ADASYN mengembangkan pendekatan ini dengan lebih adaptif yaitu menghasilkan lebih banyak sampel sintetis di area yang sulit diklasifikasikan sehingga membantu model belajar lebih baik terhadap distribusi kelas minoritas yang kompleks (He *et al.*, 2008).

Taneja *et al.* (2019) melakukan studi komparatif antara SMOTE dan ADASYN dengan menggunakan algoritma *random forest*, *light gradient boosting*, dan *extreme gradient boosting* pada *dataset* transaksi kartu kredit. Hasilnya menunjukkan bahwa metode SMOTE yang dikombinasikan dengan algoritma *random forest* memberikan performa terbaik dengan nilai *F1-score* sebesar 85%. Penelitian serupa dilakukan oleh Imani *et al.* (2025) membandingkan SMOTE dan ADASYN menggunakan algoritma *XGBoost* pada *dataset customer churn*. Hasil penelitian tersebut menunjukkan bahwa SMOTE menghasilkan *F1-score* sebesar 73% dan dinilai lebih unggul dalam mengatasi ketidakseimbangan kelas secara ekstrem terutama ketika dikombinasikan dengan *XGBoost*.

Penelitian lainnya oleh Ramadhan (2021) menunjukkan bahwa ADASYN memiliki kinerja yang lebih baik dibandingkan SMOTE dalam klasifikasi data Tipe 2 Diabetes Mellitus menggunakan algoritma *Support Vector Machine* (SVM). Hasil pengujian menunjukkan bahwa metode ADASYN-SVM memperoleh akurasi sebesar 87,3% sementara SMOTE-SVM hanya mencapai akurasi sebesar 85,4%. Hal ini mengindikasikan bahwa dalam konteks tertentu ADASYN dapat memberikan hasil yang lebih optimal dibandingkan SMOTE.

Berbagai hasil penelitian tersebut menunjukkan bahwa tidak ditemukan pola yang sepenuhnya konsisten dalam evaluasi kinerja antara SMOTE dan ADASYN. Performa kedua metode *oversampling* cenderung bervariasi tergantung pada karakteristik *dataset*, tingkat ketidakseimbangan kelas, serta algoritma klasifikasi yang digunakan. Pemilihan metode *oversampling* yang tepat harus mempertimbangkan konteks permasalahan dan sifat khusus dari *dataset* agar dapat menghasilkan model klasifikasi yang optimal dan andal.

Penelitian ini bertujuan untuk menangani permasalahan ketidakseimbangan kelas pada data curah hujan dengan membandingkan kinerja metode SMOTE dan ADASYN menggunakan algoritma *random forest*. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam memilih pendekatan yang lebih efektif untuk meningkatkan akurasi klasifikasi pada data curah hujan yang tidak seimbang. Model klasifikasi yang dihasilkan dapat digunakan untuk melihat pola curah hujan berdasarkan karakteristik variabel cuaca. Informasi tersebut dapat memberikan *insight* yang relevan dalam perencanaan musiman di berbagai sektor, seperti mitigasi bencana, pengelolaan sumber daya air, dan pertanian. Model juga

dapat dimanfaatkan sebagai sistem peringatan dini apabila nilai indikator iklim menunjukkan potensi terjadinya curah hujan ekstrem.

1.2 Rumusan Masalah

Mengacu pada paparan latar belakang di atas, berikut adalah rumusan masalah yang akan dikaji dalam penelitian ini.

1. Bagaimana penerapan algoritma *random forest* dalam klasifikasi curah hujan di Kota Bandung?
2. Bagaimana kinerja klasifikasi curah hujan di Kota Bandung menggunakan metode *random forest* dengan menerapkan teknik penanganan ketidakseimbangan kelas menggunakan SMOTE dan ADASYN?
3. Bagaimana perbandingan akurasi algoritma *random forest* sebelum dan sesudah penanganan ketidakseimbangan kelas dengan teknik SMOTE dan ADASYN dalam klasifikasi curah hujan di Kota Bandung?

1.3 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini adalah sebagai berikut.

1. Data yang digunakan merupakan data sekunder yang diperoleh dari publikasi Badan Pusat Statistik (BPS) Kota Bandung terkait data curah hujan di Kota Bandung pada Januari 2004 – Desember 2023.
2. Pengolahan data dilakukan menggunakan algoritma *random forest* dengan penanganan data tidak seimbang menggunakan metode SMOTE dan ADASYN pada *software* Python.
3. Variabel prediktor yang digunakan meliputi hari hujan, kelembapan udara, kecepatan angin, penyinaran matahari, tekanan udara, dan temperatur.

4. Label kategori curah hujan yang digunakan sebagai kelas klasifikasi yaitu Rendah, Menengah, Tinggi, dan Sangat Tinggi.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut.

1. Menerapkan klasifikasi pada curah hujan di Kota Bandung dengan metode *random forest* kemudian menganalisis berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*.
2. Menerapkan penanganan data tidak seimbang dengan teknik SMOTE dan ADASYN pada klasifikasi curah hujan di Kota Bandung dengan algoritma *random forest* kemudian menganalisis berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*.
3. Membandingkan kinerja algoritma *random forest* dalam mengklasifikasikan curah hujan di Kota Bandung baik sebelum maupun sesudah dilakukan penanganan data tidak seimbang dengan teknik SMOTE dan ADASYN.