

PREDIKSI KEKERINGAN DENGAN MENGGUNAKAN *RANDOM FOREST* DAN *XGBOOST* SEBAGAI STRATEGI MITIGASI BENCANA DI KABUPATEN MAJENE

Muh. Hijrah^{1*}, Putri Indi Rahayu², Wahyudi³

^{1,2,3} Program Studi Statistika Universitas Sulawesi Barat

*e-mail: muhhijrah@unsulbar.ac.id

DOI: 10.14710/j.gauss.15.1.98-109

Article Info:

Received: 2025-09-24

Accepted: 2026-05-18

Available Online: 2026-05-20

Keywords:

Drought; Weather; Machine Learning; Random Forest; XGBoost.

Abstract: The growing incidence of drought attributable to climate change poses serious challenges to food systems and agricultural livelihoods in Majene Regency, West Sulawesi. To address this, the present study compares the predictive *accuracy* of *Random Forest* and *XGBoost* in classifying drought occurrences from meteorological variables, using 300 monthly observations recorded by BMKG between 2000 and 2024. Exploratory data analysis revealed that 60% of drought events are concentrated in August and the target variable is severely imbalanced (16.3% drought vs. 83.7% non-drought). At the *default threshold* of 0.50, *Random Forest* outperforms *XGBoost* in all metrics. However, because imbalanced data causes both models to underdetect drought at the *default threshold* (*sensitivity* = 0.267), *threshold* optimization was applied by lowering the decision *threshold* to $\tau = 0.15$. The result of *threshold* optimization was significantly improved *Random Forest sensitivity* from 0.267 to 0.867, detecting 13 out of 15 drought events in the test set while maintaining an AUC-ROC of 0.916. Feature importance analysis consistently identifies X3 as the dominant predictor in both models. *Random Forest* with *threshold* $\tau = 0.15$ is recommended as the primary model for drought early warning systems in Majene.

1. PENDAHULUAN

Indonesia sebagai negara kepulauan dengan kekayaan biodiversitas yang luar biasa, kini menghadapi tantangan besar akibat perubahan iklim. Kenaikan suhu tahunan yang terus meningkat memperburuk pemanasan global, memperparah kondisi yang sudah ada (Ainurrohmah & Sudarti, 2022). Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) melaporkan bahwa selama 4 dekade terakhir suhu di Indonesia terus meningkat dengan anomali suhu udara rata-rata 1,2°C dari sebagian besar stasiun pengamatan. Namun yang menjadi ironi adalah Indonesia menempati posisi ke-7 sebagai negara penghasil emisi gas rumah kaca terbesar di dunia (Hafdz, et al., 2023). Dampak krisis ini semakin terasa melalui peningkatan frekuensi dan intensitas terjadinya bencana alam (Neya, et al., 2024). Sementara itu dalam sektor lainnya juga mengancam ketahanan pangan, merusak lingkungan, dan meningkatkan risiko kesehatan masyarakat.

Kabupaten Majene, Provinsi Sulawesi Barat, merupakan daerah yang rentan bencana hidrometeorologi. Sebagian wilayahnya mengalami banjir saat curah hujan tinggi, sementara sebagian lainnya mengalami kekeringan pada musim kemarau. Kedua bencana ini merusak lahan pertanian sehingga mengancam ekonomi dan ketahanan pangan daerah (Ehteram & Shabanian, 2023). Kekeringan secara khusus dikategorikan sebagai bencana hidrologi yang memiliki dampak kumulatif dan seringkali sulit dipulihkan dalam jangka pendek (McKee et al., 1993).

Prediksi kejadian kekeringan secara akurat merupakan prasyarat penting dalam penyusunan sistem peringatan dini dan perencanaan mitigasi bencana. Ketidakpastian yang diakibatkan oleh fluktuasi cuaca yang ekstrem mempersulit proses prediksi konvensional

(Prasetya & Djamal, 2019). Metode machine learning terbukti mampu menangkap pola non-linear yang kompleks pada data iklim dan memberikan performa prediksi yang lebih baik dibandingkan model statistik tradisional (Breiman, 2001; Chen & Guestrin, 2016).

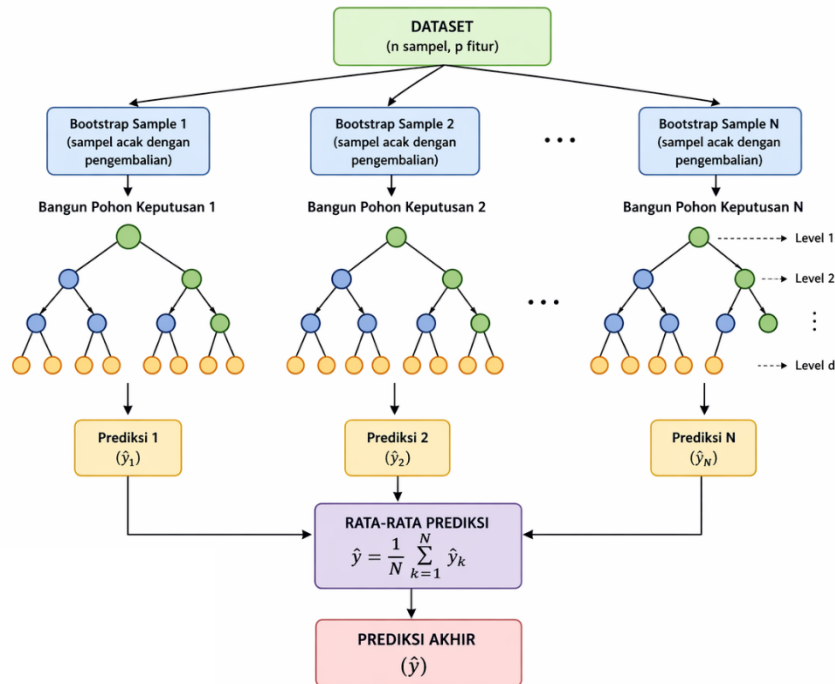
Beberapa penelitian telah membuktikan keunggulan machine learning dalam prediksi cuaca dan iklim. Metode yang sering digunakan meliputi *Random Forest*, ANN, LSTM, dan *XGBoost* (Dwiyanti & Priyanto, 2023). Namun, tantangan khusus pada data kekeringan adalah ketidakseimbangan kelas (*imbalanced class*) yang membuat model cenderung bias ke kelas mayoritas. Pemilihan *Random Forest* dan *XGBoost* dalam penelitian ini didasarkan pada tiga pertimbangan utama. Pertama, dari sisi kesesuaian data: kedua metode terbukti unggul pada data tabular berdimensi rendah dengan hubungan non-linear, sebagaimana karakteristik data BMKG Majene, dibandingkan metode berbasis jaringan saraf dalam (ANN, LSTM) yang memerlukan data dalam jumlah besar (Cholissodin et al., 2020). Kedua, dari sisi penanganan ketidakseimbangan kelas: *Random Forest* menyediakan mekanisme *class weighting* melalui *bootstrap sampling*, sementara *XGBoost* memiliki parameter *scale pos weight* keduanya lebih adaptif terhadap imbalanced data dibandingkan regresi logistik atau Support Vector Machine konvensional (He & Garcia, 2009). Ketiga, dari sisi kebutuhan optimasi *threshold*: berbeda dengan penelitian sebelumnya yang hanya mengevaluasi model pada *threshold default* 0,50, penelitian ini secara eksplisit mengintegrasikan optimasi *threshold* berbasis cost-sensitive learning (Elkan, 2001). Permasalahan ini memerlukan pendekatan evaluasi dan optimasi *threshold* yang lebih hati-hati (He & Garcia, 2009; Elkan, 2001). Penelitian ini bertujuan membandingkan kinerja *Random Forest* dan *XGBoost* dalam memprediksi kekeringan di Kabupaten Majene, termasuk optimasi *threshold* keputusan untuk meningkatkan kemampuan deteksi kekeringan sebagai bagian dari strategi mitigasi bencana.

2. TINJAUAN PUSTAKA

Machine learning adalah bagian dari kecerdasan buatan yang menitikberatkan pada pemanfaatan data melalui pengembangan sistem yang dapat belajar secara otomatis tanpa pemrograman ulang secara terus-menerus oleh manusia (Cholissodin et al., 2020). Secara konseptual, machine learning merupakan bidang yang mempelajari penerapan algoritma dan model statistik agar komputer mampu menyelesaikan tugas tertentu tanpa instruksi yang dirinci secara langsung (Dinata & Hasdyna, 2020).

Seiring perkembangan teknologi dan ketersediaan data dalam jumlah besar (*big data*), machine learning semakin banyak diterapkan dalam berbagai bidang, seperti kesehatan, keuangan, pertanian, hingga mitigasi bencana. Metode-metode dalam machine learning umumnya diklasifikasikan menjadi *supervised learning*, *unsupervised learning*, dan *reinforcement learning*, yang masing-masing memiliki karakteristik dan tujuan berbeda. Keunggulan utama pendekatan ini terletak pada kemampuannya dalam mengidentifikasi pola kompleks serta melakukan prediksi dengan tingkat akurasi yang tinggi, sehingga sangat relevan untuk mendukung pengambilan keputusan berbasis data (Bishop, 2006).

Random Forest adalah salah satu metode *ensemble learning* berbasis pohon keputusan yang paling populer. Algoritma ini bekerja dengan membangun banyak pohon keputusan yang independen selama fase pelatihan. Setiap pohon dilatih pada subset data yang dipilih secara acak. Prediksi akhir diperoleh dengan menggabungkan hasil dari semua pohon (misalnya, melalui voting untuk masalah klasifikasi atau rata-rata untuk masalah regresi). Pendekatan ini, yang dikenal sebagai *bagging* (*Bootstrap Aggregating*), efektif dalam mengurangi varians dan mencegah *overfitting*, menghasilkan model yang kokoh dan akurat.



Gambar 1. Ilustrasi *Random Forest*

Pada kasus regresi, nilai prediksi akhir dari algoritma *Random Forest* diperoleh dengan cara merata-ratakan hasil prediksi yang dihasilkan oleh seluruh pohon keputusan yang terbentuk dalam model.

$$\hat{y}_i = \frac{1}{N_{tree}} \sum_{k=1}^{N_{tree}} \hat{y}_k$$

dimana \hat{y}_i merupakan nilai prediksi akhir, N_{tree} menunjukkan jumlah total pohon dalam *Random Forest*, dan \hat{y}_k adalah hasil prediksi dari pohon ke- k (Lestari & Astuti, 2022). Kemampuan model *Random Forest* dalam memprediksi data baru yang tidak termasuk dalam data pelatihan dievaluasi menggunakan *mean-squared generalization error*. Ukuran ini dinyatakan sebagai berikut:

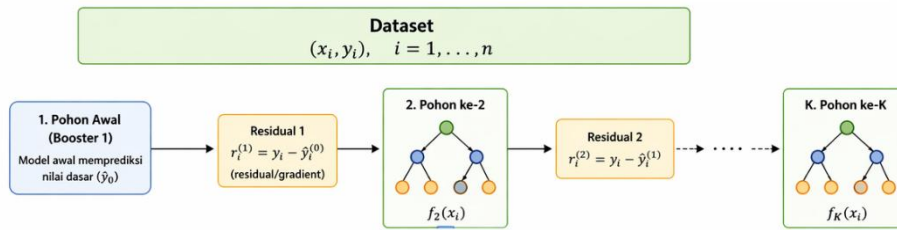
$$E_{(X,Y)}[(Y - h(X))^2]$$

dengan $h(X)$ sebagai fungsi prediksi yang diperoleh dari proses pelatihan, Y sebagai variabel target, dan $E_{(X,Y)}$ merupakan nilai harapan dari distribusi data. Nilai ini mencerminkan seberapa baik model mampu menggeneralisasi pola dari data latih ke data yang belum pernah diamati sebelumnya (Breiman, 2001).

Selain itu, dalam proses pembentukan pohon, *Random Forest* menerapkan pemilihan variabel secara acak guna meningkatkan keragaman antar pohon. Proses ini dilakukan dengan memilih sejumlah variabel m dari total variabel independen p , dengan ketentuan $m \leq p$. Untuk kasus regresi (Hastie et.al., 2008), jumlah variabel yang dipilih pada setiap proses pemisahan (*mtry*) dihitung menggunakan persamaan berikut:

$$m_{try} = \frac{p}{3}$$

XGBoost atau *Extreme Gradient Boosting* merupakan salah satu algoritma machine learning yang mengadopsi pendekatan *gradient boosting* untuk menghasilkan model berbasis pohon yang efisien dan memiliki performa tinggi, terutama pada skala data besar. Dalam konteks regresi, *XGBoost* digunakan untuk memprediksi nilai kontinu, dimana setiap pohon keputusan berperan dalam memodelkan target sekaligus menghitung residual dari model sebelumnya (Nugraha & Irawan, 2023).



Gambar 2. Ilustrasi *XGBoost*

Berdasarkan ilustrasi *XGBoost*, fungsi $f_t(x_i)$ merepresentasikan model pohon pada iterasi ke- t , dengan nilai prediksi yang dinyatakan sebagai penjumlahan kontribusi dari seluruh pohon hingga iterasi tersebut:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i)$$

Fungsi objektif dalam *XGBoost* terdiri atas dua komponen utama, yaitu fungsi kerugian (*loss function*) dan komponen regularisasi. Secara umum, fungsi objektif dapat didekati sebagai berikut:

$$Obj^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

dengan g_i sebagai turunan pertama (*gradient*) dan h_i sebagai turunan kedua (*Hessian*) dari fungsi kerugian terhadap prediksi sebelumnya. Kompleksitas model dikontrol melalui fungsi regularisasi yang melibatkan jumlah daun pohon (T) dan bobot setiap daun (w_j), yang dirumuskan sebagai:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Dengan memasukkan komponen regularisasi, fungsi objektif dapat dituliskan dalam bentuk yang lebih sederhana sebagai:

$$Obj^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Selanjutnya, dengan mengelompokkan data berdasarkan node daun, fungsi objektif dapat direformulasikan menjadi:

$$Obj^{(t)} = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T$$

dengan G_j sebagai jumlah gradien dan H_j sebagai jumlah Hessian pada node ke- j . Bobot optimal untuk setiap daun diperoleh dengan meminimalkan fungsi objektif, sehingga diperoleh:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

Jika nilai ini disubstitusikan kembali, maka nilai optimal fungsi objektif menjadi:

$$Obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T$$

Dalam proses pembentukan pohon, *XGBoost* menggunakan nilai *gain* untuk menentukan pemisahan terbaik pada setiap node. Nilai *gain* dihitung sebagai:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

Nugraha & Irawan (2023) mengungkapkan bahwa nilai ini mencerminkan peningkatan kualitas model akibat proses pemisahan node. Pemisahan dengan nilai *gain* tertinggi akan dipilih sebagai split terbaik, dan proses ini dilakukan secara berulang hingga struktur pohon optimal terbentuk. Secara keseluruhan, *XGBoost* membangun prediksi melalui pendekatan *boosting* secara bertahap. Model akhir dinyatakan sebagai:

$$\hat{y}_i = \hat{y}_0 + \eta \sum_{k=1}^K f_k(x_i)$$

dimana \hat{y}_0 adalah prediksi awal, $f_k(x_i)$ adalah pohon ke- k , K jumlah total pohon, dan η merupakan *learning rate* yang mengatur kontribusi masing-masing pohon dalam pembaruan model (Zhang, et al., 2022).

Evaluasi model merupakan tahapan penting dalam *machine learning* untuk menilai kinerja model dalam melakukan prediksi terhadap data baru. Proses ini umumnya menggunakan *confusion matrix* yang terdiri dari empat komponen utama, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) (Powers, 2020).

Berdasarkan komponen tersebut, berbagai metrik evaluasi dapat dihitung untuk memberikan gambaran performa model secara komprehensif. Salah satu metrik yang paling umum digunakan adalah akurasi (*accuracy*), yang mengukur proporsi prediksi yang benar terhadap seluruh jumlah data. Secara matematis, akurasi dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Meskipun sederhana dan mudah dipahami, akurasi kurang efektif pada dataset yang tidak seimbang (*imbalanced data*) karena dapat memberikan nilai tinggi meskipun model tidak mampu mengenali kelas minoritas dengan baik (He & Garcia, 2009).

Selanjutnya, *sensitivitas* (*sensitivity*) atau *recall* mengukur kemampuan model dalam mengidentifikasi seluruh data positif yang sebenarnya. Sensitivitas sangat penting dalam kasus dimana kegagalan mendeteksi kelas positif memiliki konsekuensi besar. Rumus sensitivitas adalah:

$$Sensitivity = \frac{TP}{TP + FN}$$

Sebaliknya, spesifisitas (*specificity*) mengukur kemampuan model dalam mengidentifikasi data negatif secara benar. Metrik ini penting untuk memastikan bahwa model tidak menghasilkan terlalu banyak kesalahan positif (*false positive*). Secara matematis:

$$Specificity = \frac{TN}{TN + FP}$$

Untuk memperoleh keseimbangan antara sensitivitas dan presisi, digunakan *F1-score*, yang merupakan rata-rata harmonik dari *precision* dan *recall*. *F1-score* sangat berguna pada kasus klasifikasi dengan distribusi kelas yang tidak seimbang (Sokolova & Lapalme, 2009):

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

dimana:

$$Precision = \frac{TP}{TP + FP}$$

F1-score sangat berguna terutama pada kasus klasifikasi dengan distribusi kelas yang tidak seimbang karena memberikan ukuran yang lebih seimbang antara kesalahan tipe I dan tipe II (Sokolova & Lapalme, 2009).

Selain itu, *Area Under the Curve* (AUC) yang berasal dari kurva ROC (*Receiver Operating Characteristic*) digunakan untuk mengukur kemampuan model dalam membedakan antara kelas positif dan negatif pada berbagai nilai ambang (*threshold*). Kurva ROC sendiri dibangun dari hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR), yang masing-masing dirumuskan sebagai:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Nilai AUC berada pada rentang 0 hingga 1, dimana nilai yang mendekati 1 menunjukkan kemampuan diskriminasi model yang sangat baik (Hanley & McNeil, 1982). Secara keseluruhan, penggunaan metrik evaluasi seperti *accuracy*, *precision*, *specificity*, *recall*, *F1-score*, dan AUC secara bersama-sama sangat penting untuk memperoleh gambaran kinerja model yang lebih menyeluruh, khususnya pada kasus dengan distribusi data yang tidak seimbang atau ketika kesalahan klasifikasi memiliki konsekuensi yang berbeda.

3. METODE PENELITIAN

Data yang digunakan pada penelitian ini merupakan data sekunder yang diambil dari BMKG Sulawesi Barat. Fitur dan Target Variabel yang digunakan pada penelitian ini adalah 5 fitur dari faktor cuaca dan 1 target variabel yakni kejadian kekeringan di Kabupaten Majene yang dijelaskan pada tabel berikut:

Tabel 1. Variabel penelitian

Fitur dan Target Variabel	Satuan
Lama Penyinaran Matahari (X_1)	%
Kecepatan Angin Rata-rata (X_2)	m/d
Kelembaban Relatif (X_3)	%
Temperatur Maksimum (X_4)	$^{\circ}C$
Temperatur Minimum (X_5)	$^{\circ}C$
Target Variabel (Y)	0 atau 1

Adapun definisi operasional dari variabel penelitian yang digunakan adalah sebagai berikut:

1. Lama Penyinaran Matahari (X_1) yang menunjukkan variabilitas lama penyinaran matahari dimana hal ini ditinjau dari periode pengamatan dalam hal ini bulanan.
2. Kecepatan Angin (X_2) yang menunjukkan rata-rata jarak tempuh angin atau pergerakan udara persatuan waktu.
3. Kelembaban Relatif (X_3) yang menunjukkan banyaknya uap air yang terkandung dalam udara atau atmosfer.
4. Temperatur Maksimum (X_4) yang menyatakan ukuran dingin atau panasnya keadaan yang paling panas dibandingkan temperatur-temperatur lainnya.

5. Temperatur Minimum (X_5) yang menyatakan ukuran dingin atau panasnya keadaan yang paling dingin dibandingkan temperatur-temperatur lainnya.
6. Target Variabel (Y) adalah kejadian terjadinya kekeringan nilai 1 jika terjadi kekeringan dan nilai 0 jika tidak terjadi kekeringan. Variabel ini diperoleh dengan menggunakan data curah hujan untuk mendapatkan nilai *Standard Precipitation Index* (SPI). SPI pertama kali diperkenalkan oleh Thomas B. McKee pada tahun 1993 sebagai metode yang sederhana, fleksibel, dan dapat diterapkan pada berbagai skala waktu untuk mengidentifikasi kejadian kekeringan meteorologis (McKee et al., 1993). Nilai SPI yang lebih kecil dari 0 dikategorikan sebagai kejadian kekeringan.

Adapun tahapan analisis data yang dilakukan pada penelitian ini sebagai berikut:

1. Mengumpulkan data cuaca historis yang relevan untuk Kabupaten Majene, mencakup variabel seperti suhu, curah hujan, kelembapan, tekanan udara, dan kecepatan angin, dengan periode waktu tahun 2000 hingga 2024.
2. Melakukan eksplorasi dan visualisasi data untuk mengetahui deskripsi data secara menyeluruh.
3. Melakukan *pre-processing* data (pembersihan data, penanganan data hilang, transformasi data, membuat fitur baru (*feature engineering*), dan normalisasi dan standarisasi data).
4. Melakukan pemilihan dan pembagian data. Data akan dibagi menjadi 70% data untuk pelatihan dan 30% data untuk pengujian.
5. Pengembangan model dengan *Random Forest* dan *XGBoost*.
6. Melakukan evaluasi model.
7. Melakukan optimasi *threshold Random Forest* dan *XGBoost*
8. Melakukan penarikan kesimpulan dari hasil analisis yang didapatkan.

Data bulanan BMKG memiliki karakteristik yang sesuai untuk kedua metode ensemble berbasis pohon. Pertama, data bersifat tabular berdimensi rendah (5 fitur prediktor), yang merupakan kekuatan utama *Random Forest* dan *XGBoost* dibandingkan metode berbasis jaringan saraf. Kedua, distribusi variabel target yang tidak seimbang dengan proporsi kekeringan ($Y = 1$) hanya 16,3% dapat ditangani melalui mekanisme bootstrap sampling berbobot pada *Random Forest* dan parameter *scale pos weight* pada *XGBoost*. Ketiga, hubungan non-linear antara prediktor dan kekeringan lebih tepat dimodelkan dengan pohon keputusan. Keempat, interaksi antar variabel cuaca ditangkap secara alami oleh struktur pohon (Breiman, 2001; Chen & Guestrin, 2016).

4. HASIL DAN PEMBAHASAN

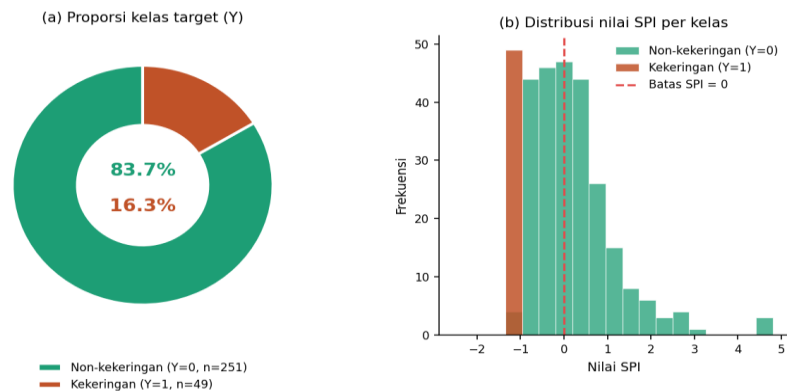
Penelitian ini merupakan pengembangan signifikan dari studi sebelumnya yang cenderung hanya menggunakan satu model *machine learning*. Kebaruan utama dalam penelitian ini adalah perbandingan kinerja antara dua model *ensemble learning* terkemuka, yaitu *XGBoost* dan *Random Forest*, dalam konteks spesifik prediksi kekeringan di Kabupaten Majene. Penelitian ini akan memberikan wawasan yang mendalam tentang karakteristik model mana yang paling sesuai untuk menangani pola data cuaca lokal yang kompleks khususnya di Kabupaten Majene, Sulawesi Barat.

Tabel 2. Statistik Deskriptif

Ukuran Statistik	Y	X_1	X_2	X_3	X_4	X_5
Mean	0,163	6,173	5,094	79,717	31,887	24,661

Median	0	6,164	4,705	80,147	31,803	24,705
SD	0,370	1,217	1,802	3,238	0,828	0,564
Min	0	3,019	2,767	70,167	30,1	22,677
Max	1	9,540	13,452	88,968	34,171	26,584

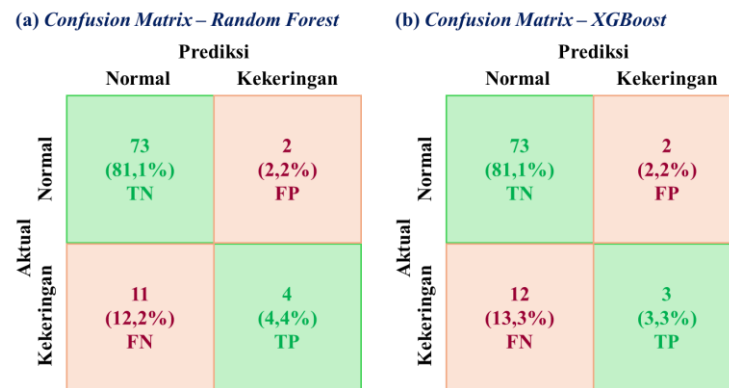
Berdasarkan Tabel 2, variabel respons Y memiliki rata-rata 0,163 yang mengindikasikan ketidakseimbangan kelas (*imbalanced data*): hanya 16,3% observasi merupakan kejadian kekeringan (49 dari 300 bulan) dengan rasio 1:5,1 antara kelas minoritas dan mayoritas. Ketidakseimbangan kelas ini menjadi pertimbangan utama dalam pemilihan metode dan optimasi *threshold* (He & Garcia, 2009). Variabel X3 (kelembaban relatif) memiliki rentang terlebar (70,167–88,968%) dengan variasi sedang (SD = 3,238), sementara X2 (kecepatan angin) menunjukkan keragaman tertinggi (SD = 1,802). Variabel X4 dan X5 relatif homogen (SD < 1), mengindikasikan stabilitas suhu sepanjang tahun.



Gambar 3 (a) Proporsi kelas target Y (kekeringan vs. non-kekeringan); (b) distribusi nilai SPI

Gambar 3(a) memperlihatkan ketidakseimbangan kelas yang signifikan: 83,7% observasi bukan kekeringan dan hanya 16,3% adalah kejadian kekeringan. Gambar 1(b) menunjukkan bahwa nilai SPI saat kekeringan ($Y = 1$) terpusat pada rentang $-0,5$ hingga $-2,5$, sedangkan saat kondisi normal SPI cenderung positif. Pemisahan distribusi ini mengkonfirmasi bahwa SPI merupakan dasar pelabelan yang valid dan konsisten (McKee et al., 1993). Variabel prediktor X_1 hingga X_5 menunjukkan variasi yang cukup beragam. Variabel X_3 memiliki nilai rata-rata paling tinggi (79,717) dengan simpangan baku sebesar 3,238, yang menunjukkan bahwa data pada variabel ini relatif terkonsentrasi di sekitar nilai mean dengan variasi sedang. Sementara itu, variabel X_2 memiliki rentang nilai yang cukup lebar (minimum 2,767 hingga maksimum 13,452) serta simpangan baku terbesar (1,802), yang mengindikasikan tingkat keragaman data yang lebih tinggi dibandingkan variabel lainnya. Variabel X_4 dan X_5 menunjukkan variasi yang relatif kecil dengan simpangan baku masing-masing sebesar 0,828 dan 0,564, sehingga data pada kedua variabel ini cenderung lebih homogen. Selain itu, perbedaan antara nilai mean dan median pada sebagian besar variabel relatif kecil, yang menunjukkan bahwa distribusi data cenderung simetris dan tidak terlalu dipengaruhi oleh pencilan (*outlier*).

Secara keseluruhan, hasil statistik deskriptif ini memberikan gambaran awal mengenai karakteristik data yang akan dianalisis lebih lanjut. Adanya indikasi ketidakseimbangan pada variabel respon serta variasi antar variabel prediktor menjadi pertimbangan penting dalam pemilihan metode analisis *Random Forest* dan *XGBoost* agar mampu menangani distribusi data yang tidak seimbang dan kompleksitas pola yang ada.



Gambar 4. *Confusion matrix* (a) *Random Forest* dan (b) *XGBoost* pada data uji ($n = 90$)

Gambar 4 memperlihatkan bahwa kedua model menghasilkan *confusion matrix* yang tidak jauh berbeda pada *threshold default* (0,5): 73 TN (non-kekeringan diprediksi benar), 2 FP (non-kekeringan salah diprediksi kekeringan), 11 FN (kekeringan tidak terdeteksi) pada *Random Forest* dan 12 FN pada *XGBoost*, serta 4 TP (kekeringan terdeteksi benar) pada *Random Forest* dan 3 TP pada *XGBoost*. *Specificity* yang tinggi (TN = 73 dari 75 kasus normal, atau 97,3%) menunjukkan bahwa model sangat akurat dalam mengidentifikasi bulan normal. Namun, hanya 3 pada *XGBoost* dan 4 pada *Random Forest* dari 15 kejadian kekeringan yang terdeteksi pada *threshold* ini.

Tabel 3. Hasil Evaluasi Kinerja Model *Random Forest* dan *XGBoost* (*Threshold*=0,5)

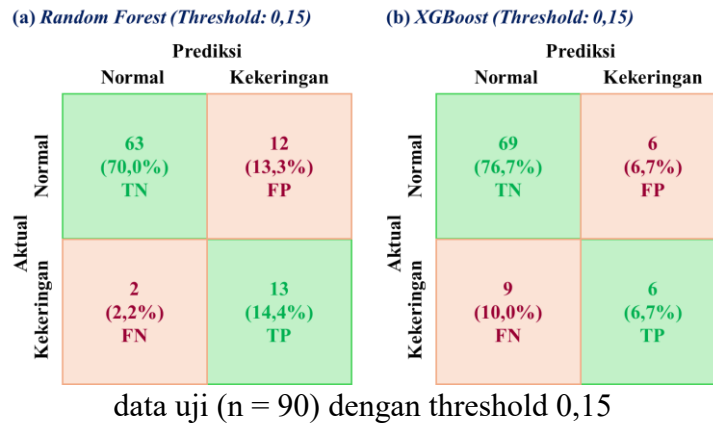
Metrik Evaluasi	<i>Random Forest</i>	<i>XGBoost</i>
<i>Accuracy</i>	0,856	0,844
<i>Sensitivity</i>	0,267	0,200
<i>Specificity</i>	0,973	0,973
<i>F1-Score</i>	0,381	0,300
AUC-ROC	0,795	0,817

Berdasarkan Tabel 3, terlihat kinerja antara model *Random Forest* dan *XGBoost* dalam mengevaluasi data klasifikasi tidak menunjukkan perbedaan yang signifikan. Secara umum, *Random Forest* menunjukkan performa yang sedikit lebih baik pada sebagian besar metrik, khususnya pada *accuracy*, *sensitivity*, dan *F1-score*, sedangkan *XGBoost* unggul pada nilai AUC.

Hasil evaluasi pada *threshold default* $\tau = 0,50$ memberikan gambaran yang menunjukkan bahwa akurasi tinggi tidak selalu bermakna performa yang baik dalam konteks prediksi bencana. Kedua model menghasilkan *accuracy* di atas 0,84, namun hanya mampu mendeteksi kurang dari sepertiga kejadian kekeringan sesungguhnya pada data uji. Kondisi ini mencerminkan bias klasik pada data tidak seimbang. Model perlu dioptimalkan untuk meminimalkan kesalahan keseluruhan, namun mengabaikan distribusi *cost* yang asimetris antara kedua jenis kesalahan prediksi (Elkan, 2001). Dalam sistem peringatan dini kekeringan, *cost* melewatkan satu kejadian kekeringan (*false negative*) secara empiris jauh lebih besar daripada *cost* satu alarm berlebih (*false positive*) sebuah prinsip yang semestinya tercermin dalam konfigurasi operasional model. Oleh karena itu, pendekatan *cost-sensitive*

melalui penurunan *threshold* keputusan perlu diterapkan untuk menyelaraskan perilaku model dengan kebutuhan nyata mitigasi bencana di Kabupaten Majene (Hastie et al., 2009; He & Garcia, 2009).

Gambar 5. *Confusion matrix* (a) *Random Forest* dan (b) *XGBoost* pada



Setelah dilakukan optimasi *threshold* pada kedua model seperti yang ditunjukkan oleh Gambar 5. *Confusion matrix* tersebut memperlihatkan pergeseran distribusi prediksi yang substansial. Pada *Random Forest*, penurunan *threshold* mengubah distribusi dari 73 TN, 2 FP, 11 FN, 4 TP menjadi 63 TN, 12 FP, 2 FN, 13 TP. Penurunan FN dari 11 menjadi 2 kasus (berkurang 81,8%) merupakan pencapaian kritis dalam konteks mitigasi bencana. Pada *XGBoost*, pergeseran lebih moderat yakni dari 73 TN, 2 FP, 12 FN, 3 TP menjadi 69 TN, 6 FP, 9 FN, 6 TP. Meskipun FP *XGBoost* lebih sedikit (6 vs 12), FN-nya masih tinggi (9 vs 2), menjadikan *Random Forest* $\tau = 0,15$ sebagai model yang lebih unggul untuk sistem peringatan dini (He & Garcia, 2009; Elkan, 2001).

Tabel 4. Hasil Evaluasi Kinerja Model *Random Forest* dan *XGBoost* (Threshold=0,15)

Metrik Evaluasi	<i>Random Forest</i>	<i>XGBoost</i>
<i>Accuracy</i>	0,844	0,833
<i>Sensitivity</i>	0,867	0,400
<i>Specificity</i>	0,840	0,920
F1-Score	0,650	0,444
AUC-ROC	0,916	0,873

Berdasarkan Tabel 4, *Random Forest* dengan *threshold* $\tau = 0,15$ merupakan model terbaik untuk sistem peringatan dini kekeringan di Kabupaten Majene. Dengan *sensitivity* = 0,867, model ini mampu mendeteksi 13 dari 15 kejadian kekeringan pada data uji. Meskipun menghasilkan lebih banyak *false positive* (12 kasus) dibandingkan *threshold default* (2 kasus), peningkatan *alarm* berlebih ini merupakan *trade-off* yang dapat diterima mengingat biaya kekeringan yang tidak terdeteksi (*false negative*) jauh lebih besar dari biaya kesiapsiagaan berlebih dalam konteks mitigasi bencana (Elkan, 2001).

5. KESIMPULAN

Berdasarkan eksplorasi data dan evaluasi model, disimpulkan beberapa hal berikut. Pertama, data kekeringan Kabupaten Majene periode 2000–2024 menunjukkan pola musiman yang jelas: 60% kejadian kekeringan terkonsentrasi di bulan Agustus. Terdapat ketidakseimbangan kelas yang signifikan (16,3% vs. 83,7%) yang memerlukan strategi khusus dalam pemodelan dan evaluasi (He & Garcia, 2009).

Optimasi threshold berbasis *cost-sensitive learning* dan *maximizing F1-Score* (Elkan, 2001; Sokolova & Lapalme, 2009) menghasilkan *threshold optimal* $\tau = 0,15$ untuk *Random Forest*. Dengan threshold ini, Sensitivity meningkat dari 0,267 menjadi 0,867 (13 dari 15 kekeringan terdeteksi), FN berkurang dari 11 menjadi 2 kasus, dan *F1-Score* meningkat dari 0,381 menjadi 0,650, sementara AUC-ROC tetap 0,916. *Trade-off* peningkatan FP dari 2 menjadi 12 kasus dapat diterima mengingat *cost* kekeringan yang tidak terdeteksi jauh lebih besar dari *alarm cost* berlebih.

Random Forest dengan *threshold* $\tau = 0,15$ direkomendasikan sebagai model utama sistem peringatan dini kekeringan di Kabupaten Majene, mengingat kemampuannya mendeteksi 86,7% kejadian kekeringan pada data uji. Implementasinya dapat mendukung penyusunan kalender tanam adaptif, alokasi cadangan pangan strategis, dan pengelolaan sumber daya air yang lebih resilien terhadap perubahan iklim. Penelitian lanjutan disarankan untuk mengeksplorasi teknik SMOTE guna lebih meningkatkan keseimbangan kelas, serta memperluas cakupan spasial ke seluruh wilayah Sulawesi Barat.

ACKNOWLEDGEMENT

Penelitian ini didanai oleh Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) Tahun 2025 melalui skema Hibah Penelitian Dosen Pemula (PDP). Penulis menyampaikan apresiasi dan terima kasih yang sebesar-besarnya atas dukungan pendanaan yang telah diberikan sehingga penelitian ini dapat terlaksana dengan baik.

Penulis juga mengucapkan terima kasih kepada Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) atas penyediaan data yang digunakan dalam penelitian ini. Selain itu, penulis mengapresiasi dukungan dari Universitas Sulawesi Barat yang telah memberikan fasilitas serta lingkungan akademik yang kondusif dalam pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan kepada semua pihak yang telah memberikan bantuan, baik secara langsung maupun tidak langsung, dalam proses penyusunan dan penyelesaian penelitian ini.

DAFTAR PUSTAKA

- Ainurrohmah, S. & Sudarti, S. 2022. Analisis Perubahan Iklim dan Global Warming yang Terjadi sebagai Fase Kritis. *Jurnal Phi: Jurnal Pendidikan Fisika dan Fisika Terapan*, 8(1), hlm. 1–10.
- Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1), hlm. 5–32. doi: 10.1023/A:1010933404324.
- Chen, T. & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, hlm. 785–794. doi: 10.1145/2939672.2939785.
- Cholissodin, I., Sutrisno, Soebroto, A. A., Hasanah, U. & Febiola, Y. I. 2020. *AI, Machine Learning & Deep Learning (Teori & Implementasi)*. Malang: FILKOM UB.
- Dwiyanti, Z.A. & Prianto, C. 2023. Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest. *Jurnal Tekno Insentif*, 17(2), hlm. 127–137.
- Ehteram, M. & Shabanian, H. 2023. Unveiling the SALSTM-M5T Model and Its Python Implementation for Precise Solar Radiation Prediction. *Energy Reports*, 10, hlm. 3402–3417.
- Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, hlm. 973–978.
- Hafdz, D., Wardhana, A. & Prawira, M.R. 2023. The Analysis of Indonesia's Climate Change Policies in Response to the 2021 IPCC Assessment Report. *Jurnal Internasional Ilmu Politik dan Hubungan Internasional*, 1, hlm. 42–53.

- Hanley, J.A. & McNeil, B.J. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1), hlm. 29–36.
- Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.
- He, H. & Garcia, E.A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), hlm. 1263–1284.
- Lestari, E.S. & Astuti, I. 2022. Penerapan Random Forest Regression untuk Memprediksi Harga Jual Rumah. *Jurnal Ilmiah FIFO*, 14(2), hlm. 131. doi: 10.22441/fifo.2022.v14i2.003.
- McKee, T.B., Doesken, N.J. & Kleist, J. 1993. The Relationship of Drought Frequency and Duration to Time Scales. *Proceedings of the 8th Conference on Applied Climatology*, hlm. 179–183.
- Neya, T. et al. 2024. Climate Change Impact Assessment and Disaster Risk Financing Strategies in Mali. *International Journal of Environment and Climate Change*, 14(3), hlm. 126–138.
- Nugraha, A.C. & Irawan, M.I. 2023. Komparasi Deteksi Kecurangan pada Data Klaim Asuransi Menggunakan SVM dan XGBoost. *Jurnal Sains dan Seni ITS*, 12(1). doi: 10.12962/j23373520.v12i1.107032.
- Powers, D.M.W. 2020. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1), hlm. 37–63.
- Prasetya, E.P. & Djamal, E.C. 2019. Rainfall Forecasting for the Natural Disasters Preparation Using Recurrent Neural Networks. *Proceedings of the International Conference on Electrical Engineering and Informatics (ICEEI)*, hlm. 52–57.
- Sokolova, M. & Lapalme, G. 2009. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 45(4), hlm. 427–437.
- Zhang, P., Jia, Y. & Shang, Y. 2022. Research and Application of XGBoost in Imbalanced Data. *International Journal of Distributed Sensor Networks*, 18(6). doi: 10.1177/15501329221106935.