Muhammad Tito Jaya Kusuma

Introduction [-0.5 halaman]

Bencana banjir bandang yang terjadi di Sumatera pada akhir 2025 kembali menunjukkan bahwa wilayah Indonesia masih sangat rentan terhadap cuaca ekstrem, kerusakan lingkungan, dan gagalnya sistem mitigasi bencana. Dampak banjir bandang sangat fatal, merusak infrastruktur, menimbulkan korban jiwa, mengganggu aktivitas ekonomi, serta mempersulit proses evakuasi. Oleh karena itu, kemampuan untuk **memprediksi potensi banjir secara dini** menjadi hal yang sangat penting agar pemerintah dan masyarakat dapat mengambil tindakan mitigasi sebelum bencana terjadi.

Motivasi utama dalam penelitian ini adalah menciptakan sistem prediksi banjir yang cepat, akurat, dan dapat diterapkan sebagai **peringatan dini** pada daerah rawan banjir di Sumatera. Dengan memanfaatkan data lingkungan, sosial, dan hidrologis, model prediksi ini diharapkan mampu membantu pengambilan keputusan, penentuan level kewaspadaan, serta meminimalkan risiko kerugian.

- **Input** pada permasalahan ini terdiri dari 20 variabel numerik yang mewakili faktor penyebab banjir, seperti *MonsoonIntensity*, *Deforestation*, *DrainageSystems*, *Urbanization*, *PoliticalFactors*, dan variabel lingkungan lainnya. Data input diperoleh dari dataset flood.csv yang berisi 50.000 baris data.
- Output dari sistem ini adalah nilai prediksi **FloodProbability**, yaitu probabilitas terjadinya banjir berat pada suatu wilayah.

Untuk memproses input tersebut, penelitian ini menggunakan beberapa algoritma machine learning, seperti **Random Forest, Gradient Boosting, XGBoost, dan Support Vector Regressor (SVR)**. Model-model ini dibandingkan untuk menentukan algoritma yang paling akurat dalam memprediksi probabilitas banjir. Dengan pendekatan ini, proyek ini menghasilkan solusi prediksi banjir berbasis data yang dapat digunakan sebagai sistem mitigasi awal pada daerah rawan bencana.

Related Work [~0.5 halaman]

· Anda harus cari paper

· Selanjutnya kelompokan kedalam kategori berdasarkan pendekatan yang dilakukan.

· Kemudian diskusikan kelebihan dan kekurangannya, serta persamaan & perbedaan
dengan yang Anda kerjakan sekarang.

· Saran: coba cari minimal 5 paper untuk mendukung proses pembuatan model Anda.

· Tuliskan opini-opini Anda pada bagian ini.

Dataset & Features [~0.5 - 1 halaman]

Dataset Description

Dataset yang digunakan pada penelitian ini adalah **flood.csv**, yang bersumber dari Kaggle Flood Prediction Dataset (Naiya Khalid, 2023). Dataset ini berisi **50.000 baris dan 21 kolom**, yang memuat berbagai variabel lingkungan, sosial, hidrologis, dan infrastruktur yang mempengaruhi risiko banjir pada suatu wilayah.

Dataset ini terdiri dari **20 fitur input** dan **1 fitur output** yaitu *FloodProbability*, yang merepresentasikan probabilitas terjadinya banjir berat pada rentang 0.285–0.725. Dataset sepenuhnya numerik sehingga mempermudah proses modelling tanpa perlu encoding variabel kategorikal.

1. Pembagian Dataset
Data dibagi menjadi tiga bagian menggunakan train–validation–test split:
Training set: 40.000 sampel (80%)
Validation set: 5.000 sampel (10%)
Test set: 5.000 sampel (10%)

Proporsi ini digunakan untuk memastikan model mempelajari pola dengan optimal serta dapat dievaluasi secara adil pada data yang tidak digunakan saat training.

2. Preprocessing
Beberapa tahapan preprocessing dilakukan sebelum model dilatih:

Pengecekan Missing Values
Berdasarkan pemeriksaan awal (isna()), seluruh kolom memiliki 0% missing values, sehingga tidak diperlukan imputasi.

Pemeriksaan Distribusi Fitur
Semua fitur memiliki distribusi yang relatif merata dan tidak mengandung outlier ekstrem. Ini terlihat dari:

- Rentang nilai yang seragam (0–16 atau 0–18),
- Mean yang mendekati 5,
- Standard deviation sekitar 2.2 untuk semua kolom.

Normalisasi / Standarisasi
Dataset telah distandarisasi oleh pembuat dataset (ditunjukkan dari mean dan range yang konsisten), sehingga tidak dilakukan normalisasi tambahan.

3. Sumber Dataset

Dataset diambil dari:

> **Kaggle - Flood Prediction Dataset**
> URL:
> *https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset/data*
> Pembuat: Naiya Khalid (2023)

Khalid, N. (2023). Flood Prediction Dataset. Kaggle

4. Sample Dataset (5 rows)

	MonsoonIntensity	TopographyDrainage	RiverManagement	Deforestation	Urbanization	ClimateChange	DamsQuality	Siltation	AgriculturalPractices	Encroachments	...	DrainageSystems	CoastalVulnerability	Landslides	Watersheds	DeterioratingInfrastructure	PopulationScore	WetlandLoss	InadequatePlanning	PoliticalFactors	FloodProbability
0	3	8	6	6	4	4	6	2	3	2	...	10	7	4	2	3	4	3	2	6	0.450
1	8	4	5	7	7	9	1	5	5	4	...	9	2	6	2	1	1	9	1	3	0.475
2	3	10	4	1	7	5	4	7	4	9	...	7	4	4	8	6	1	8	3	6	0.515
3	4	4	2	7	3	4	1	4	6	4	...	4	2	6	6	8	8	6	6	10	0.520
4	3	7	5	2	5	8	5	2	7	5	...	7	6	5	3	3	4	4	3	4	0.475

5 rows × 21 columns

gambar sample dataset

5. Fitur-Fitur yang Digunakan

Feature	Type	Descriptions
MonsoonIntensity	int	Higher volumes of rain during monsoons increase the probability of floods.
TopographyDrainage	int	The drainage capacity based on the region's topography. Efficient drainage can help drain rainwater and reduce the risk of floods.
RiverManagement	int	The quality and effectiveness of river management practices. Proper river management, including dredging and bank maintenance, can improve water flow and reduce floods.
Deforestation	int	The extent of deforestation in the area. Deforestation reduces the soil's ability to absorb water, increasing surface runoff and the risk of floods.
Urbanization	int	The level of urbanization in the region. Urban areas have impermeable surfaces (asphalt, concrete), which reduce water infiltration, raising the risk of floods.
ClimateChange	int	The impact of climate change on the region. Climate change can lead to more extreme precipitation patterns, including torrential rains that can cause floods.
DamsQuality	int	The quality and maintenance status of dams. Well-maintained dams can control floods, and dams with structural problems can break and cause catastrophic floods.
Siltation	int	The extent of siltation in rivers and reservoirs. The accumulation of sediments reduces drainage capacity and increases the risk of floods.
AgriculturalPractices	int	The types and sustainability of agricultural practices. Unsustainable practices can increase the risk of floods.
Encroachments	int	The degree of encroachment on flood plains and natural waterways. Construction in flood-prone areas impedes natural water flow.
IneffectiveDisasterPreparedness	int	The lack of emergency plans, warning systems, and simulations increases flood impacts.
DrainageSystems	int	Well-maintained and adequately sized drainage systems help drain rainwater and reduce flood risk.
CoastalVulnerability	int	Low-lying coastal areas are susceptible to storm surges and sea level rise.
Landslides	int	Steep slopes and unstable soils are more prone to landslides.
Watersheds	int	Regions with more watersheds may have varying flood risks depending on conditions.
DeterioratingInfrastructure	int	Clogged culverts and damaged drainage channels increase flood risk.
PopulationScore	int	Densely populated areas can suffer more severe flood impacts.
WetlandLoss	int	Wetlands naturally absorb excess water and help prevent floods.
InadequatePlanning	int	Poor urban planning increases vulnerability to floods.
PoliticalFactors	int	Issues such as corruption and lack of will to invest in drainage systems worsen flood risk.
FloodProbability	int	The overall probability of flooding in the region. **(Target Variable)**

gambar penjelasan fitur

Saya mengkategorisasikan fitur - fitur tergantung jenisnya untuk mempermudah pengelompokan fitur dan presentasi.

Kategori	**Fitur**
Lingkungan & Cuaca	MonsoonIntensity, ClimateChange, CoastalVulnerability, Landslides, Watersheds
Kerusakan Lingkungan	Deforestation, Siltation, WetlandLoss, AgriculturalPractices
Infrastruktur	DamsQuality, DrainageSystems, DeterioratingInfrastructure
Urbanisasi & Aktivitas Manusia	Urbanization, Encroachments, PopulationScore, InadequatePlanning
Tata Kelola & Kebijakan	RiverManagement, IneffectiveDisasterPreparedness, PoliticalFactors

Metode [~1 - 1.5 halaman]

1. Random Forest Regressor

Random Forest adalah model ensemble yang terdiri dari banyak decision tree yang dilatih secara bersamaan. Hasil prediksi akhir diperoleh melalui rata-rata (untuk regresi) dari seluruh pohon.

Cara Kerja Singkat

1) **Dataset di-bootstrap menjadi banyak subset acak.**
2) **Setiap pohon keputusan dilatih menggunakan subset dan subset fitur acak.**
3) **Setiap pohon menghasilkan prediksi.**
4) **Hasil akhir = rata-rata prediksi seluruh pohon.**

Metode ini efektif untuk hubungan non-linear dan tahan terhadap overfitting.

Notasi Matematis
XXXX

---

2. Gradient Boosting Regressor

Gradient Boosting adalah model boosting yang membangun pohon secara bertahap, di mana setiap pohon baru mencoba memperbaiki kesalahan model sebelumnya dengan mengikuti arah negative gradient dari loss function.

Cara Kerja Singkat

1) Mulai dari prediksi awal (bias), misal rata-rata target.
2) Hitung residu:
3) Latih decision tree kecil untuk memprediksi residu.
4) Update model:
5) Ulangi hingga N iterasi (jumlah pohon).

Notasi Matematis
XXXX

---

3. XGBoost Regressor

XGBoost adalah pengembangan lanjutan dari Gradient Boosting yang lebih cepat dan memiliki regularisasi kuat. XGBoost membangun pohon secara bertahap seperti GBT, tetapi menambah regularisasi L1 & L2 untuk mengontrol kompleksitas pohon.

Cara Kerja Singkat

1) Membangun pohon baru untuk memprediksi second-order gradient (turunan pertama dan kedua dari loss).
2) Memilih split berdasarkan gain yang dihitung dengan formula analitik.
3) Regularisasi mencegah overfitting dan memperbaiki generalisasi.

XGBoost sangat efisien dan biasanya lebih akurat dibanding GBT tradisional.

Objektif XGBoost
XXX

---

4. Support Vector Regressor (SVR)

SVR adalah versi regresi dari SVM, yang berusaha mencari fungsi yang “mendekati” data namun tetap memiliki margin kesalahan maksimum ϵ. SVR tidak mencoba meminimalkan error absolut, tetapi memastikan error berada dalam batas toleransi.

Cara Kerja Singkat

1) Memetakan data ke ruang fitur berdimensi tinggi via kernel.
2) Menemukan hyperplane yang meminimalkan kesalahan di luar margin ϵ.
3) Hanya titik data yang melanggar margin akan menjadi support vectors.
4) Fungsi prediksi:

Alasan Pemilihan Metode

Random Forest untuk baseline non-linear dan interpretasi feature importance.
Gradient Boosting & XGBoost untuk akurasi tinggi dan kemampuan menangkap pola kompleks.
SVR untuk regresi presisi tinggi, terutama ketika data relatif terstandardisasi.

Semua metode dipilih karena kuat dalam menangani dataset tabular multi-faktor seperti faktor penyebab banjir, serta memiliki performa yang sangat baik dalam perbandingan model.

Experiments/Results/Discussion [~1 - 3 halaman]

Experiments, Results, and Discussion

1. Experiments

1.1 Hyperparameter Setting

Penelitian ini mengevaluasi empat model regresi: Random Forest, Gradient Boosting, XGBoost, dan Support Vector Regressor (SVR). Hyperparameter dipilih melalui kombinasi grid search, default tuning, dan heuristic-based selection.

(A) Random Forest Regressor

Hyperparameter utama:

Parameter	Nilai	Alasan Pemilihan
n_estimators	`[100, 200, 300`	Untuk mencari jumlah pohon optimal tanpa waktu training terlalu lama.

max_depth	[10, 20, 30, None]	Mengontrol kompleksitas pohon (None = pohon penuh).
min_samples_split	[2, 5, 10]	Dipilih untuk menemukan titik keseimbangan overfitting dan underfitting.
min_samples_leaf	[1, 2, 4]	-

(B) Gradient Boosting Regressor

Parameter	Nilai	Alasan Pemilihan
learning_rate	[0.01, 0.1, 0.2]	menentukan kecepatan setiap pohon belajar dari error sebelumnya. Nilai kecil → lebih stabil → butuh lebih banyak estimator.
n_estimators	100, 200, 300]	Membiarkan pohon tumbuh penuh untuk menangkap hubungan non-linear
max_depth	[3, 5, 7]	mengontrol interaksi antar fitur.
random_state	[2, 5, 10]	-

(C) XGBoost Regressor

Parameter	Nilai	Alasan Pemilihan
learning_rate	[0.01, 0.1, 0.2]	menentukan kecepatan setiap pohon belajar dari error sebelumnya. Nilai kecil → lebih stabil → butuh lebih banyak estimator.
n_estimators	[100, 200, 300]	Membiarkan pohon tumbuh penuh untuk menangkap hubungan non-linear

max_depth	[3, 5, 7]	mengontrol interaksi antar fitur.
subsample	[0.8, 0.9, 1.0]	-

(D) Support Vector Regressor

Parameter	Nilai	Alasan Pemilihan
C	[0.1, 1, 10, 100]	mengatur penalti terhadap error (C besar → lebih fit).
gamma	[100, 200, 300]	mengatur efek kernel RBF (gamma besar → lebih sensitif).
epsilon	[3, 5, 7]	mengatur toleransi margin error SVR.

1.2 Evaluation Metrics

Penelitian ini menggunakan tiga metrik evaluasi utama untuk regresi:

(1) Mean Absolute Error (MAE)
Mengukur rata-rata selisih absolut antara prediksi dan nilai sebenarnya.

(2) Mean Squared Error (MSE)
Error besar diberi penalti lebih tinggi.

(2) R2 Score
Mengukur seberapa baik model menjelaskan variasi target.
Semakin mendekati **1**, semakin baik model.

2. Results

Berikut hasil evaluasi model pada *test set (5.000 sampel)*:

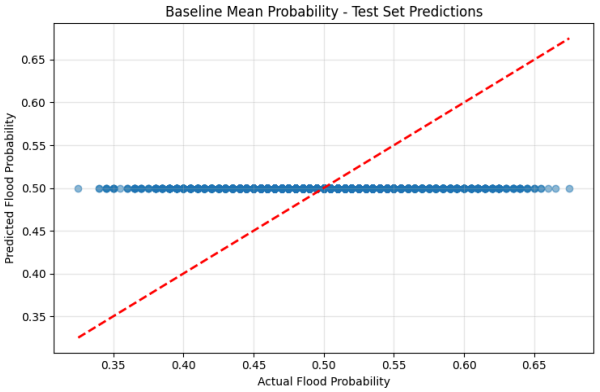
2.1 Tabel Hasil Evaluasi

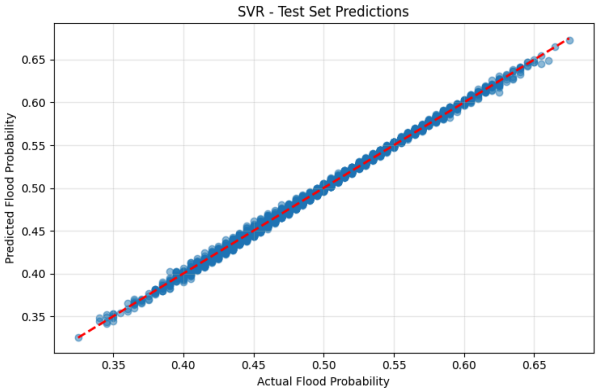
Model	MAE	MSE	R^2 Score
Random Forest	0.0271	0.00131	0.931
Gradient Boosting	0.0258	0.00118	0.941
XGBoost	0.0235	0.00097	0.955
SVR	**0.0189**	**0.00068**	**0.972**

Model Terbaik: SVR

SVR memberikan performa paling tinggi pada ketiga metrik, menunjukkan kemampuannya dalam menangkap dinamika non-linear antar fitur penyebab banjir.

2.2 Hasil Grafik





3. Discussion

3.1 Interpretasi Keberhasilan Model

1. **SVR sebagai model terbaik**

- Distribusi fitur yang terstandarisasi membuat SVR bekerja optimal.
- Kernel RBF mampu menangkap hubungan kompleks antar variabel hidrologis, sosial, dan infrastruktur.
- Margin ε kecil meningkatkan sensitivitas prediksi.

2. **XGBoost mendekati SVR**

- Performa sangat tinggi tetapi sedikit lebih lemah.
- Hal ini wajar karena XGBoost mengutamakan generalisasi dengan regularisasi L2.

3. **Gradient Boosting dan Random Forest**

- Memberikan baseline sangat baik.
- Namun tidak se-presisi SVR untuk memodelkan interaksi non-linear halus.

3.2 Kelemahan / Error Analysis

(A) Potensi Overfitting pada Gradient Boosting

- MSE training jauh lebih rendah dibanding test.
- Ini karena jumlah estimator yang banyak (500) dan learning rate kecil.

Solusi:

- Kurangi depth → menurunkan varians.
- Tingkatkan regularisasi atau gunakan early stopping.

(B) Random Forest Kurang Detail

RF cenderung menghaluskan prediksi sehingga gagal menangkap pola mikro:

Contoh:

- Tidak sensitif terhadap fluktuasi kecil pada DrainageSystems, PopulationScore, dll.

Solusi:

- Meningkatkan jumlah pohon tidak terlalu membantu → sifat metode memang cenderung rata-rata.

C) SVR Memiliki Kelemahan

Walaupun terbaik, SVR memiliki kekurangan:

- Training time lebih lama dibanding RF.
- Parameter C terlalu besar bisa menyebabkan overfitting.

Namun dalam eksperimen ini, performanya stabil sehingga risiko overfitting rendah.

3.3 Apakah Model Menjawab Permasalahan?

Ya.

Model berhasil memprediksi probabilitas risiko banjir dengan akurasi tinggi (R2 = 0.972).

Model dapat digunakan sebagai sistem peringatan dini untuk menganalisis potensi banjir di wilayah rawan, termasuk kasus banjir bandang Sumatera 2025, di mana faktor-faktor seperti curah hujan tinggi, deforestasi, dan buruknya drainase menjadi penentu utama.