

## PEMODELAN DAN PERBANDINGAN ALGORITMA MACHINE LEARNING UNTUK PREDIKSI BANJIR: STUDI KASUS KOTA MEDAN

**Putri Maulidina Fadilah**

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan  
[putrimaulidina@unimed.ac.id](mailto:putrimaulidina@unimed.ac.id) \*

**Putri Harliana**

Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan  
[harliana@unimed.ac.id](mailto:harliana@unimed.ac.id)

**Imelda Wardani Br Rambe**

Program Studi Pendidikan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan  
[imeldawardanibrambe@unimed.ac.id](mailto:imeldawardanibrambe@unimed.ac.id)

**Alvi Sahrin Nasution**

Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan  
[alvisahrin@unimed.ac.id](mailto:alvisahrin@unimed.ac.id)

### Abstrak

Banjir merupakan salah satu bencana hidrometeorologis yang sering terjadi di Indonesia, termasuk di Kota Medan, dan menimbulkan kerugian sosial ekonomi yang signifikan. Penelitian ini bertujuan untuk mengembangkan dan membandingkan model prediksi kejadian banjir menggunakan tiga algoritma *machine learning*, yaitu Random Forest (RF), Support Vector Machine (SVM), dan Decision Tree (CART). Data yang digunakan merupakan data hidrometeorologis Kota Medan periode 2021–2024 yang diperoleh dari BMKG dan Geoportal BNPB, meliputi variabel curah hujan (RR), suhu (TN, TX, TAVG), kelembapan (RH\_AVG), kecepatan angin (FF\_X), arah angin (DDD\_X), serta jumlah kecamatan dan rumah terdampak. Tahapan penelitian meliputi pra-pemrosesan data, pemodelan, serta evaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan AUC. Hasil penelitian menunjukkan bahwa Random Forest memiliki kinerja terbaik dengan akurasi sebesar 86,24% dan AUC sebesar 0,851, yang menunjukkan kemampuan prediksi yang sangat baik. Analisis *feature importance* menunjukkan bahwa curah hujan (RR) merupakan faktor paling berpengaruh terhadap kejadian banjir, diikuti oleh temperatur minimum (TN) dan temperatur rata-rata (TAVG). Metode ROSE terbukti efektif dalam mengatasi ketidakseimbangan kelas dengan meningkatkan *recall* tanpa mengorbankan akurasi. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa algoritma Random Forest lebih andal dan stabil dibandingkan SVM dan Decision Tree dalam konteks analisis data hidrometeorologis, serta berpotensi menjadi dasar pengembangan sistem peringatan dini banjir berbasis kecerdasan buatan di masa depan.

**Kata Kunci:** Prediksi Banjir, Random Forest, Support Vector Machine, Decision Tree, ROSE.

### Abstract

Flooding is one of the most frequent hydrometeorological disasters in Indonesia, including in Medan City, causing significant socioeconomic losses. This study aims to develop and compare flood prediction models using three machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (CART). The research uses hydrometeorological data from BMKG and BNPB Geoportal covering the period 2021–2024, which includes rainfall (RR), temperature (TN, TX, TAVG), humidity (RH\_AVG), wind speed (FF\_X), wind direction (DDD\_X), and the number of affected districts and houses. The research stages include data preprocessing, model development, and evaluation using accuracy, precision, recall, F1-score, and AUC metrics. The results show that the Random Forest model achieved the best performance, with an accuracy of 86.24% and an AUC of 0.851, indicating excellent predictive ability. The feature importance analysis revealed that rainfall (RR) was the most influential factor in flood occurrence, followed by minimum temperature (TN) and average temperature (TAVG). The ROSE method effectively handled class imbalance, improving recall without significantly reducing accuracy. Overall, the findings demonstrate that the Random Forest algorithm is more reliable and stable than SVM and Decision Tree in hydrometeorological data analysis and can serve as a foundation for developing AI-based early warning systems for flood risk mitigation in Medan City.

**Keywords:** Flood Prediction, Random Forest, Support Vector Machine, Decision Tree, ROSE.

## PENDAHULUAN

Banjir merupakan salah satu jenis bencana alam yang memiliki frekuensi tinggi di Indonesia, termasuk di wilayah Kota Medan. Peristiwa ini sering menimbulkan dampak ekonomi dan sosial yang cukup besar. Berdasarkan data nasional, sekitar 76% dari total kejadian bencana di Indonesia dikategorikan sebagai bencana hidrometeorologis, di mana banjir menjadi salah satu yang paling dominan (Sekolah Tinggi Meteorologi Klimatologi dan Geofisika et al., 2022). Fenomena tersebut menimbulkan kerugian besar, seperti kerusakan infrastruktur, penurunan produktivitas masyarakat, serta terganggunya aktivitas sosial ekonomi penduduk. Kondisi topografi yang rendah, intensitas curah hujan tinggi, serta sistem drainase yang belum optimal menjadi faktor utama pemicu terjadinya banjir di wilayah perkotaan. Oleh karena itu, dibutuhkan suatu sistem yang mampu melakukan prediksi dan pemantauan banjir secara lebih efektif dan berbasis data.

Dalam konteks mitigasi bencana, penerapan teknologi *machine learning* telah memberikan hasil yang menjanjikan (Roihan et al., 2020). Pendekatan ini mampu meningkatkan akurasi dalam memprediksi potensi banjir serta mengurangi risiko kerugian yang ditimbulkan (Sandiwarno, 2024). Pemodelan prediksi banjir menggunakan algoritma *machine learning* telah menjadi fokus utama bagi para peneliti untuk meningkatkan akurasi prediksi dan manajemen risiko terkait. Di antara berbagai algoritma yang tersedia, Support Vector Machine (SVM), Decision Tree (DT), dan Random Forest (RF) adalah metode yang paling umum digunakan. Setiap metode memiliki kelebihan dan kelemahan spesifik yang dapat mempengaruhi keakuratan serta efektivitasnya dalam memprediksi banjir. Masing-masing algoritma memiliki karakteristik dan aplikabilitas yang berbeda dalam konteks model prediksi banjir. Random Forest umumnya menawarkan kinerja yang lebih baik dalam hal akurasi dan stabilitas, sementara SVM dan Decision Tree memiliki tempatnya masing-masing tergantung pada sifat data dan konteks spesifik dari studi yang dilakukan.

Penelitian oleh Nayak et al. menunjukkan bahwa Random Forest memiliki tingkat akurasi yang lebih tinggi ketika dibandingkan dengan SVM dan

Decision Tree dalam konteks pemodelan yang lebih kompleks (Nayak et al., 2021). Studi oleh Ssebyala et al. juga memberikan wawasan tentang pengelolaan data tidak seimbang dalam prediksi risiko kesehatan terkait cuaca ekstrem, di mana Random Forest terbukti lebih efektif dibandingkan beberapa teknik lainnya (Ssebyala et al., 2024), meskipun tidak secara langsung berkaitan dengan prediksi banjir.

Di sisi lain, SVM memiliki kekuatan tersendiri, terutama ketika dihadapkan pada data dengan dimensi tinggi dan tidak linier (Hassan & Saeed, 2024). Meskipun demikian, penelitian oleh Hidayaturohman et al. menunjukkan bahwa SVM bisa memiliki kinerja yang kurang optimal jika data yang digunakan tidak cukup representatif, yang dapat memengaruhi kapasitas generalisasi dibandingkan dengan Random Forest (Hidayaturohman et al., 2023).

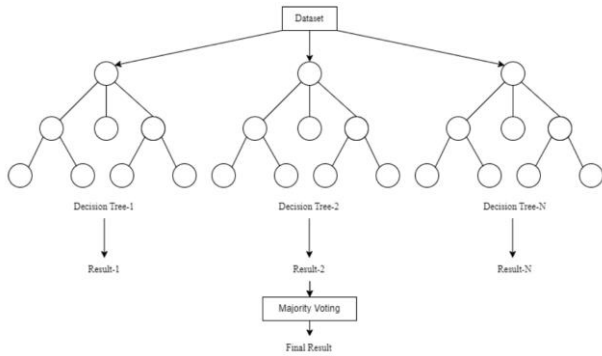
Sementara itu, Decision Tree menawarkan kemudahan dalam interpretasi karena struktur pohonnya yang jelas dan algoritma yang relatif sederhana. Namun, meskipun mudah digenerate dan dipahami, Decision Tree cenderung rawan mengalami overfitting, terutama ketika dataset yang digunakan cukup kecil atau memiliki noise yang tinggi. Penelitian oleh Ibrahim dan Abdulazeez menunjukkan bahwa keakuratan model Decision Tree dapat tergerus oleh model Random Forest yang lebih kompleks dan robust (Ibrahim & Abdulazeez, 2021). Penelitian lebih lanjut diperlukan untuk terus mengembangkan teknik-teknik ini, dengan fokus pada peningkatan algoritma dan efisiensi data yang lebih baik agar lebih terbukti efektif dalam aplikasi praktis. Dengan mengetahui metode mana yang paling efektif dalam memprediksi banjir, diharapkan dapat meningkatkan kesiapsiagaan masyarakat dan pemerintah daerah dalam upaya mitigasi dan pengurangan risiko bencana banjir.

## KAJIAN TEORI

### RANDOM FOREST

Algoritma Random Forest (RF) merupakan metode *ensemble learning* yang menggabungkan beberapa *Decision Tree* untuk meningkatkan akurasi dan stabilitas prediksi. Setiap pohon (*tree*) dibangun dari subset data yang berbeda hasil dari proses *bootstrap sampling* dan subset acak dari fitur yang digunakan pada setiap pemisahan (*split*).

Ilustrasi algoritma Random Forest dapat dilihat pada Gambar 1 (Purwati & Pristyanto, 2024).



Gambar 1. Ilustrasi Algoritma Random Forest

Tahapan utama algoritma Random Forest dapat diformulasikan sebagai berikut: (1) Pengambilan sampel bootstrap; (2) Pembentukan pohon keputusan; (3) Agregasi hasil prediksi.

Diberikan dataset  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ , dibentuk  $B$  sampel bootstrap  $D_1, D_2, \dots, D_B$  dengan pengambilan sampel secara acak dengan pengembalian.

$$D_b = \text{Bootstrap}(D), \quad b=1,2,\dots,B \tag{1}$$

Setiap  $D_b$  digunakan untuk membangun satu *Decision Tree*.

Pada setiap node, dipilih  $m$  variabel acak ( $m < p$ ) dari total  $p$  prediktor. Kemudian, ditentukan pemisahan terbaik berdasarkan nilai impurity minimum dengan menggunakan indeks Gini atau Entropy.

Indeks Gini:

$$G(t) = 1 - \sum_k (p_k)^2 \tag{2}$$

Entropy:

$$H(t) = - \sum_k (p_k \log_2 p_k) \tag{3}$$

di mana  $p_k$  adalah proporsi observasi kelas  $k$  pada node  $t$ , dan node dengan *impurity* minimum dipilih sebagai pemisahan terbaik.

Setelah semua pohon terbentuk, hasil prediksi digabungkan dengan metode voting mayoritas (klasifikasi):

$$\hat{y} = \text{mode}\{\hat{y}_1(x), \hat{y}_2(x), \dots, \hat{y}_B(x)\} \tag{4}$$

atau rata-rata (regresi):

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B \hat{y}_b(x) \tag{5}$$

**SUPPORT VECTOR MACHINE**

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang mencari hyperplane optimal yang memisahkan data dari dua kelas dengan margin maksimum (Purwati & Pristyanto,

2024). Diberikan himpunan pelatihan  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  dengan  $y_i \in \{-1, +1\}$ , formulasi optimasi SVM dinyatakan sebagai:

$$\text{Minimalkan: } \frac{1}{2} \|w\|^2 + C \sum \xi_i \tag{6}$$

dengan kendala:  $y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$  di mana  $w$  adalah vektor bobot,  $b$  adalah bias,  $C$  adalah parameter regularisasi, dan  $\xi_i$  adalah variabel toleransi.

Fungsi keputusan didefinisikan sebagai:

$$f(x) = \text{sign}(w^T x + b) \tag{7}$$

**DECISION TREE**

Decision Tree adalah metode pembelajaran terawasi (supervised learning) yang secara rekursif membagi dataset menjadi subset berdasarkan atribut paling informatif, membentuk struktur seperti pohon keputusan (Ibrahim & Abdulazeez, 2021). Kriteria pemisahan bertujuan untuk meminimalkan impurity di setiap node:

Information Gain =

$$\text{Entropy}(\text{Parent}) - \sum (|D_i| / |D|) * \text{Entropy}(D_i) \tag{8}$$

Entropy(D) =

$$- \sum_k (p_k \log_2 p_k) \tag{9}$$

di mana  $p_k$  adalah proporsi kelas  $k$  dalam dataset  $D$ . Proses pemisahan berlanjut hingga kondisi penghentian tercapai, misalnya kedalaman maksimum atau jumlah minimum sampel per node.

**RANDOM OVER-SAMPLING EXAMPLES (ROSE)**

Ketidakeimbangan kelas (*class imbalance*) merupakan permasalahan umum dalam analisis data bencana, termasuk prediksi kejadian banjir, di mana jumlah data kelas minoritas (kejadian banjir) jauh lebih sedikit dibandingkan dengan kelas mayoritas (tidak banjir). Ketidakeimbangan ini dapat menyebabkan model *machine learning* menjadi bias terhadap kelas mayoritas, sehingga menurunkan kemampuan model dalam mengenali kejadian banjir yang sesungguhnya (minoritas).

Untuk mengatasi masalah tersebut, penelitian ini menggunakan pendekatan ROSE (Random Over-Sampling Examples) yang dikembangkan oleh Lunardon, Menardi, dan Torelli. Metode ini bekerja dengan menghasilkan *synthetic balanced samples* melalui kombinasi proses *oversampling* pada kelas minoritas dan *undersampling* pada kelas mayoritas (Lunardon et al., 2014).

Secara matematis, prinsip ROSE dapat dijelaskan sebagai berikut:

$$D^* = \{(x_i^*, y_i^*)\}_{i=1}^{N^*} \tag{10}$$

dengan  $x_i^* \sim f(x | y), y_i^* \in \{0,1\}$ .

di mana  $f(x | y)$  adalah fungsi densitas bersyarat dari data fitur terhadap label kelas,  $D^*$  adalah himpunan data hasil resampling yang memiliki proporsi kelas lebih seimbang,  $N^*$  adalah ukuran total dataset baru setelah proses resampling.

**EVALUASI MODEL PREDIKSI**

Evaluasi model dilakukan untuk mengukur seberapa baik algoritma *machine learning* (Random Forest, Decision Tree, dan SVM) dalam mengklasifikasikan kejadian banjir dan tidak banjir. Penilaian performa dilakukan dengan menggunakan data uji (testing data) yang tidak digunakan dalam proses pelatihan model.

Ukuran performa utama dalam penelitian ini meliputi Akurasi (Accuracy), Presisi (Precision), Recall (Sensitivitas), Spesifisitas (Specificity), F1-Score, dan Area Under Curve (AUC) dari kurva Receiver Operating Characteristic (ROC).

Hasil klasifikasi dinilai berdasarkan *confusion matrix*, yang menggambarkan hubungan antara label sebenarnya dan prediksi model. Tabel 1 berikut menunjukkan komponennya (Widjiyati, 2021):

Tabel 1. Confusion Matrix

Prediksi / Aktual	Positif (1)	Negatif (0)
Positif (1)	True Positive (TP)	False Positive (FP)
Negatif (0)	False Negative (FN)	True Negative (TN)

Selanjutnya, akurasi digunakan untuk mengukur proporsi prediksi benar terhadap total data (Purwati & Pristyanto, 2024; Widjiyati, 2021):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

Presisi untuk mengukur ketepatan model dalam memprediksi kelas positif (Azimah & Rizky Nova Wardani, 2022):

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

Recall atau sensitivitas untuk mengukur kemampuan model dalam mengenali kejadian positif:

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

Spesifisitas untuk mengukur kemampuan model dalam mengenali kejadian negatif:

$$Specificity = \frac{TN}{TN+FP} \tag{14}$$

F1-Score merupakan rata-rata harmonik antara Precision dan Recall (Handayani & Fauzan, 2024):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

Kemudian, Area Under Curve (AUC) untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif berdasarkan kurva ROC. Semakin besar nilai AUC (mendekati 1), semakin baik performa klasifikasi. ROC curve menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR), dengan rumus:

$$TPR = \frac{TP}{TP+FN} \tag{16}$$

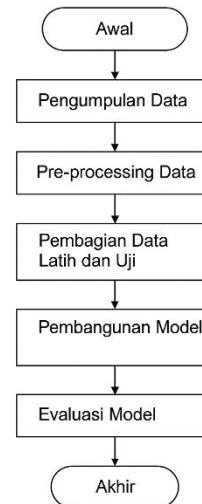
$$FPR = \frac{FP}{FP+TN} \tag{17}$$

Nilai AUC dihitung sebagai integral di bawah kurva ROC:

$$AUC = \int_0^1 TPR(FPR) d(FPR) \tag{18}$$

**METODE**

Penelitian ini menggunakan pendekatan kuantitatif dengan fokus pada pengembangan model prediksi banjir berbasis *Random Forest* dan perbandingannya dengan SVM dan Decision Tree. Penelitian ini dirancang dalam tiga tahap utama, yaitu: (1) Pengumpulan dan pra-pemrosesan data (data preprocessing); (2) Pembangunan model; (3) Evaluasi model prediksi banjir. Pendekatan *comparative modeling* digunakan untuk membandingkan kinerja ketiga algoritma berdasarkan metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan AUC. Model dengan performa terbaik akan digunakan sebagai dasar pengembangan prediksi banjir Kota Medan.



Gambar 2. Diagram Alir penelitian

Populasi dalam penelitian ini adalah seluruh kejadian banjir yang tercatat di Kota Medan dalam rentang waktu tahun 2021–2024, sebagaimana dilaporkan oleh Badan Nasional Penanggulangan Bencana (BNPB) dan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG).

Sampel penelitian diambil berdasarkan ketersediaan data hidrometeorologis harian yang meliputi variabel Curah hujan harian (RR, dalam mm), Suhu minimum (TN, °C), Suhu maksimum (TX, °C), Suhu rata-rata (TAVG, °C), Kelembapan rata-rata (RH\_AVG, %), Kecepatan angin maksimum (FF\_X, m/s), Arah angin saat kecepatan maksimum (DDD\_X), Jumlah rumah terdampak (RUMAH\_TERENDAM), Jumlah kecamatan terdampak (KEJ\_BANJIR\_KEC), dan Tanggal kejadian (TANGGAL). Variabel dependen atau target model adalah kejadian banjir (BANJIR) yang bersifat biner, dengan label 1 untuk “terjadi banjir” dan 0 untuk “tidak terjadi banjir.”

Data penelitian diperoleh melalui pendekatan dokumentasi dan observasi sekunder dari dua sumber utama, yaitu: (1) Geoportal BNPB yang menyediakan data spasial dan temporal kejadian banjir di tingkat kecamatan, termasuk jumlah kejadian dan wilayah terdampak; (2) BMKG Medan yang menyediakan data parameter cuaca harian seperti curah hujan, suhu, kelembapan, dan kecepatan angin.

Proses pengumpulan data dilakukan dengan tahapan sebagai berikut: (1) Mengunduh dan menggabungkan dataset dari kedua instansi; (2) Melakukan *data cleaning*; (3) Menentukan variabel target dengan kriteria BANJIR = 1 jika KEJ\_BANJIR\_KEC > 0, dan 0 jika tidak.

Tabel 2. Data Penelitian

Tgl	Kej_Banjir_Kec	TN	TX	TAVG	...	FF_AVG	DDDCAR
01-01-2024	0	25	32,4	29	...	1	C
02-01-2024	0	24,8	31,7	29,6	...	2	W
03-01-2024	0	25,4	30	28,5	...	1	W
...	...	...	...	...	...	...	...
29-12-2024	0	24	33	27,8	...	1	C

30-12-2024	0	24,1	30,3	26,2	...	1	C
31-12-2024	0	24,4	30,4	26,7	...	1	W

Analisis dilakukan menggunakan perangkat lunak RStudio untuk pemodelan dan analisis statistik, dengan pustaka utama: randomForest, e1071, rpart, caret, ROSE, pROC, dan dplyr. Kemudian data mentah yang telah dikumpulkan, dibersihkan dari nilai yang tidak valid (misalnya nilai “8888” atau “-”) dan diisi menggunakan metode *median imputation*. Data selanjutnya dibagi menjadi dua subset menggunakan metode *stratified sampling* dengan proporsi 70% data latih (training) dan 30% data uji (testing) menggunakan fungsi `createDataPartition()` dari paket `caret`.

Untuk mengatasi ketidakseimbangan kelas antara data banjir dan tidak banjir, digunakan metode ROSE (Random Over-Sampling Examples) yang menghasilkan data sintetis baru dari kelas minoritas tanpa menyebabkan *overfitting*. Selanjutnya, pemodelan dan pelatihan masing-masing metode yaitu Random Forest, SVM, dan Decision Tree menggunakan data latih dan evaluasi model optimal menggunakan data uji. Kinerja model diuji menggunakan confusion matrix dan metrik statistik untuk selanjutnya dilakukan perbandingan.

## HASIL DAN PEMBAHASAN

### PRA-PEMROSESAN DATA

Tahap pra-pemrosesan menghasilkan data yang akan digunakan dalam keadaan sudah bersih dan bebas dari data hilang. Hal ini menandakan bahwa tidak ada missing value yang dapat memengaruhi performa model. Kemudian permasalahan kelas tidak seimbang juga telah teratasi. Sebelum dilakukan balancing, distribusi kelas sangat tidak seimbang – terdapat 247 data non-banjir (kelas 0) dan hanya 10 data banjir (kelas 1). Setelah dilakukan balancing menggunakan metode ROSE (Random Over-Sampling Examples), distribusi menjadi lebih seimbang dengan 130 data non-banjir dan 127 data banjir. Ini memastikan bahwa model tidak bias terhadap kelas mayoritas.

**PEMODELAN**

Tahap pemodelan dilakukan dengan melatih ketiga model dengan parameter default terlebih dahulu, lalu dilakukan penyetelan jumlah pohon keputusan (ntree) dan jumlah variabel yang dipilih di setiap percabangan (mtry) pada Random Forest, hyperparameter tuning pada SVM dan tuning cp pada Decision Tree agar memperoleh hasil yang optimal. Hasil pelatihan model disimpan dalam objek klasifikasi yang mencakup informasi penting seperti error rate, jumlah pohon, serta matriks klasifikasi. Setelah model dilatih, dilakukan evaluasi terhadap performa model menggunakan data uji.

Selain evaluasi performa, dilakukan pula analisis feature importance untuk mengetahui variabel-variabel yang paling berpengaruh dalam model. Hasilnya menunjukkan bahwa variabel RR (curah hujan harian) merupakan prediktor paling dominan terhadap kejadian banjir, diikuti oleh suhu minimum (TN) dan suhu rata-rata (TAVG).

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
tn	15.016246	28.867565	30.484779	26.427347
tx	-1.905133	6.535235	3.579349	6.394985
tavg	6.444888	21.687381	20.434914	18.087364
rh_avg	5.245003	17.656278	17.173067	13.988081
rr	29.772254	38.836521	43.371073	35.664704
ss	-3.273066	5.487538	1.708976	5.858727
ff_x	5.508172	5.909752	7.905867	7.993890
ddd_x	3.724935	10.479366	10.117111	7.853684
ff_avg	1.524388	6.812579	5.709570	5.752512

Gambar 3. Feature Importance

Berdasarkan hasil feature importance, kontribusi setiap variabel dapat dilihat. Selanjutnya dilakukan reduksi dimensi untuk menghilangkan variabel yang tidak signifikan. Beberapa variabel yang dihilangkan antara lain variabel tx, ss, ff\_avg. Sedangkan variabel lainnya tetap digunakan dalam model. Hal ini dilakukan untuk meningkatkan efisiensi model, serta mengurangi risiko *overfitting* tanpa menurunkan performa prediksi.

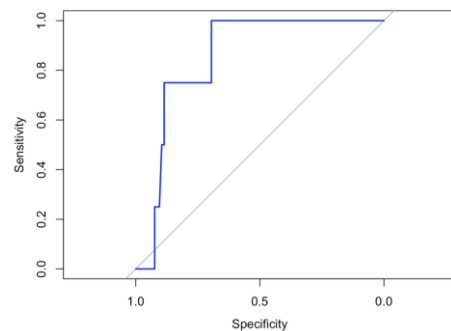
**EVALUASI MODEL**

Hasil akhir pemodelan ditunjukkan pada Tabel 3.

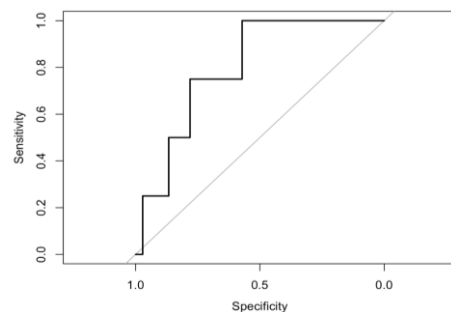
Tabel 3. Evaluasi Hasil Prediksi Model

Model	Akurasi	Recall	Precision	F1-Score	AUC
Random Forest	<b>0.8624</b>	<b>0.75</b>	0.176	0.286	<b>0.851</b>
SVM (Radial)	0.853	0.50	0.125	0.200	0.798
Decision Tree (CART)	0.780	0.50	0.083	0.143	0.636

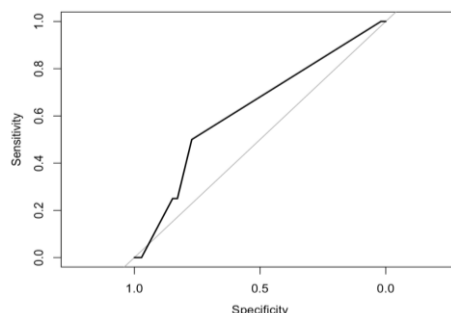
Hasil evaluasi yang didapatkan menunjukkan bahwa Random Forest memiliki akurasi tertinggi (86,24%) dan nilai AUC sebesar 0,851, menandakan kemampuan model dalam membedakan antara kejadian banjir dan non-banjir sangat baik. Sementara itu, model SVM menghasilkan performa cukup baik dengan AUC sebesar 0,798, namun presisi yang rendah mengindikasikan banyak prediksi positif palsu (false positives). Adapun model Decision Tree memiliki performa paling rendah dengan AUC 0,636, yang menunjukkan kemampuan diskriminatifnya terbatas.



Gambar 4. Kurva ROC Random Forest



Gambar 5. Kurva ROC SVM



Gambar 6. Kurva ROC Decision Tree

Kurva ROC menunjukkan bahwa garis ROC untuk model Random Forest yang ditunjukkan oleh Gambar 4 menjauhi diagonal dan mendekati sumbu kiri atas, yang berarti performanya sangat baik. Model SVM yang diperlihatkan pada Gambar 5

juga memperlihatkan pola kurva yang baik namun tidak setajam Random Forest, sedangkan Decision Tree pada Gambar 6 mendekati diagonal, menandakan model tersebut kurang optimal dalam klasifikasi.

Hasil ini juga memperkuat temuan sebelumnya bahwa algoritma ensemble seperti Random Forest lebih unggul dibandingkan algoritma tunggal seperti Decision Tree dan SVM, terutama ketika dihadapkan pada data yang tidak seimbang dan bersifat non-linear. Random Forest memanfaatkan konsep voting dari banyak decision tree, sehingga menghasilkan prediksi yang lebih stabil dan tahan terhadap overfitting. Selain itu, penerapan ROSE balancing secara signifikan meningkatkan sensitivitas model terhadap kelas minoritas (kejadian banjir), tanpa menurunkan akurasi keseluruhan secara drastis.

#### UCAPAN TERIMA KASIH

Terima kasih kepada LPPM Universitas Negeri Medan yang telah membantu dalam pendanaan penelitian ini. Pendanaan ini sangat berarti bagi peneliti untuk kelancaran pengembangan penelitian.

#### PENUTUP

#### SIMPULAN

Berdasarkan hasil analisis yang telah dilakukan, penelitian ini menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dibandingkan dengan Support Vector Machine (SVM) dan Decision Tree dalam memprediksi kejadian banjir di Kota Medan. Model Random Forest berhasil mencapai tingkat akurasi sebesar 86,24% dengan Area Under the Curve (AUC) sebesar 0,851, yang mengindikasikan kemampuan klasifikasi yang sangat baik dalam membedakan antara kondisi banjir dan non-banjir.

Analisis *variable importance* juga menunjukkan bahwa curah hujan (RR) merupakan faktor paling dominan yang memengaruhi kejadian banjir, diikuti oleh temperatur minimum (TN) dan temperatur rata-rata (TAVG). Temuan ini sejalan dengan teori hidrometeorologi yang menegaskan bahwa intensitas curah hujan merupakan variabel kunci dalam pembentukan risiko banjir di wilayah tropis seperti Medan.

Selain itu, penerapan metode ROSE (Random Over-Sampling Examples) terbukti efektif dalam mengatasi masalah ketidakseimbangan kelas (imbalance class), yang secara signifikan meningkatkan nilai *recall* tanpa menurunkan akurasi model secara drastis.

Secara keseluruhan, algoritma Random Forest terbukti lebih andal dan stabil dalam konteks analisis data hidrometeorologis Kota Medan, sehingga dapat dijadikan dasar untuk pengembangan sistem pemantauan dan peringatan dini banjir berbasis kecerdasan buatan yang lebih akurat dan adaptif di masa mendatang.

#### SARAN

Saran untuk penelitian selanjutnya sebaiknya dapat menambahkan variabel spasial seperti ketinggian wilayah, jarak ke sungai, dan jenis penggunaan lahan untuk meningkatkan akurasi model. Sejalan dengan meningkatnya akurasi juga dapat dilakukan integrasi model dengan sistem dashboard pemantauan interaktif berbasis real-time (misalnya dengan data IoT atau BMKG API) dapat memperluas manfaat praktis bagi pemerintah daerah dan masyarakat.

Validasi eksternal juga perlu dilakukan dengan dataset tahun berikutnya untuk menilai konsistensi performa model terhadap perubahan iklim dan kondisi lingkungan.

#### DAFTAR PUSTAKA

- Azimah, F., & Rizky Nova Wardani, K. (2022). Sistem Pendeteksi Gejala Awal Covid-19 dengan Penggunaan Metode AI Project Cycle. *Journal Locus Penelitian Dan Pengabdian*, 1(6), 405–418. <https://doi.org/10.36418/locus.v1i6.135>
- Handayani, P., & Fauzan, A. C. (2024). Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random. *Klik : Kajian Ilmiah Informatika Dan Komputer*, 4(6), 3064–3072. <https://doi.org/10.30865/klik.v4i6.1909>
- Hassan, S. A., & Saeed, M. S. (2024). A Comparative Study Evaluated the Performance of Two-class Classification Algorithms in Machine Learning. *Kurdistan Journal of Applied Research*, 8(2), 43–51. <https://doi.org/10.24017/science.2023.2.5>

- Hidayaturrohman, Q. A., Clarke, H. G., Taflan, G. Y., & Sancaktar, I. (2023). *A comparative study of machine learning approaches to heart disease prediction: An empirical analysis*. In Review. <https://doi.org/10.21203/rs.3.rs-3098962/v1>
- Ibrahim, I., & Abdulazeez, A. (2021). The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, 2(01), 10–19. <https://doi.org/10.38094/jastt20179>
- Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, 6(1), 79. <https://doi.org/10.32614/RJ-2014-008>
- Nayak, S. K., Beura, M., Siddique, M., & Mishra, S. P. (2021). Analysis of Indian Food Based on Machine learning Classification Models. *Journal of Scientific Research and Reports*, 1–7. <https://doi.org/10.9734/jsrr/2021/v27i730407>
- Purwati, S. E., & Pristyanto, Y. (2024). Model Random Forest and Support Vector Machine for Flood Classification in Indonesia. *Sinkron*, 8(4), 2261–2268. <https://doi.org/10.33395/sinkron.v8i4.13973>
- Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, 5(1). <https://doi.org/10.31294/ijcit.v5i1.7951>
- Sandiwarno, S. (2024). Penerapan Machine Learning Untuk Prediksi Bencana Banjir. *Jurnal Sistem Informasi Bisnis*, 14(1), 62–76. <https://doi.org/10.21456/vol14iss1pp62-76>
- Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Rachmawardani, A., Wijaya, S. K., Universitas Indonesia, Shopaheluwakan, A., & Badan Meteorologi Klimatologi dan Geofisika. (2022). SISTEM PERINGATAN DINI BANJIR BERBASIS MACHINE LEARNING: STUDI LITERATUR. *METHOMIKA Jurnal Manajemen Informatika Dan Komputerisasi Akuntansi*, 6(6), 188–198. <https://doi.org/10.46880/jmika.Vol6No2.p188-198>
- Ssebyala, S. N., Kintu, T. M., Muganzi, D. J., Dresser, C., Demetres, M. R., Lai, Y., Mercy, K., Li, C., Wang, F., Setoguchi, S., Celi, L. A., & Ghosh, A. K. (2024). Use of machine learning tools to predict health risks from climate-sensitive extreme weather events: A scoping review. *PLOS Climate*, 3(1), e0000338. <https://doi.org/10.1371/journal.pclm.0000338>
- Widjiyati, N. (2021). Implementasi Algoritme Random Forest Pada Klasifikasi Dataset Credit Approval. *Jurnal Janitra Informatika Dan Sistem Informasi*, 1(1), 1–7. <https://doi.org/10.25008/janitra.v1i1.118>