Jurnal Teknologi Sistem Informasi dan Aplikasi **ISSN: 2654-3788**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2654-4229**
Vol. 7, No. 3, Juli 2024 (1267-1279) **DOI: 10.32493/jtsi.v7i3.41632**

Model Optimasi Random Forest dengan PSO-CHI-SM dalam Mengatasi High Dimensional dan Imbalanced Data Banjir Kota Samarinda

Ilham Taufiq[1], Taghfirul Azhima Yoga Siswa*[2], Wawan Joko Pranoto[3]

Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia, 75124
e-mail: [1]2011102441152@umkt.ac.id, [*2]tay758@umkt.ac.id, [3]wjp337@umkt.ac.id

Submitted Date: July 03rd, 2024 Reviewed Date: July 18th, 2024
Revised Date: July 20th, 2024 Accepted Date: July 24th, 2024

Abstract

Flooding is a natural disaster that frequently affects our country. Samarinda City, in particular, continues to experience frequent flooding events with 18 incidents in 2018, 33 incidents in 2020, and 32 incidents in 2021. To predict flood disasters, it is necessary to utilize technology known as machine learning for analyzing and classifying floods. However, classification often encounters issues with high-dimensional data and class imbalance. This study aims to determine the extent to which the accuracy of flood disaster classification improves by using the Random Forest algorithm with PSO for optimization, Chi-Square feature selection, and SMOTE oversampling to balance classes. The data used in this study comprises flood data from 2021-2023 obtained from BMKG and BPBD Samarinda City, with a total of 1095 records and 11 attributes. The validation technique used is 5-fold cross-validation, and the evaluation uses a confusion matrix. The results of the Chi-Square feature selection identified Rainfall, Maximum Wind Direction, Most Frequent Wind Direction, Humidity, Sunshine Duration, and Wind Speed as the most influential features based on Chi-Square scores and P-values. The average accuracy obtained from the proposed classification model using 5-fold cross-validation reached 96.02%.

Keywords: Classification; Flood; Random Forest; Imbalance; Chi-Square; Optimization

Abstrak

Banjir merupakan bencana alam yang seringkali melanda tanah air. Kota Samarinda sendiri merupakan Kota yang saat ini masih sering mengalami kejadian banjir dengan 18 kejadian pada tahun 2018, 33 kejadian tahun 2020 dan 32 kejadian pada tahun 2021. Untuk dapat memprediksi bencana banjir maka dibutuhkan pemanfaatan teknologi yang dikenal dengan machine learning dalam menganalisis dan mengklasifikasikan bencana banjir. Namun, dalam klasifikasi seringkali ditemukan masalah data berdimensi tinggi dan ketidakseimbangan kelas. Penelitian ini bertujuan untuk mengetahui seberapa meningkat akurasi klasifikasi terhadap bencana banjir jika menggunakan algoritma Random Forest dengan PSO sebagai optimasi, seleksi fitur Chi-Square dan oversampling SMOTE untuk menyeimbangkan kelas. Data yang digunakan dalam penelitian ini merupakan data banjir periode 2021-2023 yang didapatkan dari BMKG dan BPBD Kota Samarinda dengan 1095 total record dan 11 atribut. Teknik validasi yang digunakan adalah 5-fold cross-validation dan menggunakan confusion matrix sebagai evaluasi. Hasil penerapan seleksi fitur Chi-Square mengidentifikasi Curah-hujan, Arah-angin-maksimum, Arah-angin-terbanyak, Kelembaban, Lama-penyinaran-matahari, dan Kecepatan-angin sebagai fitur paling berpengaruh dari perangkingan berdasarkan skor Chi-Square dan P-value. Akurasi rata-rata yang didapatkan dari model klasifikasi yang diusulkan dengan teknik validasi 5-fold cross-validation mencapai 96.02%.

Keywords: Klasifikasi; Banjir Random Forest; Imbalance; Chi-Square; Optimization

1 Pendahuluan

Banjir merupakan bencana alam yang seringkali melanda tanah air. Data yang didapatkan dalam 3 tahun terakhir dari Data Informasi Bencana Indonesia (DIBI), total bencana banjir yang terjadi di Indonesia sebanyak 4580 kejadian dan paling banyak terjadi pada 2020 dalam hampir satu dekade terakhir, dengan 1531 kejadian (Annur, 2023; BNPB, 2024).

Kota Samarinda mempunyai 27 aliran sungai dan memiliki dataran rendah dengan kondisi drainase yang tidak cukup baik. Hal tersebut dapat menjadi salah satu penyebab beberapa wilayah di Kota Samarinda yang saat ini masih sering mengalami banjir. Data yang ada menunjukkan bahwa masih tingginya kejadian bencana banjir yang dialami pada hampir seluruh desa/kelurahan di Kota Samarinda, diantaranya 18 kejadian pada tahun 2018, 33 kejadian pada tahun 2020, dan 32 kejadian pada tahun 2021 (BPS, 2024). Bencana banjir yang terjadi dapat mengakibatkan bermacam kerugian, baik itu yang menyebabkan jatuhnya korban jiwa, kerugian ekonomi, ataupun kerusakan properti. Selain dari kerugian tersebut, banjir juga dapat menyebabkan erosi tanah dan hilangnya unsur hara. Jika situasi seperti ini tetap berlanjut maka dikhawatirkan akan terjadi kerusakan yang tidak dapat diperbaiki terhadap sumber daya air dan tanah (Vafakhah et al., 2020).

Untuk dapat memprediksi bencana banjir maka dibutuhkan pemanfaatan teknologi yang dikenal dengan machine learning dalam menganalisis dan mengklasifikasikan bencana banjir tersebut. Dengan harapan dapat memberikan kontribusi terhadap pihak berwenang dalam menghasilkan informasi dan pengetahuan yang bermanfaat dalam pengambilan keputusan untuk memprediksi bencana banjir sehingga dapat membuat keputusan yang lebih tepat dalam meningkatkan kesiapan dan respons terhadap banjir. Machine learning disebut sebagai sebuah algoritma komputasi yang membutuhkan data input untuk menghasilkan output yang diinginkan tanpa pemrograman eksplisit, kemudian dapat meningkatkan dirinya sendiri secara otonom (Grady et al., 2022). Berdasarkan kemampuan machine learning tersebut, maka dalam penelitian ini penulis bermaksud untuk memanfaatkan teknologi machine learning dalam mengklasifikasikan bencana banjir Kota Samarinda.

Klasifikasi merupakan salah satu metode dalam machine learning yang digunakan untuk melakukan prediksi dengan mengelompokkan titik data berdasarkan kriteria yang telah ditentukan (Sharma et al., 2021). Beberapa penelitian sudah pernah dilakukan sebelumnya dalam mengklasifikasikan banjir dengan berbagai pendekatan, mulai dari algoritma Naïve Bayes, Decision Tree, SVM, dan Random Forest(Abu El-Magd, 2022; Sharma et al., 2021; Zhang et al., 2021). Beberapa penelitian juga pernah melakukan klasifikasi terhadap data banjir dengan algoritma Random Forest yang mendapatkan akurasi sebesar 71% dan 58% yang masing-masing memiliki fitur atau dimensi sebanyak 11 dan 10 fitur(Vafakhah et al., 2020; Zhang et al., 2021). Penelitian lain juga pernah melakukan klasifikasi pada data curah hujan yang menggunakan algoritma Random Forest dengan 12 dimensi atau fitur dan mendapatkan akurasi sebesar 73% (Akbar & Sanjaya, 2023).

Dari penelitian yang dilakukan sebelumnya, dapat dikatakan bahwa dataset yang digunakan merupakan dataset berdimensi tinggi. Dataset yang memiliki banyak variabel atau atribut dan digunakan untuk analisis dapat dikatakan sebagai dataset yang berdimensi tinggi (Kurniabudi et al., 2022). Sebagai contoh, jika dataset memiliki banyak variabel, misalnya puluhan atau bahkan ratusan maka dapat dikategorikan sebagai data yang berdimensi tinggi. Menangani data berdimensi tinggi menghadirkan sejumlah tantangan dalam analisis, termasuk kompleksitas perhitungan, risiko overfitting dan kesulitan dalam memvisualisasikan data (Diba, 2023).

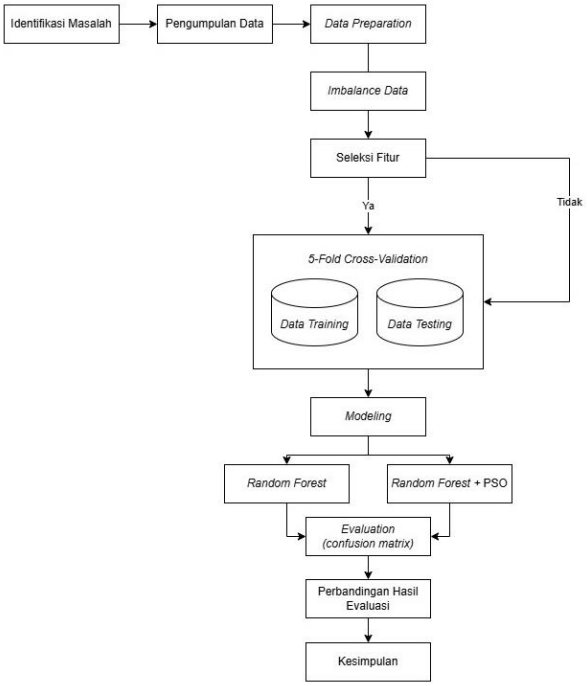
Algoritma klasifikasi yang digunakan dalam penelitian ini adalah Random Forest. Algoritma ini dianggap sebagai algoritma terbaik dalam membuat model klasifikasi karena kinerjanya yang sangat baik dalam hal akurasi (Ijaz et al., 2021). Random Forest juga memiliki kelebihan dalam menangani dataset dengan fitur yang banyak, namun fitur yang banyak tersebut sebaiknya diminimalkan untuk efisiensi (Speiser et al., 2019). Seleksi fitur Chi-Square juga akan digunakan untuk mendapatkan fitur-fitur yang relevan dalam klasifikasi. Seleksi fitur Chi-Square dianggap dapat memberikan peningkatan akurasi(Aiyelokun et al., 2023; Hasan & Al Mehedi Hasan, 2020; Komal Kumar et al., 2019; Saputra & Siswa, 2022; Williamson et al., 2022). Kemudian, untuk mengatasi masalah ketidakseimbangan kelas

terhadap dataset banjir, maka dalam penelitian ini juga akan menggunakan teknik oversampling SMOTE agar performa model yang dihasilkan dapat optimal (Nawi et al., 2020; Priscillia et al., 2022; Razali et al., 2020). Selain itu, optimasi Particle Swarm Optimization (PSO) juga akan digunakan pada penelitian ini, PSO dianggap dapat memberikan peningkatan akurasi terhadap algoritma klasifikasi (Darabi et al., 2021; Dwiasnati & Yudo Devianto, 2022; Khan et al., 2019; Yoga & Prihandoko, 2018).

Berdasarkan penjelasan sebelumnya, maka penelitian ini bertujuan untuk memanfaatkan teknologi machine learning dalam mengklasifikasikan bencana banjir. Selain itu, penelitian ini juga memiliki tujuan utama untuk mengidentifikasi indikator apa saja yang mempengaruhi terjadinya bencana banjir, juga bertujuan mengimplementasikan algoritma random forest dengan PSO sebagai optimasi, seleksi fitur chi-square dan teknik oversampling SMOTE untuk prediksi bencana banjir Kota Samarinda.

2 Metode

Penelitian yang dilakukan mencakup serangkaian langkah-langkah yang dirancang untuk mencapai tujuan penelitian. Proses penelitian yang dilakukan dimulai dengan identifikasi masalah, pengumpulan data, analisis data, hingga tahap evaluasi. Berikut merupakan flowchart dari penelitian ini.



Gambar 1. *Flowchart*

2.1 Pengumpulan Data

Data yang digunakan merupakan data banjir periode 2021-2023. Data yang didapatkan dari BPBD (Badan Penanggulangan Bencana Daerah) dan BMKG (Badan Meteorologi Klimatologi dan Geofisika) Kota Samarinda dengan jumlah atribut masing-masing sebanyak 10 dan 11 atribut

Tabel 1. Data BMKG

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
01-01-2021	23	33,2	26,5	88	1,8	3,3	4	280	2	W
02-01-2021	23,2	30,8	27,1	88	7	6,4	2	140	1	C
03-01-2021	24,8	32,7	27,3	84	2	1,2	5	290	2	NW
…	…	…	…	…	…	…	…	…	…	…
29-12-2023	23,7	32,7	28,7	77	3,5	5,9	5	60	2	NE
30-12-2023	24,2	32,6	28,3	82		10,4	4	60	1	NE
31-12-2023	24,6	32,4	28,3	84	0	6,9	4	90	2	E

Tabel 2. Data BPBD

NO	TANGGAL	JAM KEJADIAN	JENIS BENCANA	LOKASI/ WILAYAH KELURAHAN/ KECAMATAN	LUAS AREA M²	JUMLAH OBYEK YANG TERKENA BENCANA	Korban			JUMLAH			KERUGIAN (RP)	KETERANGAN
							L	S	H	M	K	IWA		
1	03 Januari 2021	-	Banjir	Jl. Irigasi RT. 50 Kel. Rawa Makmur Kec. Palaran (Dataran Rendah) Wilayah Handil Bakti RT. 1, RT. 2, RT 3 (Dataran Rendah)	± -	Jalan mejadi Susah Untuk Di lalui Dan Menganggu aktivitas warga	-			-	-	-	Rp. -	Genangan Air Penyebab Air Sungai Mahakam pasang Dan Lokasi Banjir adalah Dataran Rendah
2	Selasa, 05 Januari 2021	-	Pohon Tumbang	Jl. Kesehatan Dalam Kel. Temindung Permai	-	Jalan mejadi Susah Untuk Di lalui Dan Menganggu aktivitas warga	-	-	-	-	-	-		Penyebab : Hujan Deras dan Angin Kencang
…	…	…	…	…	… …	…	…	…	…	…	…	…	… …	…
19	Rabu, 13 Desember 2023	Pukul 21.33 Wita	Pohon Tumbang	Jl. Gunung Tabur Kel. Gunung Kelua Kec. Samarinda Ulu		Akses jalan tertutup								Penyebab Hujan deras disertai angin kencang Upaya: Melakukkan Pemangkasan
20	Jum'at, 15 Desember 2023	Pukul 15.00 Wita	Pohon Tumbang	Jl. Balai Kota Samarinda Kel. Bugis Kec. Samarinda Kota		Dampak: - Mengenai kanopi Parkiran Bus Pemkot Samarinda								Dampak: Angin Kencang Upaya: Melakukkan Pemangkasan

2.2 Data Preparation

1. Data Integration

Proses digabungkannya kedua data yang telah didapatkan dengan data BPBD yang mempunyai fitur tanggal, jam kejadian, jenis bencana, lokasi wilayah, luas area, objek terkena bencana, korban, kerugian, dan keterangan. Kemudian data dari BMKG yang mempunyai fitur temperatur minimum, temperatur maksimum, temperatur rata-rata, kelembaban, curah hujan, lamanya penyinaran matahari, kecepatan angin maksimum, kecepatan angin rata-rata, dan arah angin terbanyak.

Tabel 3 Data Integration

No	Atribut	Tipe Data	Keterangan
1	Tanggal	*date*	Tanggal Kejadian
2	Jam Kejadian	*String*	Jam kejadian
3	Jenis Bencana	*string*	Bencana alam yang terjadi
4	Lokasi wilayah	*string*	Tempat terjadinya banjir
5	Luas Area M2	*numeric*	Luas area yang terdampak
6	Objek Terkena Bencana	*string*	Kerugian fasilitas yang terdampak bencana
7	Korban	*numeric*	Jumlah korban terdampak bencana
8	Kerugian	*numeric*	Nominal kerugian
9	Keterangan	*numeric*	Detail kejadian bencana
10	Tn	*numeric*	Temperatur minimum (°C)
11	Tx	*numeric*	Temperatur maksimum (°C)
12	Tavg	*numeric*	Temperatur rata-rata (°C)
13	RH_avg	*numeric*	Kelembaban rata-rata (%)
14	RR	*numeric*	Curah hujan (mm)
15	Ss	*numeric*	Lamanya penyinaran matahari (jam)
16	ff_x	*numeric*	Kecepatan angin maksimum (m/s)
17	ddd_x	*numeric*	Arah angin saat kecepatan maksimum (°)
18	ff_avg	*numeric*	Kecepatan angin rata-rata (m/s)
19	ddd_car	*string*	Arah angin terbanyak (°)

2. Data Selection

Data awal dari proses penggabungan data BMKG dan BPBD memiliki 20 kolom atau atribut, kemudian terdapat 9 atribut yang dianggap kurang relevan dan tidak digunakan untuk melakukan prediksi banjir. Sehingga kolom yang awalnya berjumlah 20, setelah dilakukan proses pemilihan, data yang digunakan menjadi 10 kolom yang dijadikan sebagai atribut dan 1 kolom yang dijadikan sebagai kelas.

Tabel 4 Data Selection

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car	terjadi_banjir
01-01-2021	23	33,2	26,5	88	1,8	3,3	4	280	2	W	tidak banjir
02-01-2021	23,2	30,8	27,1	88	7	6,4	2	140	1	C	tidak banjir
03-01-2021	24,8	32,7	27,3	84	2	1,2	5	290	2	NW	banjir
…	…	…	…	…	…	…	…	…	…	…	…
29-12-2023	23,7	32,7	28,7	77	3,5	5,9	5	60	2	NE	tidak banjir
30-12-2023	24,2	32,6	28,3	82		10,4	4	60	1	NE	tidak banjir
31-12-2023	24,6	32,4	28,3	84	0	6,9	4	90	2	E	tidak banjir

3. Data Transformation

Proses untuk mengubah nilai terhadap data menjadi format yang sesuai untuk kepentingan analisis. Tahap ini diperlukan karena pada penggunaannya, library sklearn hanya dapat menerima atribut dengan nilai numerik. Transformasi data yang dilakukan memiliki tujuan untuk mengubah data yang tadinya kategorikal menjadi numerik (Putra et al., 2020). Untuk melakukan transformasi dibutuhkan library LabelEncoder() dari sklearn.

Tabel 5 Perbandingan Sebelum dan Sesudah Data Transformation

No	Arah-angin-terbanyak	Terjadi-banjir
1	W	tidak banjir
2	C	tidak banjir
3	NW	banjir
…	…	…
1092	C	tidak banjir
1093	NE	tidak banjir
1094	E	tidak banjir

No	Arah-angin-terbanyak	Terjadi-banjir
1	8	0
2	0	0
3	4	1
…	…	…
1092	0	0
1093	3	0
1094	1	0

4. Data Cleaning

Pada penelitian ini, data cleaning digunakan untuk menangani nilai yang kosong atau hilang. Penanganan data yang kosong dilakukan menggunakan fungsi dropna() dari python untuk menghapus baris yang memiliki nilai kosong pada data.

```
Jumlah data kosong untuk setiap atribut:
Tanggal                      0
Temperatur-minimum          73
Temperatur-maksimum         10
Temperature-rata-rata        6
Kelembaban                   7
Curah-hujan                133
Lama-penyinaran-matahari     8
Kecepatan-angin              2
Arah-angin-maksimum          2
Kecepatan-angin-rata-rata    2
Arah-angin-terbanyak         2
terjadi_banjir               0
```

Gambar 2 Jumlah Nilai Kosong Sebelum Data Cleaning

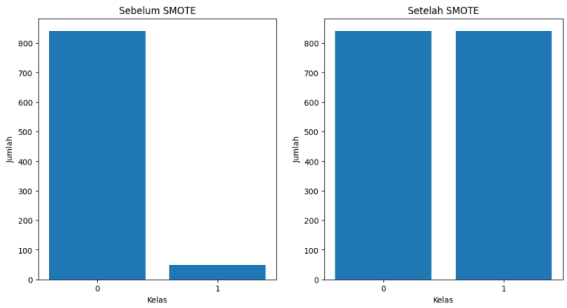
```
jumlah data kosong untuk setiap atribut:
Tanggal                      0
Temperatur-minimum           0
Temperatur-maksimum          0
Temperature-rata-rata        0
Kelembaban                   0
Curah-hujan                  0
Lama-penyinaran-matahari     0
Kecepatan-angin              0
Arah-angin-maksimum          0
Kecepatan-angin-rata-rata    0
Arah-angin-terbanyak         0
terjadi_banjir               0
```

Gambar 3 Jumlah Nilai Kosong Setelah Data Cleaning

5. Data Balancing

Untuk dapat mengatasi ketidakseimbangan kelas pada penelitian ini digunakan teknik oversampling SMOTE. SMOTE bekerja dengan mengidentifikasi kelas minoritas pada dataset, setelah itu SMOTE akan memilih sampel inividu secara acak. Untuk setiap sampel yang telah dipilih SMOTE akan mencari tetangga terdekat untuk sampel dalam feature space. Setelah mendapatkan tetangga terdekat, kemudian SMOTE menciptakan sampel sintetis diantara sampel yang dipilih dan melakukan operasi perkalian dengan bilangan acak antara 0 dan 1

lalu menambahkan hasilnya ke sampel. Sampel-sampel sintetis yang baru tersebut kemudian ditambahkan ke dataset dan meningkatkan jumlah sampel dalam kelas minoritas.



Gambar 4 Perbandingan Sebelum dan Sesudah SMOTE

2.3 Pembagian Data

Teknik k-fold cross-validation digunakan untuk membagi data training dan data testing. Data training difungsikan sebagai dasar dalam pemodelan, sementara data testing digunakan untuk melakukan evaluasi kinerja model klasifikasi yang dibuat. Teknik k-fold pada penelitian ini digunakan dengan pengaturan Cv = 5, yang berarti data dipisahkan menjadi 5 kelompok atau lipatan. Setiap lipatan akan secara bergantian digunakan sebagai data training dan data testing. Setelah lima kali percobaan, nilai rata-rata dari semua hasil pengujian diambil untuk evaluasi yang lebih akurat (Kustiyahningsih et al., 2020).

2.4 Modeling

1. Model Random Forest

Model random forest dibuat dengan library RandomForestClassifier dari scikit-learn. Dalam prosesnya, teknik 5-fold cross-validation diterapkan untuk pembagian data training dan data testing. Adapun rumus dalam perhitungan random forest dan ilustrasi cara kerjanya adalah sebagai berikut:

$$f(x) = \sum_{i=1}^{N} h_i(x) \quad (1)$$

Keterangan:
$f(x)$: Prediksi akhir dari random forest untuk input x
N : Jumlah total pohon keputusan
$h_i(x)$: Prediksi dari pohon keputusan ke-*i* untuk input *x*.
$\sum$: Proses penjumlahan prediksi dari setiap pohon keputusan dalam random forest

2. Model Random Forest + PSO

Pemodelan ini menerapkan particle swarm optimization dalam melakukan pengoptimalan parameter algoritma klasifikasi dengan tujuan meningkatkan performa keseluruhan model dengan menyesuaikan parameter yang dibutuhkan.

3. Model Random Forest + Chi-Square

Dalam tahapan ini dilakukan perangkingan terhadap fitur yang ada, mulai dari fitur yang memiliki pengaruh terbesar terhadap hasil prediksi hingga pada fitur yang memiliki pengaruh terkecil atau bahkan tidak memiliki pengaruh sama sekali terhadap prediksi. Adapun rumus perhitungan chi-square berdasarkan pada data banjir adalah sebagai berikut:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Keterangan:
x^2 : Nilai Statistik chi-square
O_i : Nilai Observasi untuk kategori ke-*i*.
E_i : Nilai ekspektasi untuk kategori ke-*i*.
$\sum$: Penjulahan untuk semua kategori

4. Model Random Forest + Chi-Square + PSO

Pemodelan ini menggunakan tiga algoritma yang telah diselesaikan tahapannya masing-masing yaitu random forest, chi-square, dan PSO. Adapun tahapannya dimulai dari melakukan import terhadap library yang dibutuhkan seperti module SelectKBest, dan chi2 untuk seleksi fitur, module RandomForestClassifier untuk pemodelan random forest, dan module PSO dari library pyswarm untuk optimasi. Kemudian melakukan proses data preparation dan pembagian data, selanjutnya diterapkan algoritma seleksi fitur untuk mendapatkan fitur yang relevan dan menerapkan optimasi. Kemudian mulai melatih model klasifikasi random forest dengan parameter yang telah dioptimalkan.

2.5 Evaluasi

Evaluasi digunakan untuk melakukan pengukuran akurasi hasil dari model algoritma klasifikasi yang diimplementasikan dengan

menggunakan teknik confusion matrix. Confusion matrix digunakan untuk mengukur performa atau kinerja model dengan menghitung nilai acuracy, precision, recall, dan F1-Score. Akurasi merupakan tingkat keakuratan model dalam klasifikasi atau kedekatan prediksi dengan nilai sebenarnya. Berikut formulasi akurasi.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{X } 100\% \quad (3)$$

Keterangan:

TP (True Positive)	:	Jumlah data point berlabel yes yang nilainya diidentifikasi benar
TN (True Negative)	:	Jumlah data point berlabel no yang nilainya diidentifikasi salah
FP (False Positive)	:	Jumlah data point berlabel yes yang nilainya diidentifikasi salah
FN (False Negative)	:	Jumlah data point berlabel no yang nilainya diidentifikasi benar

3 Hasil dan Pembahasan

3.1 Pemodelan Algoritma Random Forest dengan Python

Pemodelan ini menggunakan algoritma Random Forest yang diimplementasikan dengan bahasa pemrograman python. Kemudian mendapatkan hasil evaluasi akurasi dan confusion matrix yang diperoleh dari tiap fold dalam pembagian data training dan testing dengan fold=5 pada klasifikasi data banjir.

Tabel 6 Evaluasi Random Forest dengan Python

Fold	TP	FP	TN	FN	Accuracy
1	172	9	148	8	94.96%
2	155	9	169	4	96.14%
3	173	4	155	4	97.62%
4	162	3	164	7	97.02%
5	151	3	177	5	97.62%

Tabel 6 memperlihatkan hasil evaluasi algoritma Random Forest dengan Python yang menggunakan teknik k-fold cross-validation dengan k=5. Hasil akurasi paling tinggi didapatkan pada fold ketiga dan kelima dengan akurasi yang sama yaitu 97.62%.

Tabel 7 Akurasi Rata-rata Model Random Forest Dari Tiap Fold

Average value of each fold	TP	FP	TN	FN	Accuracy
	813	28	813	28	96.67%

Nilai TP (True Positive) rata-rata yang didapatkan adalah 813 yang berarti model klasifikasi telah dengan benar memprediksi 813 kejadian sebagai ‘banjir’ yang sebenarnya adalah ‘banjir’. Nilai FP (False Positive) rata-rata yang didapatkan adalah 28 yang berarti model klasifikasi telah salah memprediksi 28 kejadian sebagai ‘banjir’ yang sebenarnya adalah ‘tidak banjir’. Nilai TN (True Negative) rata-rata yang didapatkan adalah 813 yang berarti model klasifikasi telah dengan benar memprediksi 813 kejadian sebagai ‘tidak banjir’ yang sebenarnya adalah ‘tidak banjir’. Kemudian nilai FN (False Negative) rata-rata yang didapatkan adalah 28 yang berarti model klasifikasi telah salah memprediksi 28 kejadian sebagai ‘tidak banjir’ yang sebenarnya adalah ‘banjir’. Dari hasil perhitungan tersebut maka didapatkan akurasi dari rata-rata confusion matrix pada tiap fold adalah 96.67%.

3.2 Pemodelan Algoritm Random Forest + PSO Dengan Python

Parameter yang dioptimalkan dalam algoritma Random Forest menggunakan PSO yaitu n_estimators, max_depth, min_samples_split, dan min_samples_leaf. Nilai optimal yang diperoleh untuk n_estimators adalah 149, nilai max_depth 21, nilai min_samples_split 2, dan min_samples_leaf adalah 1.

Tabel 8 Evaluasi Random Forest + PSO dengan Python

Fold	TP	FP	TN	FN	Accuracy
1	171	10	147	9	94.36%
2	155	9	169	4	96.14%
3	174	4	157	2	98.51%
4	164	1	167	4	98.51%
5	150	4	177	5	97.62%

Tabel 8 menampilkan hasil evaluasi algoritma Random Forest + PSO dengan Python, hasil akurasi paling tinggi didapatkan pada fold ketiga dan keempat dengan akurasi 98.51%. Hasil

ini terlihat berbeda dengan hasil sebelumnya pada Tabel 6 yang memperlihatkan hasil evaluasi Random Forest tanpa PSO, terlihat peningkatan akurasi setelah diterapkan optimasi PSO pada fold ketiga dan keempat, pada fold kedua dan kelima mendapatkan akurasi yang sama pada Tabel 6 namun akurasi yang diperoleh tersebut masih dapat dikatakan sangat baik, kemudian pada fold pertama terdapat penurunan sebesar 0.60%. Tetapi akurasi rata-rata yang didapatkan terlihat meningkat setelah diterapkan PSO.

Tabel 9 Akurasi Rata-rata Tiap Fold Random Forest + PSO

Average	TP	FP	TN	FN	Accuracy
value of each fold	814	27	817	24	96.97%

Terdapat perbedaan nilai akurasi rata-rata yang didapatkan terhadap model Random Forest dan Random Forest + PSO, dimana pada Tabel 7, nilai akurasi rata-rata yang didapatkan sebelum menerapkan PSO yaitu 96.67%, sedangkan akurasi rata-rata yang didapatkan setelah menerapkan PSO yaitu 96.97%. Dari perbedaan tersebut dapat dilihat bahwa pada saat PSO diterapkan terhadap model Random Forest maka terjadi peningkatan akurasi, dan akurasi rata-rata yang didapatkan dapat dikatakan sangat tinggi. Nilai TP (True Positive) rata-rata yang didapatkan adalah 814 yang berarti model klasifikasi telah dengan benar memprediksi 814 kejadian sebagai ‘banjir’ yang sebenarnya adalah ‘banjir’. Nilai FP (False Positive) rata-rata yang didapatkan adalah 27 yang berarti model klasifikasi telah salah memprediksi 27 kejadian sebagai ‘banjir’ yang sebenarnya adalah ‘tidak banjir’. Nilai TN (True Negative) rata-rata yang didapatkan adalah 817 yang berarti model klasifikasi telah dengan benar memprediksi 817 kejadian sebagai ‘tidak banjir’ yang sebenarnya adalah ‘tidak banjir’. Kemudian nilai FN (False Negative) rata-rata yang didapatkan adalah 24 yang berarti model klasifikasi telah salah memprediksi 24 kejadian sebagai ‘tidak banjir’ yang sebenarnya adalah ‘banjir’.

3.3 Pemodelan Random Forest + Chi-Square Dengan Python

Chi-Square bertindak sebagai seleksi fitur yang akan melakukan perangkingan terhadap fitur atau atribut yang ada. Perangkingan dilakukan berdasarkan besarnya skor Chi-Square dan kecilnya P-value. Semakin besar skor Chi-Square dan semakin kecil P-value berdasarkan perangkingan, maka akan semakin besar juga pengaruh fitur terhadap kelas yang diuji.

Tabel 10 Perangkingan Chi-Square

Atribut	Skor	P-value	Ranking
Curah-hujan	182156,20	0,0	1
Arah-angin-maksimum	2553,98	0,0	2
Arah-angin-terbanyak	12,4277	0,0004	3
Kelembaban	11,9999	0,0005	4
Lama-penyinaran-matahari	10,1282	0,0014	5
Kecepatan-angin	7,55435	0,0059	6
Temperature-rata-rata	1,37504	0,2409	7
Kecepatan-angin-rata-rata	0,08719	0,7677	8
Temperatur-maksimum	0,00031	0,9857	9
Temperatur-minimum	7,09527	0,9932	10

Berdasarkan hasil tersebut, maka akan menggunakan atribut rangking 1 hingga 6 untuk pemodelan, karena memiliki skor Chi-Square yang tinggi dan p-value yang rendah. Atribut yang akan digunakan yaitu Curah-hujan, Arah-angin-maksimum, Arah-angin-terbanyak, Kelembaban, Lama-penyinaran-matahari, dan Kecepatan-angin.

Tabel 11 Evaluasi Random Forest + Chi-Square

Fold	TP	FP	TN	FN	Accuracy
1	173	8	143	13	93.77%
2	155	9	167	6	95.55%
3	168	9	157	2	96.73%
4	159	6	163	8	95.83%
5	150	4	174	8	96.43%

Tabel 11 memperlihatkan hasil evaluasi algoritma Random Forest yang menerapkan seleksi fitur Chi-Square dengan Python yang menggunakan teknik k-fold cross-validation dengan k=5. Hasil akurasi paling tinggi didapatkan pada fold ketiga dengan akurasi 96.73%.

Tabel 12 Akurasi Rata-rata Tiap Fold Random Forest + Chi-Square

Average value of each fold	TP	FP	TN	FN	Accuracy
	805	36	804	37	95.66%

Akurasi rata-rata yang diperoleh dari model klasifikasi Random Forest dan seleksi fitur Chi-Square pada tiap fold. Nilai TP (True Positive) rata-rata yang didapatkan adalah 805 yang berarti model klasifikasi telah dengan benar memprediksi 805 kejadian sebagai 'banjir' yang sebenarnya adalah 'banjir'. Nilai FP (False Positive) rata-rata yang didapatkan adalah 36 yang berarti model klasifikasi telah salah memprediksi 36 kejadian sebagai 'banjir' yang sebenarnya adalah 'tidak banjir'. Nilai TN (True Negative) rata-rata yang didapatkan adalah 804 yang berarti model klasifikasi telah dengan benar memprediksi 804 kejadian sebagai 'tidak banjir' yang sebenarnya adalah 'tidak banjir'. Kemudian nilai FN (False Negative) rata-rata yang didapatkan adalah 37 yang berarti model klasifikasi telah salah memprediksi 37 kejadian sebagai 'tidak banjir' yang sebenarnya adalah 'banjir'. Dari hasil perhitungan tersebut maka didapatkan akurasi dari rata-rata confusion matrix pada tiap fold adalah 95.66%.

3.4 Pemodelan Algoritma Random Forest + Chi-Square + PSO Dengan Python

Dalam pemodelan ini, setelah seleksi fitur Chi-Square diterapkan selanjutnya melakukan optimasi terhadap model klasifikasi untuk meningkatkan performa model klasifikasi Random Forest + Chi-Square. Adapun parameter yang dioptimalkan dalam algoritma Random Forest yaitu n_estimators, max_depth, min_samples_split, dan min_samples_leaf. Nilai optimal yang diperoleh untuk n_estimators adalah 149, nilai max_depth 21, nilai min_samples_split 2, dan min_samples_leaf adalah 1.

Tabel 13 Evaluasi Random Forest + Chi-Square + PSO

Fold	TP	FP	TN	FN	Accuracy
1	174	7	144	12	94.36%
2	157	7	168	5	96.44%
3	165	12	157	2	95.83%
4	160	5	163	8	96.13%
5	150	4	177	5	97.32%

Tabel 13 menampilkan hasil evaluasi algoritma Random Foret + Chi-Square + PSO dengan Python, hasil akurasi paling tinggi didapatkan pada fold kelima dengan akurasi 97.32% . Hasil ini terlihat berbeda dengan hasil sebelumnya pada Tabel 11 yang memperlihatkan hasil evaluasi Random Forest + Chi-Square tanpa PSO, terlihat peningkatan akurasi setelah diterapkan optimasi PSO pada fold pertama, kedua, keempat, dan kelima. Namun pada fold ketiga mendapatkan penurunan akurasi tetapi akurasi yang diperoleh tersebut masih dapat dikatakan sangat baik.

Tabel 14 Akurasi Rata-rata Tiap Fold Random Forest +Chi-Square + PSO

Average value of each fold	TP	FP	TN	FN	Accuracy
	806	35	809	32	96.02%

Tabel di atas merupakan akurasi rata-rata dari algoritma Random Forest + Chi-Square + PSO dari tiap fold. Nilai confusion matrix yang didapatkan berasal dari jumlah keseluruhan fold yang digunakan untuk mendapatkan nilai akurasi. Terdapat perbedaan nilai akurasi rata-rata yang didapatkan terhadap model Random Forest + Chi-Square dan Random Forest + Chi-Square + PSO, dimana pada Tabel 12, nilai akurasi rata-rata yang didapatkan sebelum menerapkan PSO yaitu 95.66%, sedangkan akurasi rata-rata yang didapatkan setelah menerapkan PSO yaitu 96.02%. Dari perbedaan tersebut dapat dilihat bahwa pada saat PSO diterapkan terhadap model Random Forest yang menggunakan seleksi fitur Chi-Square maka terjadi peningkatan akurasi, akurasi rata-rata yang didapatkan masih dapat dikatakan sangat tinggi. Nilai TP (True Positive) rata-rata yang didapatkan adalah 806 yang berarti model klasifikasi telah dengan benar memprediksi 806 kejadian sebagai 'banjir' yang sebenarnya adalah 'banjir'. Nilai FP (False Positive) rata-rata yang didapatkan adalah 35 yang berarti model klasifikasi telah salah memprediksi 35 kejadian sebagai 'banjir' yang sebenarnya adalah 'tidak banjir'. Nilai TN (True Negative) rata-rata yang didapatkan adalah 809 yang berarti model klasifikasi telah dengan benar memprediksi 809 kejadian sebagai 'tidak banjir' yang sebenarnya adalah 'tidak banjir'. Kemudian nilai FN (False Negative) rata-rata yang didapatkan adalah 32 yang berarti model

klasifikasi telah salah memprediksi 32 kejadian sebagai 'tidak banjir' yang sebenarnya adalah 'banjir'.

3.5 Perbandingan Hasil Evaluasi

Perbandingan dilakukan berdasarkan evaluasi terhadap akurasi yang diperoleh dari masing-masing model, perbandingan ini meliputi model yang hanya menggunakan algoritma Random Forest, model Random Forest dengan optimasi PSO, model Random Forest dengan seleksi fitur Chi-Square, dan model Random Forest dengan seleksi fitur Chi-Square dan optimasi PSO.

Tabel 15 Perbandingan Akurasi Model Klasifikasi Random Forest

RF				**RF + *Chi-Square***		
Fold	**RF**	**RF + PSO**	**Status**	**RF + *Chi-Square***	**RF + *Chi-Square* + PSO**	**Status**
1	94.96%	94.36%	Turun	93.77%	94.36%	Naik
2	96.14%	96.14%	-	95.55%	96.44%	Naik
3	97.62%	98.51%	Naik	96.73%	95.83%	Turun
4	97.02%	98.51%	Naik	95.83%	96.13%	Naik
5	97.62%	97.62%	-	96.43%	97.32%	Naik
Waktu Proses				**Waktu Proses**		
2.898s				0.007s		

Forest yang diperlihatkan pada Tabel 15 menunjukkan peningkatan pada beberapa fold terhadap model klasifikasi Random Forest yang dioptimasi. Model klasifikasi yang hanya menggunakan algoritma Random Forest telah memperoleh nilai akurasi yang sangat tinggi pada tiap fold-nya, pada saat diterapkan optimasi terhadap model yang hanya menggunakan algoritma Random Forest terdapat beberapa peningkatan. Perubahan juga dapat dilihat saat sebelum dan sesudah optimasi diterapkan pada model klasifikasi Random Forest yang menggunakan seleksi fitur Chi-Square. Terdapat peningkatan dan penurunan terhadap beberapa fold-nya dan dapat dikatakan bahwa akurasi yang diperoleh sudah sangat tinggi.

Tabel 16 Perbandingan Rata-rata Akurasi Model Klasifikasi Random Forest

RF			**RF + *Chi-Square***		
RF	**RF + PSO**	**Status**	**RF + *Chi-Square***	**RF + *Chi-Square* + PSO**	**Status**
96.67%	96.97%	Naik	95.66%	96.02%	Naik

Terlihat peningkatan performa akurasi dari masing-masing model klasifikasi Random Forest sebelum dan sesudah dioptimasi terhadap model yang tidak menggunakan seleksi fitur dan yang menggunakan seleksi fitur. Akurasi rata-rata yang didapatkan terhadap model yang menggunakan seleksi fitur terlihat lebih rendah dibandingkan dengan akurasi yang didapatkan terhadap model yang tidak menggunakan seleksi fitur. Namun waktu pemrosesan terhadap model yang menggunakan seleksi fitur sedikit lebih cepat dibanding dengan waktu pemrosesan terhadap model tanpa seleksi fitur.

3.6 Pembahasan

Seleksi fitur Chi-Square yang digunakan dalam pemodelan klasifikasi Random Forest dengan data banjir Kota Samarinda mengidentifikasi fitur 'Curah-hujan' sebagai fitur yang paling berpengaruh terhadap model klasifikasi dengan skor Chi-Square 182156,20 dan p-value 0,0. Diikuti oleh 'Arah-angin-maksimum' dengan skor Chi-Square 2553,98 dan p-value 0,0, dan 'Arah-angin-terbanyak' dengan skor Chi-Square 12,4277 dan p-value 0,0004. Fitur-fitur yang lain seperti 'Kelembaban', 'Lama-penyinaran-matahari', dan 'Kecepatan-angin' menempati 4 hingga 6 dalam perangkingan. Penelitian sebelumnya oleh Intan Permatasari et al. (2023) yang menggunakan seleksi fitur gain ratio mengidentifikasi 'Kelembaban', 'Temperatur-minimum', dan 'Temperatur-maksimum' sebagai fitur yang paling berpengaruh dan menunjukkan peningkatan akurasi algoritma klasifikasi KNN

dengan persentasi kenaikan akurasi sebesar 5.95%.

Penelitian oleh Saputra dan Siswa dalam melakukan prediksi keterlambatan biaya kuliah menggunakan algoritma C4.5 dan seleksi fitur chi-square, hasil akurasinya mengalami peningkatan sebesar 4.13% dari 61.40% menjadi 65.53%. Kemudian penelitian oleh Williamson dalam melakukan prediksi biopsi kanker payudara menggunakan algoritma random forest dan seleksi fitur chi-square hasil akurasinya mengalami peningkatan sebesar 0.83% dari 83.87% menjadi 84.7%. Dari kedua penelitian tersebut didapatkan bahwa penggunaan seleksi fitur chi-square terhadap algoritma klasifikasi dapat meningkatkan nilai akurasi. Namun hal tersebut berbeda dengan hasil penelitian yang dilakukan untuk melakukan klasifikasi data banjir Kota Samarinda, dimana terjadi penurunan akurasi terhadap algoritma klasifikasi random forest ketika diterapkan seleksi fitur chi-square. Melihat penelitian yang dilakukan oleh Fauzan yang melakukan klasifikasi Al-Qur'an menggunakan algoritma Support Vector Machine dan seleksi fitur chi-square, dan Abdullah yang melakukan klasifikasi gangguan spektrum autisme yang salah satunya menggunakan algoritma klasifikasi random forest dan seleksi fitur chi-square juga mengalami penurunan akurasi. Penurunan akurasi terhadap algoritma klasifikasi ketika menerapkan seleksi fitur chi-square tersebut dapat disebabkan karena penghapusan fitur berguna dari dataset. Hal tersebut menyebabkan kehilangan informasi penting yang dapat membantu model dalam melakukan klasifikasi dengan lebih akurat. Sehingga, meskipun chi-square dapat mengidentifikasi fitur yang paling relevan, penghapusan fitur yang kurang relevan secara individu tetapi penting ketika dikombinasikan dengan fitur yang lain dapat mengakibatkan penurunan performa terhadap model klasifikasi.

Penggunaan algoritma Random Forest dalam klasifikasi data banjir yang menggunakan teknik oversampling SMOTE dalam ¬Data Preprocessing-nya tanpa seleksi fitur dan optimasi berhasil mendapatkan rata-rata akurasi sebesar 96.67%, kemudian mendapatkan peningkatan akurasi sebesar 0.30% setelah menerapkan optimasi sehingga akurasinya menjadi 96.97%. Kemudian model klasifikasi Random Forest yang menggunakan teknik oversampling SMOTE dalam Data Preprocessing-nya yang menggunakan seleksi fitur Chi-Square mendapatkan rata-rata akurasi sebesar 95.66% dan mendapatkan peningkatan akurasi sebesar 0.36% sehingga akurasinya menjadi 96.02% setelah menerapkan optimasi.

Jika dibandingkan antara model klasifikasi yang menggunakan seleksi fitur dan tidak menggunakan seleksi fitur, maka model klasifikasi Random Forest yang menggunakan seleksi fitur mendapatkan performa yang lebih rendah dengan model klasifikasi Random Forest tanpa seleksi fitur. Namun, kecepatan pemrosesan yang menggunakan seleksi fitur Chi-Square lebih cepat 2.28 detik dibandingkan model yang tanpa seleksi fitur.

4 Kesimpulan

Dari hasil seleksi fitur Chi-Square yang telah diterapkan pada data banjir Kota Samarinda, maka dipilih 6 fitur yang memiliki pengaruh terbesar berdasarkan hasil dari perangkingan fitur yaitu Curah-hujan, Arah-angin-maksimum, Arah-angin-terbanyak, Kelembaban, Lama-penyinaran-matahari, dan Kecepatan-angin.

Hasil penerapan metode optimasi Particle Swarm Optimization (PSO), terbukti mampu meningkatkan akurasi terhadap model klasifikasi Random Forest pada data banjir Kota Samarinda yang menggunakan seleksi fitur ¬Chi-Square dan tanpa seleksi fitur Chi-Square. Pada model klasifikasi tanpa menggunakan seleksi fitur berhasil meningkatkan akurasi sebesar 0.30% dari yang akurasinya 96.67% menjadi 96.97%. Kemudian pada model klasifikasi yang menggunakan seleksi fitur Chi-Square berhasil meningkatkan akurasi sebesar 1.07% dari yang akurasinya 95.60% menjadi 96.67%. Dalam hal akurasi, model klasifikasi Random Forest pada data banjir Kota Samarinda tanpa seleksi fitur Chi-Square lebih sedikit unggul dibandingkan dengan model klasifikasi yang menggunakan seleksi fitur.

Referensi

Abu El-Magd, S. A. (2022). Random forest and naïve Bayes approaches as tools for flash flood hazard susceptibility prediction, South Ras El-Zait, Gulf of Suez Coast, Egypt. *Arabian*

Journal of Geosciences, *15*(3), 1–12. https://doi.org/10.1007/s12517-022-09531-3

Aiyelokun, O. O., Aiyelokun, O. D., & Agbede, O. A. (2023). Application of random forest (RF) for flood levels prediction in Lower Ogun Basin, Nigeria. *Natural Hazards*, *119*(3), 2179–2195. https://doi.org/10.1007/s11069-023-06211-7

Akbar, H., & Sanjaya, W. K. (2023). Kajian Performa Metode Class Weight Random Forest pada Klasifikasi Imbalance Data Kelas Curah Hujan. *Jurnal Sains, Nalar, Dan Aplikasi Teknologi Informasi*, *3*(1). https://doi.org/10.20885/snati.v3i1.30

Annur, C. M. (2023). *BNPB: Tren Banjir di Indonesia Cenderung Menurun dalam Tiga Tahun Terakhir*. Databoks. https://databoks.katadata.co.id/datapublish/2023/02/20/bnpb-tren-banjir-di-indonesia-cenderung-menurun-dalam-tiga-tahun-terakhir

BNPB. (2024). *Infografis*. BNPB. https://bnpb.go.id/infografis

BPS. (2024). *Jumlah Desa/Kelurahan yang Mengalami Bencana Alam1 [Banjir] Menurut Kecamatan di Kota Samarinda 2018-2021*. Badan Pusat Statistik Kota Samarinda. https://samarindakota.bps.go.id/indicator/153/207/1/jumlah-desa-kelurahan-yang-mengalami-bencana-alam-banjir-menurut-kecamatan-di-kota-samarinda.html

Darabi, H., Torabi Haghighi, A., Rahmati, O., Jalali Shahrood, A., Rouzbeh, S., Pradhan, B., & Tien Bui, D. (2021). A hybridized model based on neural network and swarm intelligence-grey wolf algorithm for spatial prediction of urban flood-inundation. *Journal of Hydrology*, *603*(PA), 126854. https://doi.org/10.1016/j.jhydrol.2021.126854

Diba, F. (2023). Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi. *Indonesian Journal of Computer Science*, *12*(4), 2152–2160. https://doi.org/10.33022/ijcs.v12i4.3329

Dwiasnati, S., & Yudo Devianto. (2022). Optimization of Flood Prediction using SVM Algorithm to determine Flood Prone Areas. *Journal of Systems Engineering and Information Technology (JOSEIT)*, *1*(2), 40–46. https://doi.org/10.29207/joseit.v1i2.1995

Grady, F., Tarigan, J. K., Wahidiyat, J. R., & Prasetyo, A. (2022). Classification of Flood Alert in Jakarta with Random Forest. *Proceedings of the 2022 IEEE 7th International Conference on Information Technology and Digital Applications, ICITDA 2022*, 1–6. https://doi.org/10.1109/ICITDA55840.2022.9971411

Hasan, K. A., & Al Mehedi Hasan, M. (2020). Classification of Parkinson's Disease by Analyzing Multiple Vocal Features Sets. *2020 IEEE Region 10 Symposium, TENSYMP 2020, June*, 758–761. https://doi.org/10.1109/TENSYMP50017.2020.9230842

Ijaz, M., Asghar, Z., & Gul, A. (2021). Ensemble of penalized logistic models for classification of high-dimensional data. *Communications in Statistics: Simulation and Computation*, *50*(7), 2072–2088. https://doi.org/10.1080/03610918.2019.1595647

Khan, T., Alam, M., Shaikh, F. A., Khan, S., Kadir, K., Mazliham, M. S., Shahid, Z., & Yahya, M. (2019). Flash floods prediction using real time data: An implementation of ANN-PSO with less false alarm. *I2MTC 2019 - 2019 IEEE International Instrumentation and Measurement Technology Conference, Proceedings*, *2019-May*, 1–6. https://doi.org/10.1109/I2MTC.2019.8826825

Komal Kumar, N., Vigneswari, D., Vamsi Krishna, M., & Phanindra Reddy, G. V. (2019). An optimized random forest classifier for diabetes mellitus. In *Advances in Intelligent Systems and Computing* (Vol. 813). Springer Singapore. https://doi.org/10.1007/978-981-13-1498-8_67

Kurniabudi, K., Harris, A., & Veronica, V. (2022). Komparasi Performa Tree-Based Classifier Untuk Deteksi Anomali Pada Data Berdimensi Tinggi dan Tidak Seimbang. *Jurnal Media Informatika Budidarma*, *6*(1), 370. https://doi.org/10.30865/mib.v6i1.3473

Kustiyahningsih, Y., Mula'ab, & Hasanah, N. (2020). Metode Fuzzy ID3 Untuk Klasifikasi Status Preeklamsi Ibu Hamil. *Teknika*, *9*(1), 74–80. https://doi.org/10.34148/teknika.v9i1.270

Nawi, N. M., Makhtar, M., Salikon, M. Z., & Afip, Z. A. (2020). A comparative analysis of classification techniques on predicting flood risk. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(3), 1342–1350. https://doi.org/10.11591/ijeecs.v18.i3.pp1342-1350

Priscillia, S., Schillaci, C., & Lipani, A. (2022). Arti fi cial Intelligence in Geosciences Flood susceptibility assessment using arti fi cial neural networks in Indonesia. *Artificial Intelligence in Geosciences*, *2*(April), 215–222. https://doi.org/10.1016/j.aiig.2022.03.002

Putra, M. I., Yusuf, A., & Yalina, N. (2020). Klasifikasi Kelancaran Kredit Dengan Metode Random Forest. *Systemic: Information System*

and Informatics Journal, *5*(2), 7–12. https://doi.org/10.29080/systemic.v5i2.713

Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. *IAES International Journal of Artificial Intelligence*, *9*(1), 73–80. https://doi.org/10.11591/ijai.v9.i1.pp73-80

Saputra, A., & Siswa, T. A. Y. (2022). Optimasi Chi Square Dan Perbaikan Teknik Prunning Untuk Peningkatan Akurasi Algoritma C4.5 Dalam Model Kasus Prediksi Keterlambatan Biaya Kuliah. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(2), 231. https://doi.org/10.26798/jiko.v6i2.648

Sharma, P., Kar, B., Wang, J., & Bausch, D. (2021). A machine learning approach to flood severity classification and alerting. *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities, ARIC 2021*, *November*, 42–47. https://doi.org/10.1145/3486626.3493432

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Vafakhah, M., Mohammad Hasani Loor, S., Pourghasemi, H., & Katebikord, A. (2020). Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. *Arabian Journal of Geosciences*, *13*(11), 1–16. https://doi.org/10.1007/s12517-020-05363-1

Williamson, S., Vijayakumar, K., & Kadam, V. J. (2022). Predicting breast cancer biopsy outcomes from BI-RADS findings using random forests with chi-square and MI features. *Multimedia Tools and Applications*, *81*(26), 36869–36889. https://doi.org/10.1007/s11042-021-11114-5

Yoga, T. A., & Prihandoko. (2018). Penerapan Optimasi Berbasis Particle Swarm Optimization (Pso) Algoritma Naïve Bayes Dan K-Nearest Neighbor Sebagai Perbandingan Untuk Mencari Kinerja Terbaik Dalam Mendeteksi Kanker Payudara. *Jurnal Bangkit Indonesia*, *7*(2), 1. http://journal.universitasmulia.ac.id/index.php/metik/article/view/62

Zhang, Z., Qiu, J., Huang, X., Cai, Z., Zhu, L., & Dai, W. (2021). Comparing and Evaluating Macao Flood Prediction Models. *IOP Conference Series: Earth and Environmental Science*, *769*(2). https://doi.org/10.1088/1755-1315/769/2/022001