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 (1194-1203) **DOI: 10.32493/jtsi.v7i3.41068**

Optimasi SVM dengan RFE dan ROS untuk Mengatasi High Dimension dan Imbalanced Data Banjir

Faldy Alfareza Pambudi[1], Taghfirul Azhima Yoga Siswa*[2], Wawan Joko Pranoto[3]

Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia, 75124
e-mail: [1]2011102441097@umkt.ac.id, [*2]tay758@umkt.ac.id, [3] wjp337@umkt.ac.id

Submitted Date: June 20th, 2024 Reviewed Date: June 22th, 2024
Revised Date: July 19th, 2024 Accepted Date: July 24th, 2024

Abstract

Floods are natural disasters that often occur in Indonesia, one of which is the city of Samarinda which experienced a significant increase in flood cases in 2018-2021. The use of machine learning, especially the Support Vector Machine (SVM) algorithm, aims to accurately predict future flood events, but the main problem faced is data imbalance and high-dimensional data. This research combines SVM with Random Oversampling (ROS) oversampling techniques and Recursive Feature Elimination (RFE) feature selection to overcome data imbalance and high-dimensional data, with the aim of increasing the classification accuracy of Samarinda City flood data. The cross validation method is with 10-fold cross-validation, and the model performance is evaluated with a confusion matrix to calculate the accuracy value. The data used was obtained from BPDB and BMKG Samarinda City for the 2021-2023 period, consisting of 11 attributes and 1095 lines of data. The research results show that RFE succeeded in identifying the five most important features, namely minimum temperature (Tn), maximum temperature (Tx), average temperature (Tavg), humidity (RH_avg) and maximum wind direction (ddd_x). With the combination of SVM, ROS, and RFE models, flood data classification accuracy increased by 0.78% from 97.14% to 97.92%.

Keywords: Flood Classification; SVM; RFE; ROS; Imbalanced Data; High Dimension

Abstrak

Banjir merupakan bencana alam yang sering terjadi di Indonesia, salah satunya Kota Samarinda yang mengalami peningkatan signifikan kasus banjir di tahun 2018-2021. Penggunaan *machine learning*, khususnya algoritma *Support Vector Machine* (SVM), bertujuan untuk memprediksi kejadian banjir mendatang dengan akurat, namun masalah utama yang dihadapi adalah ketidakseimbangan data dan data berdimensi tinggi. Penelitian ini mengkombinasikan SVM dengan teknik *oversampling Random Oversampling* (ROS) dan seleksi fitur *Recursive Feature Elimination* (RFE) untuk mengatasi ketidakseimbangan data dan data berdimensi tinggi, dengan tujuan untuk meningkatkan akurasi klasifikasi data banjir Kota Samarinda. Metode validasi silang dengan 10-*fold cross-validation*, dan performa model dievaluasi dengan *confusion matrix* untuk menghitung nilai akurasi. Data yang digunakan diperoleh dari BPDB dan BMKG Kota Samarinda periode tahun 2021-2023, yang terdiri dari 11 atribut dan 1095 baris data. Hasil penelitian bahwa RFE berhasil mengidentifikasi lima fitur terpenting yaitu, temperatur minimum (Tn), temperatur maksimum (Tx), temperatur rata-rata (Tavg), kelembapan (RH_avg) dan arah angin maksimum (ddd_x). Dengan kombinasi model SVM, ROS, dan RFE, akurasi klasifikasi data banjir meningkat sebesar 0,78% dari 97,14% menjadi 97,92%.

Kata Kunci: Klasifikasi Banjir; SVM; RFE; ROS; Imbalanced Data; High Dimension

1. Pendahuluan

Indonesia sebagai negara maritim beriklim tropis, sering mengalami bencana alam, terutama banjir. Menurut Badan Nasional Penanggulangan Bencana (BNPB), banjir adalah bencana alam kedua yang paling sering terjadi di Indonesia pada

tahun 2023, dengan 1.170 kejadian (Andrean, 2024). Banjir dapat diprediksi melalui pola curah hujan, aliran air, dan faktor lingkungan seperti berkurangnya lahan hijau (Dilla Evitasari et al., 2023). Kota Samarinda di Provinsi Kalimantan Timur sering mengalami banjir dalam beberapa tahun terakhir. Data dari Badan Pusat Statistik (BPS) Kota Samarinda menunjukkan peningkatan jumlah desa/kelurahan yang terkena banjir, dari 18 desa pada tahun 2018 menjadi 33 desa pada tahun 2020, dan 32 desa pada tahun 2021 (BPS). Maka diperlukan teknik klasifikasi dengan *data mining* atau *machine learning* untuk mengurangi risiko terjadinya banjir.

Mengklasifikasikan banjir berdasarkan faktor penyebabnya dapat meningkatkan keakuratan perkiraan frekuensi banjir di tingkat lokal dan regional serta membantu identifikasi dan penafsiran perubahan kejadian serta skala banjir (Tarasova et al., 2019). Penelitian sebelumnya menggunakan beberapa pendekatan data mining atau machine learning untuk klasifikasi banjir, termasuk algoritma *Random Forest* (Puspasari et al., 2023), KNN (Gauhar et al., 2021), *Naïve Bayes* (Nawi et al., 2020), SVM, dan C5.0 (Fitrianah et al., 2022). Meskipun penelitian-penelitian ini menunjukkan akurasi rata-rata di atas 90%, data yang digunakan adalah dataset berdimensi rendah.

Dalam *data mining* dimensi sebuah dataset dapat mempengaruhi kinerja sebuah algoritma terutama data yang berdimensi tinggi atau *high dimension data*. Data berimensi tinggi adalah struktur data yang memiliki jumlah fitur yang banyak dan melebihi jumlah observasi (Idris et al., 2022). Selain itu data berdimensi tinggi juga menimbulkan kendala dalam penerapan teknik pembelajaran mesin karena memberikan dampak negatif terhadap analisis (Fauzi et al., 2020). Penelitian terkait menunjukkan akurasi rendah pada dataset berdimesni tinggi. Misalnya, penelitian yang dilakukan oleh Ahmmed (Ahmmed et al., 2022) pada klasifikasi banjir, yang menunjukkan rata-rata hasil akurasi hanya sebesar 50% hingga 60% dengan menguji berbagai algoritma. Penelitian klasifikasi banjir yang lain, hasil akurasi terendah diperoleh algoritma SVM yaitu 67% dan ML (*Multinominal Logit*) sebesar 79% (Sharma et al., 2021). Hal ini menunjukkan bahwa dimensi dataset mempengaruhi performa algoritma.

Penelitian ini menggunakan algoritma *Support Vector Machine* (SVM) untuk klasifikasi data banjir di Kota Samarinda. SVM adalah algoritma pembelajaran mesin yang mengolah data menjadi hyperplane untuk mengklasifikasikan data linier berdasarkan kelasnya (Huang Kendrew, 2022). SVM dianggap memiliki performa terbaik dalam klasifikasi data banjir dibandingkan dengan KNN (Khan et al., 2019) dan menunjukkan kinerja yang baik dibandingkan dengan *Random Forest* dan ANN (Duwal et al., 2023). Namun, SVM memiliki kelemahan pada dataset berdimensi tinggi, yang menyebabkan penurunan kinerja dan akurasi rendah (Ahmmed et al., 2022) (Sharma et al., 2021). Karena SVM menunjukkan performa yang kurang baik ketika berhadapan dengan dataset berdimensi tinggi, maka digunakanlah seleksi fitur sebagai solusi. Tujuannya adalah untuk mengurangi jumlah fitur yang ada dengan hanya menggunakan fitur-fitur yang paling relevan jika dibandingkan dengan lainnya (M. Adib Al Karomi, Abdul Kharis, 2019). Berdasarkan masalah tersebut, penelitian ini mengusulkan seleksi fitur *Recursive Feature Elimination* (RFE) untuk mengatasi kelemahan tersebut.

Berhubungan dengan RFE para peneliti telah menggunakan kombinasi SVM dengan RFE untuk menangani dataset berdimensi tinggi. Kombinasi ini meningkatkan akurasi klasifikasi curah hujan sebesar 2% untuk data uji 70:30 dan 4% untuk 80:20 (Pratama et al., 2022). Penelitian lain menunjukkan peningkatan akurasi sebesar 14% setelah menggunakan RFE (Rustam et al., 2019). SVM dengan RFE juga menunjukkan performa yang lebih baik dibandingkan dengan metode *Chi-Square* dan *Information Gain* (Thakkar & Lohiya, 2021). Hal ini menegaskan bahwa seleksi fitur signifikan meningkatkan kinerja algoritma klasifikasi pada dataset berdimensi tinggi.

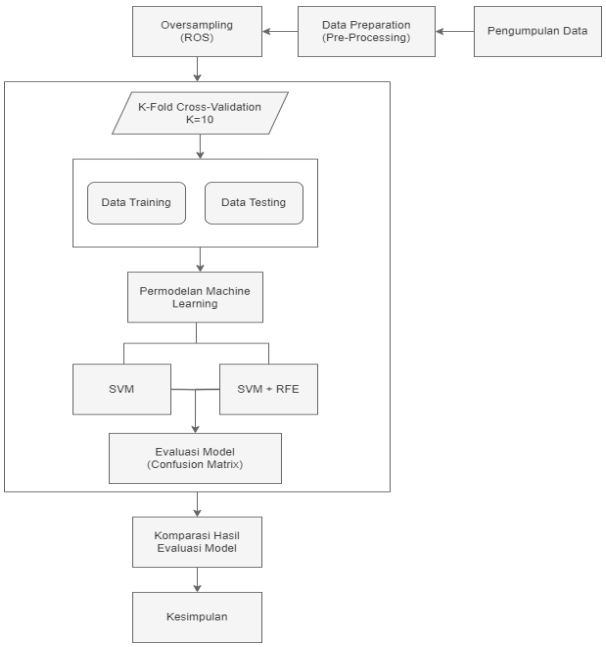
Selain data berdimensi tinggi, ketidakseimbangan data atau *imbalanced data* juga sering menjadi masalah dalam proses klasifikasi. Ketidakseimbangan distribusi kelas dapat memengaruhi kinerja algoritma klasifikasi (Pristyanto, 2019). Untuk mengatasi masalah ini, peneliti mengembangkan teknik *resampling* seperti *undersampling* atau *oversampling*, yang bertujuan menyeimbangkan distribusi kelas dalam dataset (Gumelar et al., 2021). Penelitian ini menggunakan teknik *oversampling* dengan metode *Random Oversampling* (ROS) karena mudah diterapkan dan

sering digunakan (Rifqi Fitriadi & Deni Mahdiana, 2023). Penelitian sebelumnya menunjukkan bahwa metode ROS mampu meningkatkan akurasi rata-rata sekitar 30% dengan beberapa algoritma klasifikasi (Salsadilla et al., 2023). ROS bersama SVM meningkatkan akurasi sebesar 0,89% (Uddin et al., 2023). Kombinasi ROS dengan beberapa kernel SVM, seperti yang diteliti oleh Ramadhan (Ramadhan et al., 2023), menunjukkan peningkatan akurasi signifikan, kecuali kernel *sigmoid* yang mengalami penurunan. Pada data banjir, kombinasi SVM dan ROS memberikan akurasi lebih baik dibandingkan dengan SMOTE dan SVM-SMOTE (Rahman et al., 2023).

Berdasarkan uraian diatas, penelitian ini akan menerapkan metode SVM sebagai algoritma klasifikasi utama, kemudian menggunakan *Recursive Feature Elimination* (RFE) sebagai seleksi fitur, dan menggunakan *Random Oversampling* (ROS) sebagai *oversampling*. Tujuannya adalah untuk mengidentifikasi sejauh mana peningkatan akurasi yang dapat dicapai dan bagaimana hal tersebut mempengaruhi klasifikasi data banjir di Kota Samarinda.

2. Metodologi

Penelitian ini melibatkan serangkaian langkah-langkah yang akan dilaksanakan untuk mencapai tujuan penelitian.



Gambar 1. Diagram Alur

Berdasarkan gambar 1, uraian dari metode penelitian ini dijelaskan pada subbab berikutnya.

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini berasal dari data banjir di Kota Samarinda yang didapatkan dari Badan Penanggulangan Bencana Daerah (BPBD) dan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Kota Samarinda periode tahun 2021-2023.

2.2 Data Preparation

Data yang diperoleh dari Badan Penanggulangan Bencana Daerah (BPBD) dan Badan Meteorologi Klimatologi dan Geofisika (BMKG) memerlukan pengolahan lebih lanjut sebelum dapat dimasukkan ke dalam proses pemodelan, dengan tujuan menghindari data yang tidak relevan. Proses pengolahan data ini meliputi beberapa tahap, termasuk *data integration*, *data selection*, *data transformation*, *data cleaning*, dan *data balancing* (Yoga Siswa T.A, 2023).

2.3 Pembagian Data

Proses pembagian data dilakukan dengan memisahkan dataset menjadi data latih untuk melatih model dan data uji untuk mengevaluasi performa model. Menerapkan teknik *K-Fold Cross-Validation* dengan nilai K=10 untuk menilai performa model machine learning, yang mana pengujian dengan teknik ini mampu memberikan hasil yang lebih baik (Asrol et al., 2021). Teknik ini membagi dataset menjadi 10 bagian yang digunakan secara bergiliran sebagai data latih dan data uji, memberikan penilaian lebih akurat terhadap performa model dalam mengklasifikasikan data banjir di Kota Samarinda.

2.4 Permodelan

Pada tahap ini, akan membahas tentang model yang digunakan dalam penelitian ini. Model klasifikasi yang dipilih adalah algoritma utama *Support Vector Machine* (SVM) dengan kernel RBF, dan *Recursive Feature Elimination* (RFE) sebagai seleksi fitur. Penggunaan kernel RBF dikarenakan kernel ini adalah yang paling umum digunakan untuk mengatasi masalah klasifikasi dan memiliki kinerja yang baik (Listanto et al., 2023).

1. Algoritma SVM kernel RBF dengan *Python*

SVM kernel RBF adalah SVM yang menggunakan fungsi γ (*gamma*) yang menentukan

seberapa jauh pengaruh satu titik data terhadap yang lain. Jika γ besar, setiap titik data hanya mempengaruhi tetangganya yang sangat dekat, sementara jika γ kecil, pengaruhnya lebih luas (Al-Mejibli et al., 2020).

$$K(x, xi) = exp(-\gamma \parallel x - xi \parallel^2) \quad (1)$$

Keterangan:
x, xi : Titik data pertama dan kedua.
K(x, xi) : Menunjukkan seberapa mirip dua titik data x dan xi.
exp : Fungsi eksponensial.
γ : Parameter kontrol pengaruh jarak.
$\parallel x - xi \parallel^2$: Jarak kuadrat *Euclidean* antara titik data x dan xi.

Adapun proses pembuatan modelnya sebagai berikut:
a. Dataset dibagi menjadi data latih dan data uji dengan metode 10 *Fold Cross Validation*.
b. *Import library* SVM, pisahkan fitur (X) dan target (y).
c. Menyiapkan model SVM dengan kernel *Radial Basis Function* (RBF).
d. Melakukan *cross validation* dengan melatih dan menguji model setiap *fold*.
e. Simpan dan tampilkan akurasi serta *confusion matrix* dari setiap *fold*.

2. Seleksi Fitur RFE dengan *Python*

RFE digunakan untuk memilih fitur terbaik pada dataset banjir dengan mengevaluasi setiap fitur secara berulang. Metode ini melibatkan pelatihan model, evaluasi pentingnya fitur, dan penghapusan fitur yang kurang relevan untuk menemukan kombinasi fitur terbaik (Siswa & Wibowo, 2023). Namun, RFE tidak memiliki rumus matematika yang spesifik karena bergantung pada proses iteratif pelatihan model dan peringkat fitur (Guido, 2016). Berikut proses iteratifnya:
a. Memulai dengan semua fitur dan latih modelnya.
b. Rangking semua fitur berdasarkan kepentingannya seperti yang ditentukan oleh model.
c. Eliminasi fitur yang dianggap atau paling tidak penting.
d. Mengulangi proses dengan fitur yang tersisa hingga mencapai jumlah fitur yang diinginkan.

2.5 Evaluasi

Tahap evaluasi merupakan langkah penting setelah pembentukan model untuk mengukur performa dan kualitas data latih. Pengujian dilakukan menggunakan teknik Confusion Matrix, yang digunakan untuk menghitung akurasi dalam data mining (Pratiwi, 2020). Pada penelitian ini, evaluasi performa model diukur menggunakan *accuracy* (akurasi).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\% \quad (2)$$

Keterangan:
TP (*True Positive*) : Jumlah titik data dengan label benar yang berhasil diidentifikasi benar.
TN (*True Negative*) : Jumlah titik data dengan label salah yang berhasil diidentifikasi salah.
FP (*False Positive*) : Jumlah titik data yang sebenarnya salah tetapi diberi label benar.
FN (*False Negative*) : Jumlah titik data yang sebenarnya benar tetapi diberi label salah.

3. Hasil dan Pembahasan

Tahapan ini adalah berisi pemaparan hasil pengujian model dan membahas hasil dari pengujian model tersebut. Setiap tahapan diuraikan di subbab berikutnya.

3.1 Hasil Pengumpulan Data

Dari pengumpulan data yang didapatkan adalah 9 fitur untuk data BPBD dan 11 untuk data BMKG, jumlah data keseluruhannya adalah 1095 *records*.

3.2 Hasil Data Preparation

Setelah pengumpulan data selesai, langkah berikutnya adalah melakukan pra-pemrosesan data, agar data tersebut siap untuk diproses pada tahap permodelan. Adapun uraiannya pada subbab di bawah ini.

3.2.1 Data Integration

Penggabungan data dilakukan dengan mengintegrasikan berbagai jenis data terkait banjir dari BPBD Kota Samarinda dan BMKG, yang mencakup detail kejadian banjir di Kota Samarinda dari tahun 2021 hingga 2023.

Tabel 1. Hasil Penggabungan Data

No	Fitur	Tipe Data
1	Tanggal	*date*
2	Jam Kejadian	*string*
3	Jenis Bencana	*string*
4	Lokasi Wilayah	*string*
5	Luas Area M2	*string*
6	Objek Terkena Bencana	*string*
7	Korban	*numeric*
8	Kerugian	*string*
9	Keterangan	*string*
10	Temperatur-maksimum (Tx)	*numeric*
11	Temperatur-minimum (Tn)	*numeric*
12	Temperatur-rata-rata (Tavg)	*numeric*
13	Kelembaban-rata-rata (RH_avg)	*numeric*
14	Curah-hujan (RR)	*numeric*
15	Lamanya-penyinaran-matahari (ss)	*numeric*
16	Kecepatan-angin-maksimum (ff_x)	*numeric*
17	Arah-angin-maksimum (ddd_x)	*numeric*
18	Kecepatan-angin-rata-rata (ff_avg)	*numeric*
19	Arah-angin-terbanyak (ddd_car)	*numeric*

Data yang digabungkan secara manual bertujuan untuk meninjau bentuknya setelah integrasi. Dari tabel 1, terlihat bahwa tidak semua fitur relevan untuk proses klasifikasi. Fitur seperti jenis bencana, lokasi wilayah, dan lainnya dari data BPBD tidak relevan karena tipe data *string* dan numeriknya tidak sesuai dengan potensi terjadinya banjir.

3.2.2 Data Selection

Data *selection* (pemilihan data) merupakan proses pemilihan fitur atau atribut mana saja yang relevan untuk dianalisis lebih lanjut pada klasifikasi data banjir, kemudian fitur-fitur tersebut yang tidak relevan akan dibuang atau dihilangkan.

Tabel 2. Hasil Data *Selection*

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car	Terjadi-banjir
01/01/2021	23.0	33.2	26.5	88.0	1.8	3.3	4.0	280.0	2.0	W	0
02/01/2021	23.2	30.8	27.1	88.0	7.0	6.4	2.0	140.0	1.0	C	0
…	…	…	…	…	…	…	…	…	…	…	…
30/12/2023	24.2	32.6	28.3	82.0		10.4	4.0	60.0	1.0	NE	0
31/12/2023	24.6	32.4	28.3	84.0	0.0	6.9	4.0	90.0	2.0	E	0

Tabel 2 adalah hasil fitur yang relevan terhadap penyebab banjir dipilih, sedangkan yang tidak relevan dihilangkan. Fitur "jenis bencana" dari data BPBD disesuaikan dengan data BMKG dan diubah menjadi "terjadi banjir" sebagai label. Fitur lain yang terpilih mencakup semua fitur dari data BMKG.

3.2.3 Data Cleaning

Data *cleaning* (pembersihan data) merupakan proses eliminasi atau perbaikan data yang keliru, tidak lengkap, atau tidak konsisten (Dwiasnati & Devianto, 2021).

Tabel 3. Hasil Data *Cleaning*

Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car	Terjadi-banjir
01/01/2021	23.0	33.2	26.5	88.0	1.8	3.3	4.0	280.0	2.0	W	0
02/01/2021	23.2	30.8	27.1	88.0	7.0	6.4	2.0	140.0	1.0	C	0
…	…	…	…	…	…	…	…	…	…	…	…
29/12/2023	23.7	32.7	28.7	77.0	3.5	5.9	5.0	60.0	2.0	NE	0
31/12/2023	24.6	32.4	28.3	84.0	0.0	6.9	4.0	90.0	2.0	E	0

Tabel 3 menunjukkan hasil pembersihan data, di mana baris dengan minimal satu nilai kosong dihapus. Seperti baris tanggal 30 pada tabel 2 dihapus karena nilai terdapat kosong pada kolom RR (Curah-hujan). Karena proses pembersihan ini, jumlah data berkurang dari 1095 menjadi 890.

3.2.4 Data Transformation

Data *transformation* (transformasi data) adalah proses mengubah format atau skala data. Transformasi data menggunakan fungsi *LabelEncoder*. *LabelEncoder* digunakan untuk mengubah format data yang sebelumnya dari

bertipe string menjadi numerik (Sailasya & Kumari, 2021).

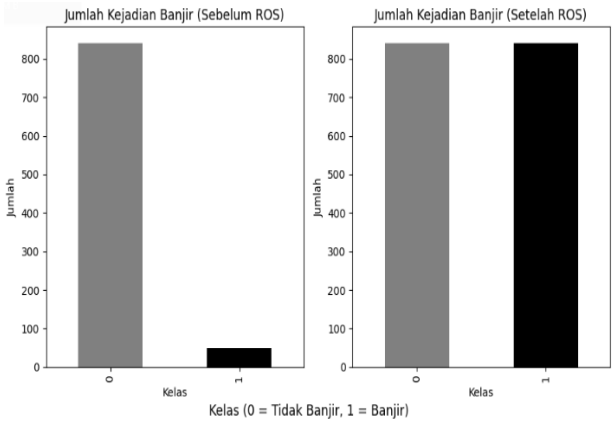
Tabel 4. Data Sebelum Dan Sesudah Diransformasi

No	(sebelum) ddd_car	No	(sesudah) ddd_car
0	W	0	8
1	C	1	0
2	NW	2	4
…	…	…	...
887	C	887	0
888	NE	888	3
889	E	889	1

Ditinjau dari tabel 4, fitur yang diubah dari *string* ke numerik adalah arah angin terbanyak (ddd_car), yang berisi simbol arah mata angin.

3.2.5 Data Balancing

Tahap terakhir dalam pra-pemrosesan data adalah penyeimbangan data, tujuannya untuk menyeimbangkan distribusi dalam dataset guna mencegah bias pada algoritma klasifikasi yang disebabkan oleh ketidakseimbangan jumlah sampel antar kelas (Gumelar et al., 2021). Proses penyeimbangan data menggunakan teknik *Random Oversampling* (ROS), yang menambahkan sampel dari kelas minoritas secara acak hingga jumlahnya setara dengan kelas mayoritas, untuk mencapai distribusi kelas yang lebih seimbang (Khushi et al., 2021).



Gambar 2. Hasil Proses Data *Balancing*

Gambar 2 yang menunjukkan bahwa setelah proses *Random Oversampling* (ROS), jumlah kelas mayoritas dan minoritas menjadi seimbang. Jumlah data untuk kategori tidak terjadi banjir dan terjadi banjir, masing-masing adalah 841 kelas, yang sebelumnya adalah 841 untuk kelas 0 dan 49 untuk kelas 1.

3.3 Hasil Permodelan

Tahapan ini akan menampilkan hasil akurasi dari pengujian model *Support Vector Machine* (SVM) yang dikombinasikan dengan seleksi fitur *Recursive Feature Elimination* (RFE) dan data yang diseimbangkan menggunakan *Random Oversampling* (ROS). Teknik 10 *Fold Cross Validation* digunakan untuk membagi dan mempelajari data. Hasil akurasi dari setiap model akan dievaluasi menggunakan confusion matrix untuk memastikan keakuratan evaluasi model.

3.3.1 Implementasi Algoritma SVM

Tahap pertama ini adalah pengujian model *Support Vector Machine* (SVM) dengan bahasa pemrograman *Python* tanpa menggunakan seleksi fitur RFE. Proses pembagian data menggunakan teknik 10 *Fold Cross Valdiation* untuk memahami dan mempelajari pola data.

Langkah pengujian model ini melibatkan mengimport library SVM dari *scikit_learn*, memisahkan fitur (X) dan target (y) dari data, menyiapkan model SVM dengan kernel RBF, dan menggunakan *StratifiedKFold* dengan 10 lipatan untuk *cross-validation*. Akurasi dan *confusion matrix* dari setiap *fold* disimpan. Proses *cross-validation* dilakukan dengan melatih dan menguji model SVM pada setiap *fold*, menghitung akurasi dan *confusion matrix*, serta menampilkan hasil dari rata-rata keselutuhan *fold*.

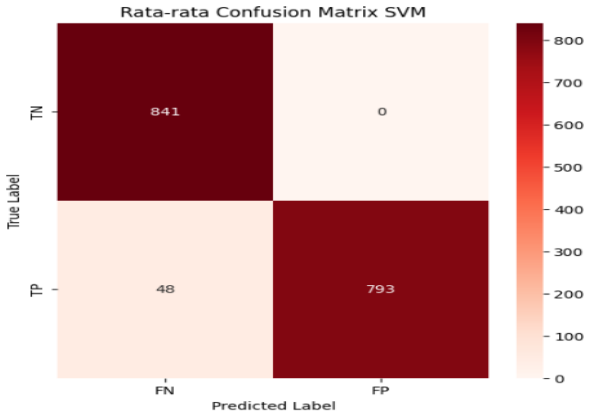
```
Rata-rata dari semua folds:
Accuracy: 0.9714497041420117
Percentage: 97.14%
```

Gambar 3. Hasil Pengujian SVM

Pada Gambar 3, hasilnya menunjukkan bahwa algoritma SVM tanpa seleksi fitur RFE menghasilkan akurasi sebesar 97,14% dari rata-rata keseluruhan *fold*. Untuk membuktikan keakuratan rata-rata hasil akurasinya adalah melakukan perhitungan manual *confusion matrix*.

Pada Gambar 4, dapat dilihat bahwa hasil pengujian SVM dengan menggunakan perhitungan manual *confusion matrix* membuktikan terdapat kesamaan hasil pada hasil pengujian model.

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v7i3.41068



Gambar 4. *Confusion Matrix* SVM

$$Accuracy = \frac{841 + 793}{841 + 793 + 0 + 48} = \frac{1634}{1682} X\ 100\% \quad (2)$$
$$= 97.14\%$$

3.3.2 Implementasi Seleksi Fitur RFE

Pada tahap ini pengujian model *SVM* dengan *Recursive Feature Elimination* (RFE) untuk melakukan reduksi fitur. RFE mereduksi fitur dengan cara mengevaluasi secara berulang-ulang dengan melatih model dan menghapus fitur yang kurang relevan atau yang paling sedikit kontribusinya. Pada tahap ini dilakukan perangkingan terhadap fitur-fitur yang ada pada RFE berdasarkan fitur yang memiliki pengaruh terbesar hingga yang memiliki pengaruh terkecil.

Proses dimulai dengan mengimpor *library* RFE, lalu menyiapkan data dalam variabel X dan y. Model RFE dengan estimator *Support Vector Classifier* (SVC) dibuat untuk memilih 5 fitur terbaik secara iteratif. *StratifiedKFold* digunakan untuk validasi silang dengan 10 lipatan, menjaga keseimbangan kelas, dan mengacak data sebelum pembagian. Sebuah *pipeline* dibuat untuk normalisasi data dengan *StandardScaler*() dan seleksi fitur dengan RFE. Setelah validasi selesai, rata-rata skor kinerja dihitung dan fitur yang paling sering dipilih oleh RFE ditampilkan.

Tabel 5. Hasil Perangkingan RFE

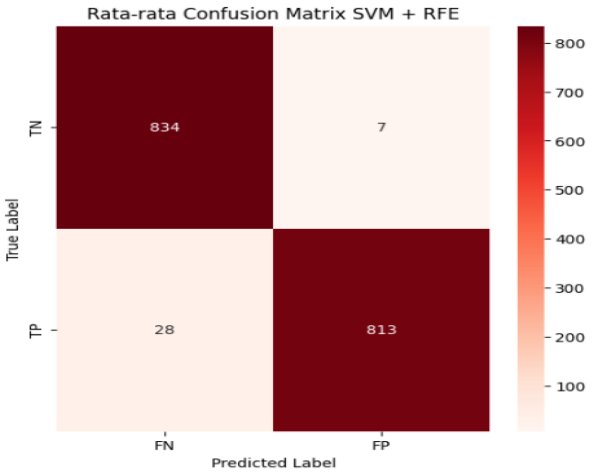
Fitur	Ranking	Nilai Frekuensi RFE
Tn	1	10
Tx	2	10
Tavg	3	10
RH_avg	4	10
ddd_x	5	10

Dari hasil perangkingan terhadap kesepuluh fitur yang ada, RFE berhasil mengidentifikasi 5 fitur terbaik dan yang paling relevan, diantaranya temperatur minimum (Tn), temperatur maksimum (Tx), temperatur rata-rata (Tavg), kelembapan (RH_avg) dan arah angin maksimum (ddd_x), sedangkan fitur-fitur yang lain dihapus.

Setelah mendapatkan fitur-fitur terbaik dari RFE, selanjutnya mengimplementasikan fitur-fitur yang dipilih oleh RFE dan diterapkan ke model SVM untuk melihat seberapa pengaruh fitur-fitur tersebut dalam meningkatkan akurasi yang diperoleh SVM.

```
Rata-rata dari semua folds:
Accuracy: 0.9791913214990137
Percentage: 97.92%
```

Gambar 5. Hasil Pengujian SVM dengan RFE



Gambar 6. *Confusion Matrix* SVM dengan RFE

$$Accuracy = \frac{834 + 813}{813 + 813 + 7 + 28} = \frac{1647}{1682} X\ 100\% \quad (2)$$
$$= 97.92\%$$

Hasilnya menunjukkan bahwa kombinasi algoritma SVM dengan seleksi fitur RFE menghasilkan akurasi sebesar 97,92% dari rata-rata keseluruhan *fold* yang sudah dihitung dengan perhitungan manual *confusion matrix*.

3.4 Pembahasan

Pemilihan fitur oleh RFE pada dataset banjir Kota Samarinda mengidentifikasi lima fitur terbaik: Tn, Tx, Tavg, RH_avg, dan ddd_x, masing-masing dengan nilai frekuensi pemilihan sebesar 10. Fitur-fitur ini memberikan hasil akurasi yang beragam. Penelitian ini juga membandingkan efektivitas RFE dengan penelitian serupa oleh (Pratama et al., 2022), yang menggunakan data

curah hujan dari BMKG. Pratama mengidentifikasi tiga fitur terbaik yaitu Tavg, ss, dan Tn, dengan model SVM dan teknik *resampling* dari *Random over-undersampling*. Hasil akurasinya bervariasi dengan teknik *splitting* data, Mulai dari *splitting* 90:20 yang menghasilkan sebesar 74% dari 74%, diikuti 80:20 sebesar 79% sebelumnya 74%, 70:30 sebesar 79% sebelumnya 77% dan 60:40 sebesar 59% yang sebelumnya 60%.

Tabel 6. Persamaan Hasil Fitur Terpilih *RFE* Penelitian ini dengan Penelitian Lain

Fitur Terpilih RFE dengan *Oversampling* ROS	**Penlitian Pratama SVM + RFE**
Temperatur-minimum (Tn)	
Temperatur-maksimum (Tx)	✓
Temperatur-rata-rata (Tavg)	
Kelembapan (RH_avg)	✓
Arah-angin-maksimum (ddd_x)	

Pada tabel 6 hasil seleksi fitur dari penelitian ini yang sama-sama menggunakan teknik *resampling*, menunjukkan kesamaan dengan hasil penelitian lain. Fitur-fitur yang terpilih, yaitu Tn dan Tavg, sama-sama ditemukan dalam penelitian tersebut. Hal ini menandakan bahwa kesamaan fitur-fitur tersebut dapat mempengaruhi klasifikasi secara signifikan.

Penerapan seleksi fitur RFE pada algoritma SVM meningkatkan akurasi klasifikasi data banjir di Kota Samarinda dari 97,14% menjadi 97,92%. RFE berhasil memilih fitur paling relevan, mengurangi dimensi data, dan meningkatkan performa model. Teknik *Random Oversampling* (ROS) membantu menyeimbangkan distribusi kelas pada dataset tidak seimbang, sehingga model SVM mampu belajar lebih baik dan mengurangi bias terhadap kelas mayoritas. Kombinasi RFE dan ROS pada SVM secara keseluruhan meningkatkan akurasi dan efisiensi klasifikasi dengan peningkatan akurasi sebesar 0,78%.

Tabel 7 menunjukkan variasi akurasi dari setiap model, termasuk model SVM dan kombinasinya dengan RFE. Beberapa model menunjukkan peningkatan, penurunan, atau tetap pada akurasi. Sementara, Tabel 8 menunjukkan rata-rata akurasi dari kedua model tersebut, yang mengalami peningkatan secara keseluruhan. Hasil ini menunjukkan bahwa penggunaan seleksi fitur (RFE) dapat memengaruhi akurasi secara positif dibandingkan dengan tidak menggunakan seleksi fitur.

Tabel 7. Perbandingan Hasil Akurasi Kedua Model

Fold	**SVM**	**SVM + RFE**	**Perubahan Hasil Model**	**Status**
1	97.04%	97.04%	0%	Tetap
2	99.41%	98.82%	-0.59%	Turun
3	93.45%	95.24%	1.79%	Naik
4	98.21%	97.62%	-1.59%	Turun
5	96.43%	97.62%	1.19%	Naik
6	99.40%	100%	0.60%	Naik
7	95.83%	95.83%	0.%	Tetap
8	95.83%	100%	4.17%	Naik
9	97.02%	97.62%	0.60%	Naik
10	98.81%	99.40%	0.59%	Naik

Tabel 8. Perbandingan Rata-rata Hasil Akurasi Kedua Model

Rata-rata Akurasi	SVM	SVM + RFE	Perubahan Hasil Model	Status
	97,14%	97,92%	0,78%	Naik

4. Kesimpulan

Pemilihan fitur menggunakan RFE berhasil mengidentifikasi lima fitur terbaik yang meningkatkan akurasi klasifikasi data banjir di Kota Samarinda, dan fitur-fiturnya adalah temperatur minimum (Tn), temperatur maksimum (Tx), temperatur rata-rata (Tavg), kelembapan (RH_avg), dan arah angin maksimum (ddd_x). Penerapan RFE dan teknik *oversampling* ROS pada algoritma SVM terbukti efektif, meningkatkan akurasi dari 97,14% menjadi 97,92%, dengan peningkatan sebesar 0,78%.

5. Saran

Selain RFE, penelitian selanjutnya disarankan mencoba metode seleksi fitur *wrapper* lain seperti *Sequential Feature Selection* (SFS) untuk hasil yang lebih optimal. Juga, teknik *resampling* lain seperti *Random Undersampling* (RUS) atau *Synthetic Minority Oversampling Technique* (SMOTE) dapat membantu meningkatkan akurasi model. Untuk penelitian selanjutnya, mencoba kernel SVM lain seperti *linear*, *sigmoid*, atau *polynomial*, dan lebih memperhatikan nilai parameter *gamma* dan *C* (*Cost*) pada kernel RBF untuk mencari hasil yang lebih optimal. Selain itu, penggunaan algoritma optimasi seperti *Particle Swarm Optimization* (PSO), *Bayesian Optimization* dapat meningkatkan kinerja model secara keseluruhan.

Referensi

Ahmmed, M. R., Monir, J., & Khushbu, S. A. (2022). Analysis of Flood Risk Prediction Using Different Machine Learning Classifiers: A Study of Predicting Flood Risk in Rural Areas, Bangladesh. *2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*, 1–6. https://doi.org/10.1109/ICCCNT54827.2022.9984449

Al-Mejibli, I. S., Alwan, J. K., & Abd, D. H. (2020). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*, *10*(5), 5497–5506. https://doi.org/10.11591/IJECE.V10I5.PP5497-5506

Andrean, F. W. (2024). *4.940 Bencana Terjadi di Indonesia Sepanjang 2023*. https://indonesiabaik.id/infografis/4940-bencana-terjadi-di-indonesia-sepanjang-2023

Asrol, M., Papilo, P., & Gunawan, F. E. (2021). Support Vector Machine with K-fold Validation to Improve the Industry's Sustainability Performance Classification. *Procedia Computer Science*, *179*(2020), 854–862. https://doi.org/10.1016/j.procs.2021.01.074

BPS. (n.d.). *Jumlah Desa/Kelurahan yang Mengalami Bencana Alam1 menurut Kecamatan di Kota Samarinda*. Retrieved April 14, 2024, from https://samarindakota.bps.go.id/indicator/153/147/1/jumlah-desa-kelurahan-yang-mengalami-bencana-alam-sup-1-sup-menurut-kecamatan-di-kota-samarinda.html

Dilla Evitasari, Y., Pranoto, W. J., & Adzmi Verdikha, N. (2023). Evaluasi Support Vector Machine Dengan Optimasi Metode Genetic Algorithm Pada Klasifikasi Banjir Kota Samarinda Evaluation Support Vector Machine With Optimization Genetic Algorithm Method On Flood Classification In Samarinda. *Jurnal Sains Komputer Dan Teknologi Informasi*, *6*(1), 49–53.

Duwal, S., Liu, D., & Pradhan, P. M. (2023). Flood susceptibility modeling of the Karnali river basin of Nepal using different machine learning approaches. *Geomatics, Natural Hazards and Risk*, *14*(1). https://doi.org/10.1080/19475705.2023.2217321

Dwiasnati, S., & Devianto, Y. (2021). Optimasi Prediksi Bencana Banjir menggunakan Algoritma SVM untuk penentuan Daerah Rawan Bencana Banjir. *Prosiding SISFOTEK*, 202–207. http://seminar.iaii.or.id/index.php/SISFOTEK/article/view/283

Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. *Jurnal Infortech*, *2*(1), 96–101. https://doi.org/10.31294/infortech.v2i1.8079

Fitrianah, D., Gunawan, W., & Puspita Sari, A. (2022). Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction. *Februari*, *21*(1), 1–11.

Gauhar, N., Das, S., & Moury, K. S. (2021). Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm. *International Conference on Robotics, Electrical and Signal Processing Techniques*, *January*, 357–361. https://doi.org/10.1109/ICREST51555.2021.9331199

Guido, S. (2016). *Introduction to Machine Learning with Python* (D. Schanafelt (ed.); October 20). O'Reilly Media, Inc.

Gumelar, G., Ain, Q., Marsuciati, R., Agustanti Bambang, S., Sunyoto, A., & Syukri Mustafa, M. (2021). Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance. *SISFOTEK : Sistem Informasi Dan Teknologi*, 250–255.

Huang Kendrew, P. P. E. (2022). *Support Vector Machine Algorithm*. Binus. https://sis.binus.ac.id/2022/02/14/support-vector-machine-algorithm/

Idris, M., Adam, R. I., Brianorman, Y., Munir, R., & Mahayana, D. (2022). Kebenaran dalam Perspektif Filsafat Ilmu Pengetahuan dan Implementasi dalam Data Science dan Machine Leaning. *Jurnal Filsafat Indonesia*, *5*(2), 173–181. https://doi.org/10.23887/jfi.v5i2.42207

Khan, T. A., Alam, M., Ahmed, S. F., Shahid, Z., & Mazliham, M. S. (2019). A Factual Flash Flood Evaluation using SVM and K-NN. *ICETAS 2019 - 2019 6th IEEE International Conference on Engineering, Technologies and Applied Sciences*. https://doi.org/10.1109/ICETAS48360.2019.9117424

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, *9*, 109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399

Listanto, F., Fatchan, M., Hadikristanto, W., Studi, P., Informatika, T., Teknik, F., Pelita, U., & Bekasi, B. (2023). Prediksi Defect Produk Casting Dengan Algoritma SVM Berbasis RBF dan Linier. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *5*(2), 109–119.

M. Adib Al Karomi, Abdul Kharis, I. (2019). Optimasi Algoritma Naive Bayes Dengan Information Gain Ratio Untuk Menangani Dataset Berdimensi Tinggi. *Seminar Nasional Edusaintek*, 37–43.

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

Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (Svm) Dan Recursive Feature Elimination (Rfe). *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *7*(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675

Pratiwi, B. P. (2020). Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi WSN Menggunakan Confusion Matrix. *Jurnal Informatika UPGRIS*, *6*(2), 66–75.

Pristyanto, Y. (2019). Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset. *Jurnal Teknoinfo*, *13*(1), 11. https://doi.org/10.33365/jti.v13i1.184

Puspasari, R. L., Yoon, D., Kim, H., & Kim, K. W. (2023). Machine Learning for Flood Prediction in Indonesia: Providing Online Access for Disaster Management Control. *Economic and Environmental Geology*, *56*(1), 65–73. https://doi.org/10.9719/eeg.2023.56.1.65

Rahman, M. A., Akter, A., Richi, F. S., Shoud, A., & Ahmed, T. (2023). A Comparative Study of Undersampling and Oversampling Methods for Flood Forecasting in Bangladesh using Machine Learning. *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, *December*. https://doi.org/10.1109/ICCCNT56998.2023.10306368

Ramadhan, N. G., Khoirunnisa, A., Kurnianingsih, & Hashimoto, T. (2023). A Hybrid ROS-SVM Model for Detecting Target Multiple Drug Types. *International Journal on Informatics Visualization*, *7*(3), 794–800. https://doi.org/10.30630/joiv.7.3.1171

Rifqi Fitriadi, & Deni Mahdiana. (2023). Systematic Literature Review of the Class Imbalance Challenges in Machine Learning. *Jurnal Teknik Informatika (Jutif)*, *4*(5), 1099–1107. https://doi.org/10.52436/1.jutif.2023.4.5.970

Rustam, Z., Syarifah, M. A., & Siswantining, T. (2019). Recursive Particle Swarm Optimization (RPSO) schemed Support Vector Machine (SVM) Implementation for Microarray Data Analysis on Chronic Kidney Disease (CKD). *IOP Conference Series: Materials Science and Engineering*, *546*(5). https://doi.org/10.1088/1757-899X/546/5/052077

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 539–545. https://doi.org/10.14569/IJACSA.2021.0120662

Salsadilla, V., Permana, I., Jazman, M., & Afdal, M. (2023). *Determining the Final Project Topic Based on the Courses Taken by Using Machine Learning Techniques*. *3*(October), 188–198.

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

Siswa, T. A. Y., & Wibowo, R. P. (2023). Komparasi Metode Seleksi Fitur Dalam Prediksi Keterlambatan Pembayaran Biaya Kuliah. *Teknika*, *12*(1), 73–82. https://doi.org/10.34148/teknika.v12i1.601

Tarasova, L., Merz, R., Kiss, A., Basso, S., Blöschl, G., Merz, B., Viglione, A., Plötner, S., Guse, B., Schumann, A., Fischer, S., Ahrens, B., Anwar, F., Bárdossy, A., Bühler, P., Haberlandt, U., Kreibich, H., Krug, A., Lun, D., … Wietzke, L. (2019). Causative classification of river flood events. *Wiley Interdisciplinary Reviews: Water*, *6*(4), 1–23. https://doi.org/10.1002/wat2.1353

Thakkar, A., & Lohiya, R. (2021). Attack classification using feature selection techniques: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 1249–1266. https://doi.org/10.1007/s12652-020-02167-9

Uddin, M. J., Ahamad, M. M., Hoque, M. N., Walid, M. A. A., Aktar, S., Alotaibi, N., Alyami, S. A., Kabir, M. A., & Moni, M. A. (2023). A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh. *Information (Switzerland)*, *14*(7), 1–19. https://doi.org/10.3390/info14070376

Yoga Siswa T.A. (2023). *Data Mining: Mengupas Tuntas Analisis Data Dengan Metode Klasifikasi Hingga Deployment Aplikasi Menggunakan Python*. Umkt Press : Universitas Muhammadiyah Kalimantan Timur.