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 (1289-1299) **DOI: 10.32493/jtsi.v7i3.41587**

Model Optimasi KNN-PSORF dalam Menangani High Dimensional Data Banjir Kota Samarinda

Anggiq Karisma Aji Restu[1], Taghfirul Azhima Yoga Siswa[2*], Wawan Joko Pranoto[3]

Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia 75124
e-mail : [1]2011102441089@umkt.ac.id, [2*]tay758@umkt.ac.id, [3]wjp337@umkt.ac.id

Submitted Date: July 01st, 2024 Reviewed Date: July 03rd, 2024
Revised Date: July 20th, 2024 Accepted Date: July 24th, 2024

Abstract

Floods are a natural phenomenon that frequently occurs in Indonesia, including in Samarinda City which has faced flood issues over the past three years, affecting thousands of homes and around 27,000 residents. Predicting flood disasters requires machine learning technology using data mining classification methods. However, classification processes often encounter issues related to high-dimensional data, which can lead to overfitting and class imbalance, thereby biasing dominant classes while neglecting minority classes. This research aims to enhance classification accuracy in Samarinda City's flood data using the K-Nearest Neighbor (KNN) algorithm combined with Relief feature selection and Particle Swarm Optimization (PSO) optimization. The validation method employed is 10-fold cross-validation, with performance evaluation using a confusion matrix. Data sourced from Samarinda City's Disaster Management Agency (BPBD) and Meteorology, Climatology, and Geophysics Agency (BMKG) spans from 2021 to 2023, comprising 19 features and a total of 1095 records. Relief feature selection identified four crucial features: maximum wind direction, wind speed, average wind speed, and maximum wind speed direction. Average evaluations with k values of 3, 5, 7, 11, 13, and 15 demonstrate that Relief feature selection and PSO optimization effectively enhance accuracy in the K-Nearest Neighbor algorithm for flood data, with KNN and PSO yielding improvements of 2-5%. Relief feature selection alone improves accuracy by 1-2%, while combining Relief with PSO provides a 2-5% enhancement. The combined KNN, Relief, PSO model is expected to deliver optimal performance in classifying Samarinda City's flood data.

Keywords: K-Nearest Neighbor; Relief; Flood; 10-Fold Cross-Validation; Classification

Abstrak

Banjir adalah fenomena alam yang sering terjadi di Indonesia, termasuk di Kota Samarinda yang mengalami masalah banjir dalam tiga tahun terakhir dengan dampak ribuan rumah sebanyak 27.000 jiwa terkena banjir. Untuk memprediksi bencana banjir dibutuhkan teknologi *machine learning* menggunakan metode klasifikasi *data mining*. Namun, pada proses klasifikasi seringkali terjadi permasalahan yang berkaitan dengan data berdimensi tinggi ini dapat menyebabkan *overfitting* dan ketidakseimbangan kelas yang menyebabkan bias pada kelas yang dominan dengan mengabaikan kelas minoritas. Penelitian ini bertujuan untuk meningkatkan nilai akurasi klasifiikasi pada data banjir Kota Samarinda menggunakan algoritma *K-Nearest Neighbor* (KNN) yang dikombinasikan seleksi fitur *Relief* dan optimasi *Particle Swarm Optimization* (PSO). Metode validasi yang digunakan adalah *10-fold cross-validation*, sementara evaluasi kinerja model dilakukan menggunakan *confusion matrix*. Data yang digunakan diperoleh dari BPBD dan BMKG Kota Samarinda pada rentang tahun 2021-2023, dengan 19 fitur dan total 1095 *record.* Hasil seleksi fitur *Relief* didapatkan empat fitur penting, yaitu arah angin maksimum, kecepatan angin, kecepatan angin rata-rata, dan arah angin maksimum. Evaluasi rata-rata dengan nilai k=3, k=5, k=7, k=11, k=13, dan k=15 menunjukkan penerapan seleksi fitur *Relief* dan optimasi PSO, efektif dalam meningkatkan akurasi pada algoritma *k-Nearest Neighbor* pada data banjir dengan hasil akurasi KNN dan PSO memberikan peningkatan sebesar 2-5%, KNN dengan seleksi fitur *Relief* memberikan peningkatan sebesar

1-2% dan KNN dengan kombinasi *Relief* dan PSO memberikan peningkatan sebesar 2-5%. Kombinasi model KNN, *Relief*, PSO diharapkan dapat memberikan peforma yang optimal dalam klasifikasi data banjir Kota Samarinda.

Kata kunci: Klasifikasi ; K-Nearest Neighbor; Seleksi fitur; Banjir; Optimasi

1 Pendahuluan

Banjir adalah fenomena alam yang sering melanda indonesia. Menurut Data Informasi Bencana Indonesia (DIBI), dalam kurun waktu tiga tahun terakhir, tercatat sebanyak 4580 kejadian banjir di Indonesia dan Jumlah tertinggi terjadi pada tahun 2020, mencapai 1531 kejadian, menjadi yang terbanyak dalam hampir satu dekade terakhir (Databoks, 2023). Kota Samarinda saat ini sedang dilanda permasalahan banjir yang cukup parah. Selain itu terdapat dua sub wilayah yang juga mempunyai masalah banjir yaitu DAS Karang Asam Besar (9,65 km2) dan DAS Karang Asam Kecil (16,25 km2) (Purwanto, 2020). Pada tahun 2020 banjir terjadi pada 10 kecamatan, 4 kelurahan dan menyebabkan sebanyak 27.000 jiwa terkena dampak banjir yang merugikan masyarakat (Ernawati et al., 2021).

Klasifikasi banjir berdasarkan penyebabnya dapat membantu memperbaiki prediksi banjir, pemahaman perubahan kejadian dan tingkat keparahan banjir (Tarasova et al., 2019). Oleh karena itu, perlu dilakukan evaluasi perbaikan akurasi dengan teknologi *machine learning* seperti metode klasifikasi data mining. *Data mining* merupakan proses yang dilakukan dengan penggabungan tekni analisis data untuk memperoleh pola penting pada suatu data (Tarigan et al., 2022).

Ketidakseimbangan kelas (*class imbalance*) dalam mechine learning menyebabkan beberapa masalah karena menganggap bahwa data didistribusikan secara rata, jadi ketika ada kelas yang tidak seimbang, mesin akan lebih bias pada kelas yang dominan dengan mengabaikan kelas minoritas, sehingga pada kelas mayoritas lebih cenderung menunjukkan nilai akurasi yang lebih baik (Yoga Siswa, 2023).

Data berdimensi tinggi (High dimensional) pada kumpulan data menyebabkan beberapa masalah dalam *Machine Learning*. Pertama, sulit bagi model pembelajaran untuk mencapai performa optimal karena semakin banyak fitur yang digunakan. Kedua, jumlah data yang besar dapat menyebabkan *overfitting* karena banyaknya konfigurasi karakteristiknya. Ketiga, data dengan dimensi yang besar susah untuk diproses secara komputasi (*computationally expensive*) (Ariyoga, 2022).

Dari beberapa penelitian yang pernah dilakukan sebelumnya dalam klasifikasi data banjir, Algoritma *k-nearest neighbour* (KNN) tanpa seleksi fitur dinilai memiliki performa lebih unggul dalam klasifikasi data banjir dengan akurasi sebesar 94,91% (Gauhar et al., 2021), dibandingkan dengan akurasi *Random Forest* sebesar 71,3 % (Hossain & Zeyad, 2023), *Support Vector Machine* 52,71% (Evitasari et al., 2023), *Naive Bayes* 82,10% (Daniel et al., 2023), 89,23% (Vafakhah et al., 2020). Sehingga pada penelitian ini akan menggunakan algoritma KNN untuk Klasifikasi pada data banjir Kota Samarinda.

Berdasarkan penelitian sebelumnya yang menerapkan algoritma KNN pada data banjir dengan dimensi data yang tinggi menunjukkan akurasi yang lebih rendah yaitu 88,94%, Ditemukan adanya permasalahan pada penelitian dengan data *High Dimensional* yang dapat menurunkan akurasi (Cumel, David Zamri, Rahmaddeni, 2022). Kemudian, pada *dataset* banjir yang akan digunakan pada penelitian ini terdapat 19 fitur, dimana untuk jumlah fitur yang tinggi seringkali merujuk pada data berdimensi tinggi. Oleh karena itu, untuk menangani masalah tersebut diterapkan *feature selection* untuk mengidentifikasi fitur-fitur yang paling relevan, dengan tujuan untuk meningkatkan performa yang lebih baik.

Pendekatan yang digunakan dari penelitian sebelumnya dalam mengatasi dimensi tinggi menggunakan *feature selection Relief* terbukti bisa memberikan peningkatan akurasi sebesar 5-10%. Dengan demikian, penerapan seleksi fitur menggunakan *Relief* dilakukan karena metode tersebut telah terbukti mampu memberikan peningkatan akurasi model KNN dari akurasi awal pada penelitian sebelumnya (Yahdin et al., 2021; Yusra et al., 2021; Kemal Musthafa Rajabi et al., 2023; Abdulrazaq et al., 2021).

Kemudian, ketidakseimbangan kelas terjadi pada dataset banjir dimana pada data yang

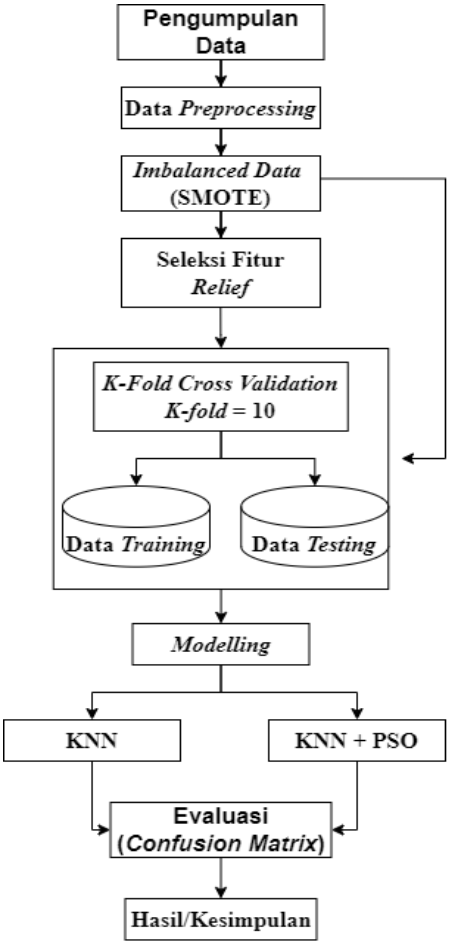
diperoleh dari BMKG dan BPBD, data yang terjadi banjir berjumlah 49 data sedangkan yang tidak banjir berjumlah 841 data, maka pada penelitian ini juga akan menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Berdasarkan penelitian sebelumnya, metode SMOTE pernah digunakan dalam mengatasi ketidakseimbangan kelas pada dataset banjir dan dianggap dapat memberikan peningkatan akurasi terhadap model klasifikasi sebesar 0.21-10% setelah diuji dengan algoritma selain KNN (Nawi et al., 2020; Priscillia et al., 2022; Nursyahfitri et al., 2022; Razali et al., 2020). *Particle Swarm Optimization* (PSO) akan di gunakan sebagai metode optimasi dalam dalam mengoptimalkan peforma algortima KNN pada penelitian data banjir, Seperti ditunjukkan oleh peneltian sebelumnya yang dapat mengoptimalkan peforma algortima, dimana penerapan optimasi tersebut dapat memberikan peningkatkan akurasi sebesar 3-11% (Dwiasnati & Yudo Devianto, 2022; Faldi et al., 2023; Arora et al., 2021). Kombinasi model KNN, PSO, *Relief* digunakan dengan tujuan Menentukan atribut yang berpengaruh pada algortima *k-nearest* neighbour (KNN) terhadap dataset banjir Kota Samarinda dan Mengevaluasi hasil kinerja algoritma *k-nearest neighbors* (KNN).

2 Metodologi Penelitian

Setiap penelitian memiliki beberapa tahapan dalam pelaksanaan penelitian, adapun tahapan yang akan dilakukan seperti pada Gambar 1. Setiap tahapan dijelaskan pada subbab berikutnya.

2.1 Pengumpulan Data

Penelitian ini akan menggunakan data banjir Kota Samarinda dari rentang tahun 2021-2023. Data yang didapatkan dari BPBD (Badan Penanggulangan Bencana Daerah) dan BMKG (Badan Meteorologi, Klimatologi, dan Geofisika), menunjukkan bahwa data tersebut terdiri atas *1095 dataset.*



Gambar 1 Tahapan Penelitian

2.2 Preprocesing Data

Pada tahap ini dilakukan Preproseing data untuk mengolah data terlebih dahulu dengan menghilangkan data yang tidak diperlukan sebelum masuk ke tahapan selanjutnya, Tahapan dalam *preprocesing data* adalah sebagai berikut:

1. *Data Integration*

   Pada tahap ini akan menggabungkan data dari sumber yang berbeda menjadi satu *set data*. Data yang akan digabungkan bersumber dari Badan Penanggulangan Bencana Daerah (BPBD) dan Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Setelah kedua data digabungkan maka didapatkan total 19 atribut dan 1 kelas.

Tabel 1. Hasil dari Data Integration

No	Atribut	Tipe Data	Keterangan
1	Tanggal	*date*	Waktu Kejadian
2	*(Tn)*	*numeric*	Temperatur minimum (°C)
3	*(Tx)*	*numeric*	Temperatur maksimum (°C)

No	Atribut	Tipe Data	Keterangan
4	*(Tavg)*	*numeric*	Temperatur rata rata (°C)
5	*(RH_avg)*	*numeric*	Kelembaban rata rata (%)
6	*(RR)*	*numeric*	Curah-hujan (mm)
7	*(ss)*	*numeric*	Lamanya penyinaran matahari (jam)
8	*(ff_x)*	*numeric*	Kecepatan angin maksimum (m/s)
9	*(ff_avg)*	*numeric*	Kecepatan angin rata-rata (m/s)
10	*(ddd_x)*	*numeric*	Arah angin maksimum (°)
11	*(ddd_car)*	*string*	Arah angin terbanyak (°)
12	Jam kejadian	*time*	Jam Terjadinya Bencana
13	Lokasi wilayah	*string*	Wilayah Terjadinya Bencana
14	Luas area M2	*numeric*	Luas area yang terdampak
15	Objek terkena bencana	*string*	Fasilitas Yang terdampak
16	Korban	*numeric*	Jumlah korban bencana
17	Kerugian	*numeric*	Nominal kerugian
18	Keterangan	*numeric*	Detail kejadian
19	Terjadi banjir	*string*	Ya/Tidak(class)

2. *Data Selection*

Tahap ini dilakukan proses pemilihan atribut yang berpengaruh terhadap penyebab banjir, Hasil *data selection* diperoleh sebanyak 11 Atribut yang terpilih dan 1 atribut sebagai target atau kelas.

Tabel 2. Hasil Data Selection

No	Atribut Awal	Atribut Hasil Seleksi	Keterangan
1	*Tgl*	Tanggal	Date
2	*Tn*	Temperatur-minimum	Atribut
3	*Tx*	Temperatur-maksimum	Atribut
4	*Tavg*	Temperatur rata-rata	Atribut
5	*RH_avg*	Kelembaban	Atribut
6	*RR*	Curah-hujan	Atribut
7	*ss*	Lamanya-penyinaran-matahari	Atribut
8	*ff_x*	Kecepatan-angin	Atribut
9	*ddd_x*	Arah-angin-maksimum	Atribut
10	*ff_avg*	Kecepatan-angin-rata-rata	Atribut
11	*ddd_car*	Arah-angin-terbanyak	Atribut
12	Terjadi Banjir	Terjadi-Banjir	Class

3. *Data Cleaning*

Pada tahap ini dilakukan proses data cleaning untuk menghapus nilai kosong pada dataset banjir dengan jumlah awal 1095 baris. Setelah melalui proses *data cleaning*, jumlahnya berkurang menjadi 890 baris. Dengan demikian, terdapat 205 baris data kosong yang telah dihapus selama proses pembersihan data, seperti yang ditunjukkan pada gambar 2 dan gambar 3.

```
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
dtype: int64
Jumlah data kosong:  245
```

Gambar 2. Nilai kosong atribut sebelum data cleaning

Selanjutnya dilakukan pemeriksaan untuk memastikan bahwa tidak ada lagi data kosong pada data banjir.

```
Jumlah nilai yang hilang setelah pembersihan:
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
dtype: int64
```

Gambar 3. Jumlah Nilai Kosong Tiap Kolom Setelah Pembersihan

4. *Transformation Data*

Pada tahap ini, dilakukan pengubahan terhadap data kategorikal menjadi *numeric.* Data yang diubah pada tahap ini meliputi ‘Arah-angin-terbanyak (ddd_car)’ dan ‘terjadi_banjir’.

Tabel 3 *Data Sebelum diTransformasi*

No	Arah angin terbanyak	Terjadi banjir
0	W	Tidak Banjir
1	C	Tidak Banjir
2	NW	Banjir
…	...	...
1092	NE	Tidak Banjir
1094	E	Tidak Banjir

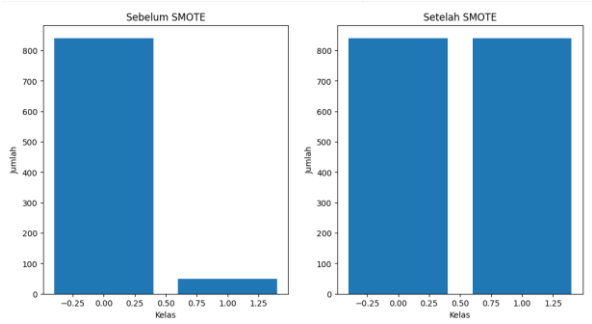
Pada Tabel 4, menunjukka bahwa data yang sebelumnya berbentuk kategorikal telah diubah menjadi numeric untuk memudahkan proses klasifikasi.

Tabel 4. Data Setelah diTransformasi

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

5. *Balancing Data*

Pada tahap ini dilakukan proses untuk menyeimbangkan distribusi kelas atau label pada dataset. Hal ini sering kali diperlukan dalam konteks masalah klasifikasi di mana terdapat ketidakseimbangan yang signifikan antara jumlah sampel yang termasuk dalam setiap kelas atau label. Proses penyeimbangan data dilakukan pada dataset banjir yang diperoleh dari BMKG dan BPBD Karena didapatkan ketidakseimbangan kelas. Dalam menangani permasalahan tersebut, metode *oversampling* SMOTE akan digunakan untuk menyeimbangkan jumlah sampel antara kelas minoritas dan kelas mayoritas dalam dataset banjir.



Gambar 4. sebelum dan sesudah penerapan SMOTE

2.3 Feature Selection

Relief sebagai metode seleksi fitur yang akan melakukan perangkingan fitur-fitur dalam dataset berdasarkan hasil skor kepentingan (*importance score*), di mana fitur dengan skor kepentingan yang lebih tinggi memiliki pengaruh lebih signifikan terhadap hasil klasifikasi. Sebaliknya, fitur dengan skor kepentingan yang rendah memiliki pengaruh yang lebih kecil terhadap klasifikasi dan mungkin kurang sesuai untuk model tersebut. Adapun rumus untuk mencari nilai bobot fitur menggunakan *Relief* seperti berikut:

a. Inisialisasi nilai awal seluruh bobot fitur = 0 dan menentukan jumlah iterasi.
b. Memilih sebuah data yang akan dijadikan sebagai titik acak atau titik pusat.
c. Mencari *miss* dan *hitter* dekat dengan cara menghitung jarak antara titik pusat dengan data yang memiliki kelas yang sama.
d. Update bobot untuk setiap fitur-fitur dengan data kategori yanag dihitung menggunakan Persamaan 1.

$$diff(A, Ri, HM) = \{0; value(A, Ri) = value(A, HM) \\ 1; otherwise \quad (1)$$

e. Sedangkan, fitur dengan data numeric akan dihitung menggunakan permasamaan 2.

$$diff(A, Ri, HM) = \frac{\text{value(A,Ri)} - \text{value(A,HM)}|}{\max(A) - \min(A)} \quad (2)$$

f. Sehingga rumus perbaruan bobot akan dihitung menggunakan persamaan 3

$$W[A] = W[A] - diff(A, Ri, H)m + \\ diff(A, Ri, M)m \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

2.4 Pembagian Data

Dalam penelitian klasifikasi menggunakan algoritma KNN, dataset dibagi menjadi dua bagian utama, yaitu data latih dan data uji. Untuk memastikan evaluasi model yang akurat dan konsisten, pengujian akan menggunakan *metode k-fold cross validation* (Nabila et al., 2021). Metode *k-fold cross validation* yang akan digunakan pada penelitian ini dengan nilai k = 10, yang menunjukkan bahwa eksperimen akan dijalankan sepuluh kali. dan setiap bagian bergantian menjadi data latih dan uji.

2.5 Modelling

a. Permodelan KNN

Permodelan ini akan *menggunakan K-Nearest Neighbor* (KNN) dalam mencari hasil evaluasi confusion matrix dan juga prediksi nilai akurasi. Kemudian, dataset dibagi menjadi data training dan data testing yang dikhususkan untuk model dalam mempelajari pola data dengan menggunakan metode *10-Fold Cross-Validation.* Adapun rumus yang biasa digunakan dalam melakukan perhitungan algoritma K-Nearest Neighbor (KNN) adalah sebagai berikut :

$$\sqrt{\sum_{i=1}^{p} (\alpha_k - b_k)^2} \quad (4)$$

Keterangan :

α_K : Sampel data
b_k : Data uji atau Data testing
p : Dimensi data
i : variable data

b. Permodelan KNN + PSO

Pemodelan ini menerapkan optimasi *Particle swarm optimization* (PSO), dengan tujuan untuk meningkatkan kinerja model secara keseluruhan dengan menyesuaikan parameter yang diperlukan. Dalam konteks algoritma KNN, parameter yang dioptimalkan meliputi jumlah jumlah tetangga (*n_neighbors*), kedalaman maksimum pohon (*max_depth*), jumlah minimum sampel untuk memisah *internal node* (*min_samples_split*), dan jumlah minimum sampel di *leaf node* (*min_samples_leaf*).

c. Permodelan KNN + *Relief*

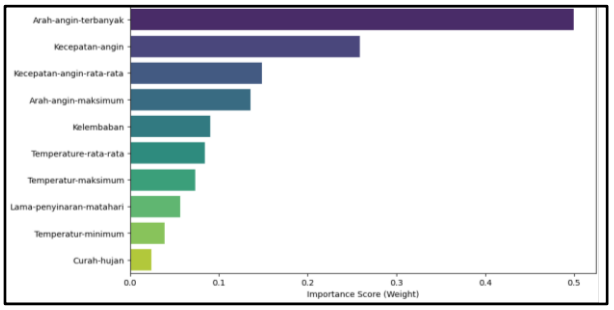
Dalam penelitian ini, algoritma *Relief* digunakan untuk seleksi fitur. Pada tahap ini, fitur-fitur yang ada akan diurutkan berdasarkan pengaruhnya terhadap hasil prediksi, dimulai dari fitur yang memiliki pengaruh terbesar hingga fitur yang memiliki pengaruh terkecil atau bahkan tidak berpengaruh sama sekali.

Tahapannya meliputi mengimpor modul relief dari *library sklearn* dan modul pandas. Modul *pandas* digunakan untuk mengubah dataset menjadi bentuk *dataframe*. Setelah itu, fitur-fitur akan dirangking berdasarkan skor yang diperoleh dari feature selection *Relief* seperti yang ditunjukkan pada tabel 5.

Tabel 5. Perangkingan Fitur berdasarkan score dan Fitur yang akan digunakan

Atribut	*importance score*	Hasil
Arah-angin-terbanyak	0.499401	Digunakan
Kecepatan-angin	0.259026	Digunakan
Kecepatan-angin-rata-rata	0.148464	Digunakan
Arah-angin-maksimum	0.135455	Digunakan
Kelembapan	0.090396	Tidak Digunakan
Temperature-rata-rata	0.084353	Tidak Digunakan
Temperature-maksimum	0.073162	Tidak Digunakan
Lama-penyiraman-matahari	0.056167	Tidak Digunakan
Temperature-minimum	0.039113	Tidak Digunakan
Curah-hujan	0.024114	Tidak Digunakan

Kemudian pada gambar 5 menunjukkan grafik setiap fitur berdasarkan nilai skor kepentingan (*importance score*) dari Fitur yang memiliki Skor terbanyak sampai yang terendah.



Gambar 5 Grafik fitur

d. Permodelan KNN + *Relief* + PSO

Dalam penelitian ini, akan dilakukan kombinasi dari tiga algoritma yang masing-masing telah menyelesaikan tahapannya, algoritma tersebut adalah KNN, *Relief*, dan PSO. Selanjutnya, penerapan optimasi PSO pada model klasifikasi dilakukan untuk meningkatkan peforma model KNN yang telah diintegrasikan menggunakan seleksi fitur *Relief*. Optimasi ini meliputi penyesuaian berbagai parameter, seperti

jumlah tetangga terdekat (*n_neighbors*), ukuran *leaf* (*leaf_size*), dan parameter jarak(p) dalam algortima KNN.

2.6 Evaluasi

Dalam penelitian ini, kinerja model KNN akan dievaluasi dengan membandingkan beberapa permodelan KNN, baik yang menggunakan metode seleksi fitur maupun yang tidak. Evaluasi model dilakukan untuk mendapatkan model terbaik melalui pengujian terhadap data uji berdasarkan *Confusion Matrix* sedangkan untuk Evaluasi menggunakan *accuracy* sebagai *metric*, yang dapat dihitung dengan persamaan berikut:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} X100\% \quad (5)$$

Keterangan:

TP (True Positive) : jumlah data yang berlabel benar diklasifikasikan sebagai benar oleh model.

TN (True Negative) : jumlah data yang berlabel salah diklasifikasikan sebagai salah oleh model.

FP (False Positive) : jumlah data yang berlabel benar seharusnya salah.

FN (False Negative) : jumlah data yang berlabel salah diklasifikasikan padahal seharusnya benar.

3 Hasil dan Pembahasan

3.1 Pengumpulan Data

Penelitian ini menggunakan total data banjir sebanyak 1095 *record*. yang didapatkan dari BPBD dan BMKG Kota Samarinda tahun 2021-2023 dalam dataset tersebut, terdapat 49 record yang menunjukkan adanya kejadian banjir, sementara 841 record menunjukkan tidak adanya kejadian banjir dengan label kelas 0 menunjukkan tidak terjadi banjir, sedangkan label kelas 1 menunjukkan terjadi banjir.

Data yang diperoleh dari BMKG yang memiliki 11 fitur sedangkan data yang diperoleh dari BPBD memiliki 9 fitur.

Data yang didapatkan dari BMKG meliputi tanggal, temperatur maksimum (Tx), temperatur minimum (Tn), temperatur rata-rata (Tavg), kelembaban rata-rata (RH-avg), curah hujan (RR), lamanya penyinaran matahari (ss), kecepatan angin maksimum (ff_x), arah angin maksimum (ddd_x), kecepatan angin rata-rata (ff_avg), dan arah angin terbanyak (ddd_car). Sedangkan data yang didapatkan dari BPBD meliputi tanggal, jam kejadian, jenis bencana, lokasi wilayah, luas area m2, objek terkena bencana, korban, kerugian, dan keterangan. hasil penggabungan data dari pendekatan *excel* tersebut dapat dilihat dalam tabel berikut : Seperti yang ditunjukkan pada Tabel 6 dan TABEL 7.

Tabel 6. 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
04-01-2021	24,6	31,8	28,1	84	2,7	5,4	3	300	2	NW
05-01-2021	24,6	31,4	27,4	83	10,5	1,8	4	300	2	W
…	…	…	…	…	…	…	…	…	…	…
27-12-2023	24,2	32	27,6	83	0,3	2,8	4	60	2	NE
28-12-2023	24	32	28	82	1	6,4	5	70	2	C
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 7. Data BPBD

No	Tanggal	Jam kejadian	Jenis Bencana	Lokasi/ Wilayah Kelurahan/ Kecamatan	Luas Area	Jumlah Objek Terkena Bencana	Korban					Jumlah	Kerugian (RP)		Keterangan
							KL	KS	KH	KM	KK	JIWA			
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

…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
14	Selasa, 31 Januari 2023	Pukul 19.45 wita	Pohon Tumbang	Jl. P. Antasari Pondok Wira 1 Kel. Teluk Lerong Ulu	-	Dampak: Menutup Bahu Jalan dan Menggangu Aktifitas Warga Sekitar	-	-	-	-	-	-	-	-	Penyebab terjadinya Hujan dengan intensitas sedang dan angin kencang

3.2 Permodelan KNN

Pada penelitian ini, tahap pemodelan menggunakan algoritma KNN menggunakan nilai k=3, k=5, k=7, k=11, k=13, dan k=15, sedangkan untuk pembagian data dilakukan menggunakan *10-fold cross-validation*. Pada Evaluasi akan menggunakan *Confusion Matrix*, Selanjutnya hasil evaluasi rata-rata *Confusion Matrix* untuk seluruh fold dari masing-masing nilai *k* dapat dilihat pada Tabel 8.

Tabel 8 Hasil Akurasi Permodelan KNN

Nilai K	TN	FP	FN	TP	*Mean Accuracy*
K=3	66	1.9	18	82	88,23%
K=5	62	2.6	22	82	85,38%
K=7	60	3.2	24	81	84,07%
K=9	58	3.6	26	80	82,52%
K=11	57	4.4	27	80	81,27%
K=13	56	4.9	28	79	80,62%
K=15	54	5.1	30	79	79,31%

3.3 Permodelan KNN + PSO

Hasil pemodelan menggunakan kombinasi KNN dan PSO menunjukkan peningkatan akurasi lebih tinggi dengan mengoptimalkan parameter dibandingkan dengan KNN tanpa optimasi. Penggunaan PSO membantu dalam menemukan parameter optimal yang dapat memaksimalkan kemampuan algortima KNN. Sehingga dapat memberikan Model yang lebih optimal dalam memprediksi data uji. Pada tabel 9 diperlihatkan hasil evaluasi algotima KNN setelah diterapkan optimasi PSO.

Tabel 9. Hasil Akurasi Permodelan KNN + PSO

Nilai K	TN	FP	FN	TP	*Mean Accuracy*
K=3	84	15	0.6	69	90,85%
K=5	83	17	0.8	67	89,42%
K=7	84	20	0.5	64	87,70%
K=9	83	22	0.9	62	86,50%
K=11	83	22	0.8	62	86,39
K=13	83	23	1.3	61	85,55%
K=15	83	24	60	1.4	84,84%

3.4 Permodelan KNN + Relief

Hasil pemodelan KNN dengan pemilihan fitur menggunakan *Relief,* Setelah menerapkan fitur-fitur yang relevan dari hasil seleksi fitur. hal ini menunjukkan bahwa *Relief* dapat memberikan pengaruh terhadap performa algoritma KNN, hal ini terbukti dengan meningkatnya akurasi sebesar 1-2% dibandingkan dengan knn tanpa seleksi fitur.

Tabel 10 Hasil Akurasi Permodelan KNN + Relief

Nilai K	TN	FP	FN	TP	*Mean Accuracy*
K=3	74	5.6	9.8	78	89.59%
K=5	72	6.1	79	12	87.69%
K=7	70	6.4	78	14	85.67%
K=9	69	6.9	77	16	84.90%
K=11	68	8.1	76	16	82.64%
K=13	66	9.8	74	18	81.09%
K=15	65	11	73	19	80.20%

3.5 Permodelan KNN + Relief + PSO

Hasil pemodelan yang menggabungkan algoritma KNN, seleksi fitur *Relief,* dan optimasi parameter melalui PSO mampu memberikan peningkatan peforma yang signifikan. Seleksi fitur dengan *Relief* mampu mengidentifikasi variabel yang paling relevan, sementara optimasi PSO berhasil menentukan parameter yang optimal untuk algoritma KNN. Hasil ini memberikan model yang tepat dalam melakukan prediksi data uji, serta meningkatkan peforma algortima secara keseluruhan.

Tabel 11. Hasil Akurasi Permodelan KNN + Relief + PSO

Nilai K	TN	FP	FN	TP	*Mean Accuracy*
K=3	66	1.9	18	82	90.84%
K=5	62	2.6	22	82	89.95%
K=7	60	3.2	24	81	87.75%
K=9	58	3.6	26	80	86.68%
K=11	57	4.4	27	80	85.55%
K=13	56	4.9	28	79	83.23%
K=15	54	5.1	30	79	82.11%

3.6 Pembahasan

Seleksi fitur *Relief* diterapkan pada data banjir Kota Samarinda untuk meningkatkan peforma model Klasifikasi *K-Nearst Neighbor* (KNN), dengan menggunakan fitur-fitur terpilih yang ditampilkan pada tabel 5. Penelitian sebelumnya juga mendukung hasil ini, seperti penelitian yang dilakukan oleh (Intan & Sari, 2023) yang menggunakan metode seleksi fitur *gain ratio* dengan mengidentifikasi beberapa fitur, seperti 'Kelembaban', 'Temperatur-minimum', dan 'Temperatur-maksimum' sebagai fitur yang paling berpengaruh, hal ini terbukti dari peningkatan akurasi algoritmai KNN sebesar 5.95%. Sementara itu, penelitian oleh (Evitasari et al., 2023) yang menerapkan metode seleksi fitur menggunakan *algoritma Genetik* (GA) menunjukkan bahwa fitur seperti : 'Kelembapan', 'Lama-penyinaran-matahari', dan 'Kecepatan-angin' memiliki pengaruh yang paling signifikan terhadap prediksi banjir, dimana hasilnya memberikan peningkatan akurasi algoritma klasifikasi SVM sebesar 13.45%.

Setelah dilakukan analisa, menunjukkan bahwa PSO secara signifikan meningkatkan akurasi algoritma KNN melalui optimasi parameter yang lebih tepat, sementara seleksi fitur dengan *Relief* juga memberikan peningkatan akurasi. Kombinasi *Relief* dan PSO memperlihatkan kombinasi dalam pengoptimalan parameter dan pemilihan fitur yang relevan, dengan peningkatan akurasi terbesar kedua setelah KNN dan PSO. Metode kombinasi ini menegaskan pentingnya penyesuaian parameter dan seleksi fitur untuk memperbaiki performa model secara signifikan, terutama pada nilai *k* yang lebih tinggi.

Tabel 12. Perbandingan Hasil Akurasi Dari Setiap Model KNN

Nilai K	*KNN*	KNN+PSO	Status	*KNN+Relief*	status	KNN+Relief+PSO	Status
K=3	88,23%	90,85%	+2.62%	89.59%	+1,36%	90.84%	+2,61%
K=5	85,38%	89,42%	+4.04%	87.69%	+2,31%	89.95%	+4,57%
K=7	84,07%	87,70%	+3,63%	85.67%	+1,60%	87.75%	+3,68%
K=9	82,52%	86,50%	+3.98%	84.90%	+2,38%	86.68%	+4,16%
K=11	81,27%	86,39	+5.12%	82.64%	+1,37%	85.55%	+4,28%
K=13	80,62%	85,55%	+4.93%	81.09%	+0,47%	83.23%	+2,61%
K=15	79,31%	84,84%	+5.53%	80.20%	+0,89%	82.11%	v2,80%

Peningkatan akurasi oleh KNN+PSO lebih tinggi dibandingkan KNN + *Relief* + PSO dapat dijelaskan oleh beberapa faktor utama. Pertama, optimasi parameter dengan PSO memungkinkan penyesuaian yang sangat tepat terhadap parameter model seperti jumlah *tetangga terdekat* (k), ukuran leaf, dan parameter jarak, dengan menggunakan pendekatan berbasis populasi untuk eksplorasi ruang parameter secara efektif. Kedua, PSO membantu mengatasi *overfitting* dengan mencari keseimbangan optimal antara bias dan *varians*, sehingga model dapat menangkap pola yang lebih *relevan* dalam data tanpa terlalu terpengaruh oleh *noise.* Ketiga, Meskipun seleksi fitur *Relief* meningkatkan akurasi dengan memilih fitur yang paling relevan dan mengurangi *noise*, sebagian peningkatannya tidak sebesar yang dicapai oleh PSO. Hal ini dikarenakan *Relief* hanya memilih fitur tanpa menyesuaikan parameter model, sementara PSO memberikan manfaat lebih efektif melalui optimasi parameter langsung. kombinasi antara seleksi fitur *Relief* dan optimasi PSO tetap memberikan manfaat, namun tidak selalu lebih baik daripada optimasi parameter langsung dengan PSO.

Dengan demikian, kemampuan PSO secara efektif menyesuaikan parameter model dalam mengatasi *overfitting* dan menjelaskan mengapa model KNN yang dioptimalkan dengan PSO secara konsisten mampu memberikan peningkatan akurasi tertinggi dibandingkan model KNN dasar, model KNN dengan seleksi fitur *Relief*, dan model KNN dengan kombinasi seleksi fitur *Relief* dan PSO.

4 Kesimpulan

Hasil dari penerapan seleksi fitur *Relief* yang diterapkan pada data banjir Kota Samarinda mengidentifikasi empat fitur dengan pengaruh signifikan berdasarkan peringkatnya, yaitu Arah -angin-terbanyak, Kecepatan-angin, Kecepatan-angin-rata-rata, dan Arah-angin-maksimum. Setelah melalui tahap seleksi fitur dan penerapan PSO, disimpulkan bahwa penerapan *Relief* efektif dalam meningkatkan akurasi dari algoritma *k-Nearest Neighbor* pada data banjir Kota Samarinda dengan hasil akurasi KNN PSO memberikan peningkatan sebesar 2-5%, dengan

seleksi fitur *Relief* mengalami peningkatan sebesar 1-2% dan KNN dengan kombinasi *Relief* dan PSO terjadi peningkatan sebesar 2-5%.

5 Saran

Untuk penelitian selanjutnya diharapkan dapat mengeksplorasi metode seleksi fitur lainnya dengan menggunakan metode lain pada algoritma *k-Nearest Neighbor* seperti ANOVA *(Analysis of Variance), Adaboost, Chi-square, Recursive Feature Elimination* (RFE) dan metode lainnya yang relevan dan informatif, sehingga dapat lebih meningkatkan akurasi dan efisiensi model klasifikasi KNN dan Melakukan studi komparatif dengan algoritma klasifikasi lain seperti *Random Forest, Support Vector Machine (SVM), Naïve Bayes,* dan *Decision Tree* untuk melihat bagaimana kombinasi seleksi fitur dan optimasi parameter dapat diterapkan pada algoritma lain dan membandingkan hasilnya.

Referensi

Abdulrazaq, M. B., Mahmood, M. R., Zeebaree, S. R. M., Abdulwahab, M. H., Zebari, R. R., & Sallow, A. B. (2021). An Analytical Appraisal for Supervised Classifiers' Performance on Facial Expression Recognition Based on Relief-F Feature Selection. Journal of Physics: Conference Series, 1804(1). https://doi.org/10.1088/1742-6596/1804/1/012055

Ariyoga, D. (2022). Perbandingan Metode Seleksi Fitur Filter, Wrapper, Dan Embedded Pada Klasifikasi Data Nirs Mangga Menggunakan Random Forest Dan Support Vector Machine .https://dspace.uii.ac.id/handle/123456789/38955

Arora, A., Arabameri, A., Pandey, M., Siddiqui, M. A., Shukla, U. K., Bui, D. T., Mishra, V. N., & Bhardwaj, A. (2021). Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India. Science of the Total Environment, 750(August). https://doi.org/10.1016/j.scitotenv.2020.141565

Cumel, David Zamri, Rahmaddeni, S. (2022). Perbandingan Metode Data Mining untuk Prediksi Banjir Dengan Algoritma Naïve Bayes dan KNN. SENTIMAS: Seminar Nasional Penelitian Dan, 40–48. https://journal.irpi.or.id/index.php/sentimas/article/view/353%0Ahttps://journal.irpi.or.id/index.php/sentimas/article/download/353/132

Daniel, I., Hartono, H., & Situmorang, Z. (2023). Analysis of Machine Learning Algorithms in Predicting the Flood Status of Jakarta City. International Conference on Information Science and Technology Innovation (ICoSTEC), 2(1), 82–87. https://doi.org/10.35842/icostec.v2i1.42

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

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

Ernawati, R., Dirdjo, M. M., & Wahyuni, M. (2021). Peningkatan Pengetahuan Siswa Terhadap Mitigasi Bencana di SD Muhammadiyah 4 Samarinda. Journal of Community Engagement in 4(2), 393–399. https://jceh.org/index.php/JCEH/article/view/258

Evitasari, Y. D., Pranoto, W. J., & Verdikha, N. A. (2023). Evaluasi Support Vector Machine Dengan Optimasi Metode Genetic Algorithm Pada Klasifikasi Banjir Kota Samarinda. Jurnal Sains Komputer Dan Teknologi Informasi, 6(1), 49–53. https://doi.org/10.33084/jsakti.v6i1.5462

Faldi, F., NurHalisha, T., Pranoto, W. J., & ... (2023). The application of particle swarm optimization (PSO) to improve the accuracy of the naive bayes algorithm in predicting floods in the city of Samarinda. Journal of Intelligent …, 6(3), 138–146. http://idss.iocspublisher.org/index.php/jidss/article/view/148%0Ahttps://idss.iocspublisher.org/index.php/jidss/article/download/148/99

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, 357–361. https://doi.org/10.1109/ICREST51555.2021.9331199

Hossain, M. S., & Zeyad, M. (2023). Prediction of Flood in Bangladesh Using Different Classifier Model. AIUB Journal of Science and Engineering, 22(1), 45–52. https://doi.org/10.53799/ajse.v22i1.365

Intan, S., & Sari, P. (2023). Analisis Pengaruh Gain Ratio Untuk Algoritma K-Nearest Neighbor Pada Klasifikasi Data Banjir Di Kota Samarinda Analysis Of The Effect Of Gain Ratio For

Algorithms K-Nearest Neighbor On Classsification Flood Data In Samarinda City. Jurnal Sains Komputer Dan, 6(1), 54–59. https://journal.umpr.ac.id/index.php/jsakti/article/view/5472%0Ahttps://journal.umpr.ac.id/index.php/jsakti/article/download/5472/3664

Kemal Musthafa Rajabi, Witanti, W., & Rezki Yuniarti. (2023). Penerapan Algoritma K-Nearest Neighbor (KNN) Dengan Fitur Relief-F Dalam Penentuan Status Stunting. INNOVATIVE: Journal Of Social Science Research, 3, 3555–3568.

Nabila, S. P., Ulinnuha, N., Yusuf, A., Informasi, S., Wonosari, J., & Timur, J. (2021). Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors. 6(1), 39–47.

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

Nursyahfitri, R., Rozikin, C., & Adam, R. I. (2022). Penerapan Metode SMOTE dalam Klasifikasi Daerah Rawan Banjir di Karawang Menggunakan Algoritma Naive Bayes. Jurnal Sistem Dan Teknologi Informasi (JustIN), 10(4), 339. https://doi.org/10.26418/justin.v10i4.46935

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.

Purwanto, P. (2020). Analisis Sistem Pengendalian Banjir Sungai Pampang Daerah Aliran Hulu Sungai Karangmumus. Jurnal Kacapuri : Jurnal Keilmuan Teknik Sipil, 3(2), 44. https://doi.org/10.31602/jk.v3i2.4066

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

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

Tarigan, P. M. S., Hardinata, J. T., Qurniawan, H., Safii, M., & Winanjaya, R. (2022). Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang. Jurnal Janitra Informatika Dan Sistem Informasi, 2(1), 9–19. https://doi.org/10.25008/janitra.v2i1.142

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

Yahdin, S., Desiani, A., Gofar, N., & Agustin, K. (2021). Application of the Relief-f Algorithm for Feature Selection in the Prediction of the Relevance Education Background with the Graduate Employment of the Universitas Sriwijaya. Computer Engineering and Applications Journal, 10(2), 71–80. https://doi.org/10.18495/comengapp.v10i2.369

Yoga Siswa, T. A. (2023). Data Mining: Mengupas Tuntas Analisis Data Dengan Metode Klasifikasi Hingga Deployment Aplikasi Menggunakan Python (T. A. Yoga Siswa (ed.)). UMKT PRESS.

Yusra, R. N., Sitompul, O. S., & Sawaluddin. (2021). Kombinasi K-Nearest Neighbor (KNN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data. InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan, 1, 0–5.