

# **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 Classification Flood Data In Samarinda City**

Septia Intan Permata Sari  
Wawan Joko Pranoto  
Naufal Azmi Verdikha  
Program Studi Teknik Informatika,  
Fakultas Sains dan Teknologi,  
Universitas Muhammadiyah Kalimantan Timur

### **ABSTRAK**

Berdasarkan data yang diperoleh dari BMKG dan BPBD Kota Samarinda, diketahui bahwa telah terjadi bencana banjir pada periode tahun 2019 - 2020 di Kota Samarinda. Penelitian ini bertujuan untuk melakukan klasifikasi data banjir di Kota Samarinda menggunakan algoritma K-Nearest Neighbor dan pembagian data menerapkan teknik 5-Fold Cross-Validation serta perhitungan rumus jarak Euclidean Distance. Kemudian, dilakukan seleksi fitur pada algoritma KNN menggunakan metode Gain Ratio guna mengetahui pengaruhnya terhadap akurasi dari KNN. Hasil penelitian menunjukkan bahwa peningkatan akurasi tertinggi setelah menerapkan Gain Ratio didapatkan oleh K=7 dengan persentase kenaikan akurasi sebesar 5,95%, diikuti oleh K=5 dengan persentase kenaikan akurasi 5,81%, K=3 dengan persentase kenaikan akurasi 5,68%, K=9 sebesar 3,61%, K=11 sebesar 2,44%, dan K=13 sebesar 1,23%. Hanya ada satu akurasi yang tidak mengalami peningkatan atau penurunan akurasi, yaitu K=15.

**Kata kunci:** *Banjir, K-Nearest Neighbor, 5-Fold Cross-Validation, Euclidean Distance, Gain Ratio*

### **ABSTRACT**

*Based on data obtained from the BMKG and BPBD Samarinda City, it is known that there has been a flood disaster in the 2019 - 2020 period in Samarinda City. This study aims to classify flood data in Samarinda City using the K-Nearest Neighbor algorithm and dividing data using the 5-Fold Cross-Validation technique and calculating the Euclidean Distance formula. Then, feature selection is carried out in the KNN algorithm using the Gain Ratio method to determine its effect on the accuracy of the KNN. The results showed that the highest increase in accuracy after applying the Gain Ratio was obtained by K=7 with a percentage increase in accuracy of 5.95%, followed by K=5 with a percentage increase in accuracy of 5.81%, K=3 with a percentage increase in accuracy of 5.68 %, K=9 of 3.61%, K=11 of 2.44%, and K=13 of 1.23%. There is only one accuracy that does not experience an increase or decrease in accuracy, namely K=15.*

**Keywords:** *Flood, K-Nearest Neighbor, 5-Fold Cross-Validation, Euclidean Distance, Gain Ratio*

### **Pendahuluan**

Banjir merupakan bencana alam yang sering terjadi di negara Indonesia. Berdasarkan data Badan Nasional Penanggulangan Bencana bahwa sejak awal tahun hingga bulan Mei tahun 2020, telah terjadi 1296 bencana alam dengan kejadian bencana alam banjir mencapai 495 kasus. Samarinda sebagai ibukota Provinsi Kalimantan Timur cukup mengawatirkan, di mana bencana banjir ini mengalami peningkatan cukup tinggi. Merujuk pada data dari situs Badan Pusat Statistik Kota Samarinda, terjadi peningkatan jumlah desa atau kelurahan yang terkena dampak banjir pada tahun 2014 sebanyak 24 kali, pada tahun 2018 sebanyak 18 kali dan pada tahun 2020 sebanyak 33 kali (BPS, 2023)

Klasifikasi banjir berdasarkan penyebabnya dapat membantu memperbaiki perkiraan frekuensi banjir, mendukung deteksi serta penafsiran untuk perubahan kejadian dan besaran banjir (Tarasova et al., 2019). Oleh karena itu, perlu diadakan evaluasi perbaikan akurasi dengan metode klasifikasi data mining.

data mining merupakan salah satu proses yang digunakan untuk membuat koleksi data besar menjadi sebuah pengetahuan yang dibutuhkan (Taghifuril Azhima Yoga Siswa, S.Kom., 2023). Data mining juga dapat digunakan untuk mendapatkan informasi terkini yang berharga dan bermanfaat dalam pengumpulan data yang melibatkan antara manusia dan perangkat komputer, baik melalui cara otomatis ataupun manual (Tarigan et al., 2022)

Penelitian terkait perbandingan algoritma Naive Bayes dan KNN, dengan hasil penilaian yang membuktikan bahwa algoritma KNN lebih baik dibanding Naive Bayes dengan tingkat akurasi lebih tinggi, yaitu 88,94% disertai dengan nilai margin error yang lebih kecil, yaitu hanya 11,06% (Zamri, 2022)

Pada penelitian lain yang dilakukan oleh (Tempola et al., 2018) perbandingan klasifikasi antara knn dan naive bayes pada penentuan status gunung berapi dengan k-fold cross validation diperoleh hasil

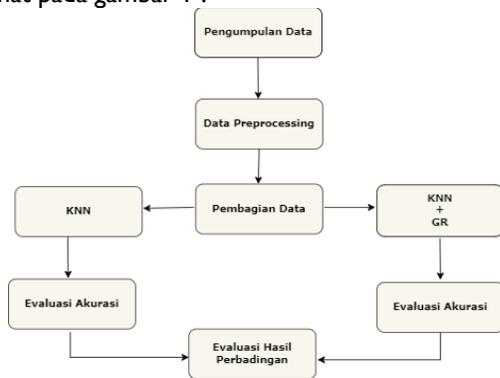
menggunakan k-nn akurasi rata-rata sistem 63,68% dan standar deviasi sebesar 7,47

Gain Ratio sendiri pernah digunakan dalam penelitian terdahulu, diantaranya pada penelitian komparasi tingkat akurasi information gain dan gain ratio pada metode K- Nearest Neighbor terbukti dapat meningkatkan nilai akurasi 70,06% sehingga dapat disimpulkan penggunaan seleksi fitur gain ratio dianggap lebih optimal (Fajriyan et al., 2022). Dalam lingkup yang akan datang, Gain Ratio dapat menjadi rekomendasi bagi peneliti lain dalam mereduksi data atribut untuk meningkatkan kinerja suatu algoritma (Hasdyna et al., 2020). Gain ratio adalah seleksi fitur yang banyak digunakan oleh peneliti karena handal dan mampu berjalan pada dimensi data yang tinggi. Tahap seleksi fitur sebenarnya melibatkan perhitungan kepentingan dari semua atribut data yang ada. atribut dengan tingkat kepentingan tinggi yang akan digunakan, sementara atribut dengan tingkat kepentingan rendah tidak akan digunakan dalam tahap berikutnya, yaitu klasifikasi. (Kurniawan et al., 2018)

Penelitian ini menggunakan metode eksperimen yang merupakan jenis penelitian kuantitatif untuk mengukur fitur yang mempengaruhi akurasi banjir. Penelitian ini menguji algoritma KNN dan melakukan seleksi fitur dengan menggunakan metode Gain Ratio, untuk mengetahui apakah terdapat kenaikan, penurunan atau perubahan nilai akurasi.

**Metode**

Alur atau tahapan dalam penelitian ini adalah sebagaimana terlihat pada gambar 1 :



**Gambar 1 Alur Penelitian Pengumpulan Data**

Tahapan pertama penelitian ini pengumpulan data dengan melakukan observasi langsung ke BPBD (Badan Penanggulangan Bencana Daerah) Kota Samarinda dan BMKG Kota Samarinda untuk data primer .Dataset ini mencakup rentang waktu dari tahun 2019 hingga 2022. hasil pengumpulan data dari kedua sumber tersebut, terdapat 1462 dataset yang dapat diakses dengan 20 atribut. Untuk memudahkan pemahaman, data banjir Kota Samarinda dibentuk tabel yang dapat mempermudah

dalam mengetahui berapa banyak atribut yang ada sebagai berikut:

**Table 1 Atribut Data BPBD Dan BMKG Kota Samarinda 2019 - 2022**

| No | Atribut               | Keterangan                             |
|----|-----------------------|--|
| 1  | Tanggal               | Waktu Kejadian                         |
| 2  | Tn                    | Temperatur minimum suhu (°C)           |
| 3  | Tx                    | Temperatur maksimum suhu (°C)          |
| 4  | Tavg                  | Temperatur rata-rata (°C)              |
| 5  | RH_avg                | Kelembapan rata-rata (%)               |
| 6  | RR                    | Curah hujan (mm)                       |
| 7  | Ss                    | Lamanya penyinaran matahari (jam)      |
| 8  | ff_x                  | Kecepatan angin maksimum (m/s)         |
| 9  | ddd_x                 | Arah angin saat kecepatan maksimum (°) |
| 10 | ff_avg                | Kecepatan angin rata-rata (m/s)        |
| 11 | ddd_car               | Arah angin terbanyak (°)               |
| 12 | Terjadi Banjir        | Ya/ Tidak                              |
| 13 | Jam Kejadian          | Jam terjadinya bencana                 |
| 14 | Jenis Bencana         | Peristiwa bencana alam yang terjadi    |
| 15 | Lokasi Wilayah        | Tempat terjadinya banjir               |
| 16 | Luas Area M2          | Luas area yang terdampak               |
| 17 | Objek Terkena Bencana | Fasilitas yang tedampak bencana        |
| 18 | Korban                | Jumlah korban yang terdampak bencana   |
| 19 | Kerugian              | Nominal kerugian yang dialami          |
| 20 | Keterangan            | Detail kejadian bencana                |

**Data Pre-Processing**

Tahap ini melibatkan lima tahapan, yakni data integration, data selection, data transformation, data

cleaning, serta data Balancing agar dapat dilanjutkan kedalam tahap pemodelan (modeling). dengan proses persiapannya sebagai berikut :

1. Prosesnya menggabungkan dua data dengan data BPBD yang memiliki fitur Tanggal dan data dari sumber yaitu BMKG memiliki fitur yaitu suhu minimal, suhu maksimal, suhu rata-rata, kelembaban rata-rata, curah hujan, kecepatan angin, kecepatan angin pada arah x, arah angin pada arah x, kecepatan angin rata-rata, arah angin kendaraan, terjadi banjir.
2. Pada tahapan data selection, dipilih fitur data yang paling sesuai terhadap tujuan penelitian. Pemilihan fitur dilakukan secara manua
3. Tahap data transformation akan mengubah nilai data ke dalam format yang diinginkan dan sesuai pada proses data mining
4. Data cleaning dilakukan menggunakan python untuk menghilangkan data yang buruk, rusak, cacat atau bahkan tidak lengkap dengan menggantikan nilai yang missing value ke nilai rata-rata
5. Pada tahap terakhir, terjadi ketidakseimbangan data ketika jumlah dalam kelas-kelas berbeda dalam dataset yang tidak seimbang. Maka dilakukan data balancing menggunakan Python dengan modul ``imblearn.over_sampling`` dan mengimpor fungsi `'SMOTE (sampling_strategy)'` untuk melakukan oversampling. Salah satu metode oversampling yang digunakan adalah SMOTE (Synthetic Minority Over-sampling Technique).

### Pembagian Data

Data dibagi menjadi dua bagian: data training dan data testing. dengan teknik K-Fold Cross Validation parameter  $Cv = 5$  yang artinya data dibagi menjadi 5 fold atau grup. Setiap lipatan bergantian antara data pelatihan dan data uji.

### Modelling

Berikut ini adalah penjelasan mengenai gambaran yang dihasilkan dari model knn :

Pada tahap awal, data melalui proses pembagian menjadi dua bagian, yaitu data training dan data testing, dengan menggunakan metode 5-fold Cross-validation. data training digunakan untuk melatih model algoritma, sedangkan data testing digunakan untuk mengevaluasi model algoritma yang telah dilatih sebelumnya. Model algoritma yang digunakan untuk melakukan klasifikasi Banjir di kota Samarinda adalah kNN. Untuk menghitung jarak antara dua data, dapat digunakan rumus persamaan Euclidean.

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

Sumber : (Jatmiko Indriyanto, 2021)

Tahap seleksi fitur dilakukan menggunakan metode Gain Ratio dengan cara kerjanya yang menentukan atribut terbaik atau tertinggi dalam

mempengaruhi hasil klasifikasi. Nilai gain ratio biasanya berkisar antara 0 hingga 1, di mana semakin tinggi nilainya, semakin besar kontribusi fitur dalam memahami pola data yang digunakan.

### Evaluasi

Pada tahap evaluasi ini mengukur akurasi hasil dari algoritma yang diimplementasikan dengan kualitas data training serta akan diuji dengan menggunakan teknik Confusion Matrix.

### Hasil Dan Pembahasan

Sebelum melakukan pemodelan, dilakukan proses data preparation dengan 4 tahapan sebagai berikut :

#### 1. Data Integration

Penelitian ini menggunakan total data banjir sebanyak 1461 record. dalam dataset tersebut, terdapat 75 record yang menunjukkan adanya kejadian banjir, sementara 1368 record menunjukkan tidak adanya kejadian banjir dengan label kelas 0 menunjukkan tidak terjadi banjir, sedangkan label kelas 1 menunjukkan terjadi banjir. pada tahap ini, dilakukan penggabungan antara data BMKG yang memiliki 10 atribut dan data BPBD Kota Samarinda yang memiliki 11 atribut. Proses penggabungan ini dirancang sedemikian rupa untuk memastikan bahwa data dari BMKG dan BPBD saling melengkapi dan menghasilkan dataset yang optimal untuk klasifikasi banjir.

#### 2. Data Selection

Tahap ini dilakukan untuk memilih atribut-atribut yang akan digunakan dalam proses pemodelan menggunakan Python dengan library pandas. Dalam tahap ini, digunakan fungsi `drop()` untuk menghapus atribut tanggal dari dataset. Tujuan dari langkah ini adalah agar data yang digunakan dalam pemodelan hanya terdiri dari atribut-atribut yang dianggap berpengaruh pada hasil klasifikasi. Tujuan dari langkah ini adalah agar data yang digunakan dalam pemodelan hanya terdiri dari atribut-atribut yang dianggap berpengaruh pada hasil klasifikasi.

#### 3. Data Transformation

Pada tahap ini, atribut Arah Angin Terbanyak diubah dari simbol arah mata angin menjadi derajat mata angin. Hal ini dilakukan karena metode seleksi fitur Gain Ratio yang digunakan dalam penelitian tidak dapat memproses atribut dalam bentuk string. Dengan mengubah atribut menjadi bentuk numerik, seperti derajat mata angin, dapat memudahkan proses pengolahan data dan analisis menggunakan pendekatan Python. Selain itu, penggunaan derajat mata angin memberikan representasi numerik yang lebih langsung dan konsisten, memungkinkan komparasi dan perhitungan matematis yang mudah. Dengan demikian, transformasi ini memungkinkan pengolahan data

yang optimal untuk mencari pola dan hubungan antara arah angin dan data target.

**Table 2 Hasil Transformasi Data Sebagian**

| No.  | Arah angin terbanyak (Sebelum Ditransformasi) | Arah angin terbanyak (Setelah Ditransformasi) |
|------|---|---|
| 1    | NW  | 315   |
| 2    | C   | 60  |
| 3    | NE  | 45  |
| 4    | SE  | 135   |
| 5    | N   | 360   |
| 6    | S   | 180   |
| 7    | SW  | 225   |
| :    | :   | :   |
| 1462 | W   | 270   |

Hasil keseluruhan dari proses *data transformation* dapat dilihat pada nilai dari atribut arah-angin-terbanyak telah ditransformasikan menjadi angka sesuai dengan mapping yang telah ditentukan.

**4. Data Cleaning**

Tahap data cleaning dilakukan melalui pendekatan python menggunakan metode `fillna()`. Tujuan dari tahapan ini adalah untuk menggantikan nilai NaN dengan nilai rata-rata pada data

**5. Data Balancing**

Dalam penelitian ini, langkah data balancing dilakukan untuk mengatasi ketidakseimbangan kelas dalam kumpulan data. Teknik yang digunakan pada langkah ini adalah oversampling. oversampling adalah meningkatkan jumlah sampel pada kelas minoritas agar sebanding dengan jumlah sampel pada kelas mayoritas,Sebanyak 75 record “TERJADI BANJIR” merupakan data mayoritas, dan “TIDAK ADA BANJIR” untuk minoritas dengan total ada 1386 record. Menggunakan Python dengan modul `imblearn.over\_sampling` dan mengimpor fungsi `SMOTE(sampling\_strategy)` untuk melakukan oversampling. Salah satu metode oversampling yang digunakan adalah SMOTE (Synthetic Minority Over-sampling Technique).

**Modelling, Seleksi Fitur & Evaluasi**

Tahapan seleksi fitur ini melakukan perhitungan untuk seluruh atribut dataset dengan menggunakan seleksi fitur gain ratio. Proses ini memungkinkan mendapatkan nilai korelasi untuk semua atribut dalam proses klasifikasi selanjutnya. nilai korelasi tertinggi adalah 0.0978,0.0969,0.968 merupakan atribut dengan kepentingan tertinggi dalam klasifikasi. Sebaliknya atribut dengan nilai korelasi 0.0942 0.0941 0.0876 0.0857 0.0830 0.0828 0.0663 merupakan atribut yang kepentingannya sangat kecil dalam proses klasifikasi atau bahkan mungkin tidak memiliki kepentingan dalam proses klasifikasi. Tabel

3 merupakan hasil perhitungan seleksi fitur gain ratio untuk pembobotan atribut data Kelembapan, Temperatur-minimum, Temperatur-maksimum.

**Table 3 Penentuan Atribut yang Digunakan**

| No  | Atribut                     | Nilai Korelasi | Hasil           |
|-----|-----------------------------|----------------|-----------------|
| 1.  | Kelembapan                  | 0.0978         | Digunakan       |
| 2.  | Temperatur-minimum          | 0.0969         | Digunakan       |
| 3.  | Temperatur-maksimum         | 0.0968         | Digunakan       |
| 4.  | Temperatur                  | 0.0942         | Tidak Digunakan |
| 5.  | Lamanya-penyinaran-matahari | 0.0941         | Tidak Digunakan |
| 6.  | Arah-angin-maksimum         | 0.0876         | Tidak Digunakan |
| 7.  | Curah hujan                 | 0.0857         | Tidak Digunakan |
| 8.  | Kecepatan-angin-rata-rata   | 0.0830         | Tidak Digunakan |
| 9.  | Kecepatan-angin             | 0.0828         | Tidak Digunakan |
| 10. | Arah-angin-terbanyak        | 0.0663         | Tidak Digunakan |

Dari hasil tabel 3 di atas dapat diketahui bahwa atribut dengan nilai korelasi tertinggi yaitu 0.0978,0.0969,0.968Artinya atribut Kelembapan, Temperatur-minimum, Temperatur-maksimum adalah atribut dengan kepentingan tertinggi dibandingkan atribut yang lain. Sedangkan atribut lain merupakan atribut dengan tingkat kepentingan terendah karena memiliki nilai korelasi 0.0942 0.0941 0.0876 0.0857 0.0830 0.0828 0.0663 dalam perhitungan menggunakan algoritma seleksi fitur gain ratio.setelah melalui tahap seleksi fitur dilakukan lagi pemodelan ulang pada algoritma KNN untuk melihat pengaruh Gain Ratio pada akurasi dari KNN.Perbandingan hasil akurasi rata-rata yang didapatkan dari keseluruhan fold sebelum dan sesudah seleksi fitur dapat dilihat pada table 4

**Table 4 Perbandingan Hasil Akurasi Rata-rata**

| Nilai K | Mean Accuracy (Sebelum Seleksi Fitur) | Mean Accuracy (Setelah Seleksi Fitur) | Status  |
|---------|---------------------------------------|---------------------------------------|---------|
| K=3     | 0.88                                  | 0.93                                  | + 5.68% |
| K=5     | 0.86                                  | 0.91                                  | + 5.81% |
| K=7     | 0.84                                  | 0.89                                  | + 5.95% |
| K=9     | 0.83                                  | 0.86                                  | + 3.61% |
| K=11    | 0.82                                  | 0.84                                  | + 2.44% |
| K=13    | 0.81                                  | 0.82                                  | + 1.23% |

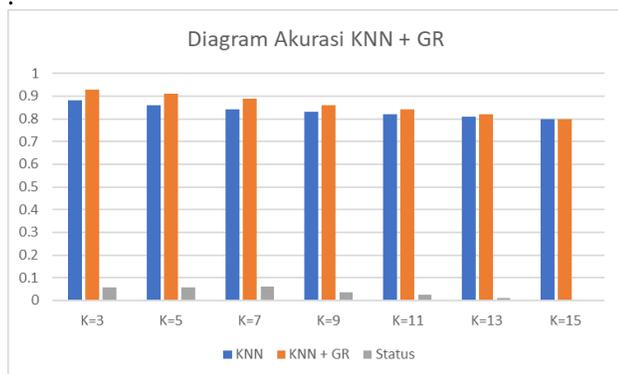
K=15                      0.80                      0.80                      0%

Bisa dilihat pada Tabel , bahwa terdapat perubahan dalam hal akurasi pada algoritma k-Nearest Neighbor setelah menerapkan seleksi fitur Gain Ratio. Nilai K= 9 sampai K=13 mengalami kenaikan, sedangkan nilai K 15 tidak mengalami perubahan atau tetap.

**Pembahasan**

Dalam penelitian ini, data banjir kota Samarinda yang dikumpulkan dari BPBD dan BMKG Kota Samarinda pada periode 2019-2022 digunakan. Data ini melalui tahap preprocessing mulai dari data integration, data selection, data transformation, data cleaning dan data balancing. Tujuan dari tahap ini adalah untuk memastikan kualitas data yang digunakan. Mengingat adanya ketidakseimbangan kelas pada data yang dikumpulkan, dilakukan penyeimbangan data dengan menggunakan teknik oversampling menggunakan modul `imblearn.over\_sampling` di Python dan fungsi 'SMOTE (sampling strategy)'. Selain itu, dataset juga dibagi menjadi data training dan data testing menggunakan teknik k-fold cross validation dengan nilai k= 5.

Dalam proses seleksi fitur, penelitian ini menggunakan metode Gain Ratio. Setelah melakukan pemodelan dengan 10 fitur yang ada, terpilih 3 fitur terbaik yaitu Kelembapan, Temperatur-minimum, dan Temperatur-maksimum.pada penerapan algoritma KNN, dilakukan dua skema. Skema pertama menggunakan data tanpa seleksi fitur Gain Ratio, sedangkan skema kedua menggunakan data setelah dilakukan seleksi fitur Gain Ratio. KNN dijalankan dengan menggunakan K=3 hingga K=15.hasil pengujian dengan menggunakan teknik confusion matrix menunjukkan peningkatan performa algoritma KNN setelah menerapkan seleksi fitur dengan Gain Ratio. Hal ini dapat dilihat pada diagram dibawah ini :



**Gambar 2 Diagram Akurasi KNN+GR**

**Kesimpulan**

Berdasarkan penelitian yang telah dilakukan, maka kesimpulan dari penelitian adalah sebagai berikut:

1. Hasil dari penerapan seleksi fitur menggunakan Gain Ratio pada data banjir kota Samarinda terdapat tiga atribut yang memiliki pengaruh paling signifikan

terhadap terjadinya banjir. Atribut-atribut tersebut adalah kelembapan, lamanya penyinaran matahari, dan kecepatan angin.

2. Setelah melalui tahap seleksi fitur, disimpulkan bahwa penggunaan Gain Ratio efektif dalam meningkatkan akurasi dari algoritma k-Nearest Neighbor pada data banjir di Kota Samarinda. Peningkatan akurasi tertinggi didapatkan oleh K=7 dengan persentase kenaikan akurasi sebesar 5,95%, diikuti oleh K=5 dengan persentase kenaikan akurasi 5,81%, K=3 dengan persentase kenaikan akurasi 5,68%, K=9 sebesar 3,61%, K=11 sebesar 2,44%, dan K=13 sebesar 1,23%. Hanya ada satu akurasi yang tidak mengalami peningkatan atau penurunan akurasi, yaitu K=15.

Pada penelitian ini, algoritma K-Nearest Neighbor (KNN) dan seleksi fitur Gain Ratio digunakan untuk mengklasifikasi dan menyeleksi fitur yang paling berpengaruh terhadap data banjir di Kota Samarinda. Dengan memadukan kedua metode tersebut, penelitian ini berhasil mencapai tingkat akurasi yang cukup baik dalam analisis data banjir. Penggunaan algoritma KNN memungkinkan klasifikasi banjir berdasarkan atribut-atribut yang ada dalam dataset, sedangkan metode seleksi fitur Gain Ratio membantu mengidentifikasi atribut-atribut yang memberikan kontribusi paling signifikan dalam klasifikasi banjir. Hasil penelitian ini memberikan wawasan yang berharga dalam pemahaman dan penanganan bencana banjir di Kota Samarinda

**Saran**

Saran untuk penelitian selanjutnya, diantaranya:

1. Pada penelitian selanjutnya, diharapkan untuk melakukan seleksi fitur dengan menggunakan metode lain pada algoritma k-Nearest Neighbor untuk data nilai Banjir di kota Samarinda. seperti ANOVA (Analysis of Variance), Adaboost, Chi-square, Recursive Feature Elimination, Principal Component Analysis (PCA), dan metode lainnya yang relevan.
2. Untuk penelitian selanjutnya, dapat memilih lima sampai tujuh fitur atau atribut paling berpengaruh berdasarkan perhitungan Gain Ratio untuk melihat perbandingan pengaruhnya pada hasil akurasi algoritma KNN.

**DAFTAR PUSTAKA**

**Lain-lain**

BPS. (2023). *Jumlah Desa /Kelurahan yang Mengalami Bencana Alam (Banjir) Menurut Kecamatan di Kota Samarinda 2018-2021*. <https://samarindakota.bps.go.id/indicator/153/207/1/jumlah-desa-kelurahan-yang-mengalami-bencana-alam-banjir-menurut-kecamatan-di-kota-samarinda.html>

**Buku**

Taghfirul Azhima Yoga Siswa, S.Kom., M. K. (2023). *DATA MINING: MENGUPAS TUNTAS ANALISIS DATA DENGAN METODE KLASIFIKASI HINGGA*

DEPLOYMENT APLIKASI MENGGUNAKAN PYTHON.  
UMKT PRESS.

**Jurnal/Prosiding/Disertasi/Tesis/Skripsi**

- 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>
- Tempola, F., Muhammad, M., & Khairan, A. (2018).

Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(5), 577. <https://doi.org/10.25126/jtiik.201855983>

- Zamri, D. (2022). ... Metode Data Mining untuk Prediksi Banjir Dengan Algoritma Naive Bayes dan KNN: Comparison of Data Mining Methods for Prediction of Floods with Naive Bayes .... *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>
- Kurniawan, M. F., Irawan, N., Tinggi, S., Informatika, M., Komputer, D., Pratama, W., & Com, M. F. (2018). Peningkatan Performa Algoritma Naive Bayes Dengan Gain Ratio Untuk Klasifikasi Kanker Payudara. *Seminar Nasional Edusaintek*, 53–61.