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

Yuliana Dilla Evitasari
Wawan Joko Pranoto
Naufal Adzmi Verdikha
Program Studi teknik Informatika,
Fakultas Sains dan Teknologi,
Universitas Muhammadiyah Kalimantan Timur
1911102441117@umkt.ac.id

ABSTRAK

Banjir merupakan bencana alam yang sering terjadi di Indonesia, terutama di kota Samarinda yang terletak di Kalimantan Timur. Penelitian ini bertujuan untuk meningkatkan akurasi dengan menerapkan metode seleksi fitur menggunakan Genetic Algorithm (GA). Melalui analisis data banjir kota Samarinda, ditemukan bahwa terdapat tiga atribut yang paling berpengaruh terhadap terjadinya banjir, yaitu kelembapan, lamanya penyinaran matahari, dan kecepatan angin. Selanjutnya, penelitian ini menggunakan algoritma Support Vector Machine (SVM) untuk mengklasifikasikan data banjir. Dengan menerapkan seleksi fitur menggunakan GA, hasil pengujian menunjukkan peningkatan akurasi algoritma SVM sebesar 13.45%. Sebelum penerapan seleksi fitur, akurasi SVM hanya mencapai 52,71%, namun setelah penerapan seleksi fitur menggunakan GA, akurasi meningkat menjadi 66,16%. Hasil ini membuktikan bahwa seleksi fitur dengan menggunakan GA efektif dalam meningkatkan akurasi prediksi banjir. Kesimpulan dari penelitian ini adalah seleksi fitur menggunakan GA dapat mengidentifikasi atribut-atribut yang paling berpengaruh terhadap terjadinya banjir di kota Samarinda. Penerapan seleksi fitur ini menghasilkan peningkatan signifikan dalam akurasi algoritma SVM untuk prediksi banjir.

Kata kunci: Klasifikasi, Support vector Machine, Genetica Algorithm, Data Mining, Banjir

ABSTRACT

Flooding is a natural disaster that often occurs in Indonesia, especially in the city of Samarinda, which is located in East Kalimantan. This research aims to improve accuracy by applying a feature selection method using Genetic Algorithm (GA). Through analyzing the flood data of Samarinda city, it was found that there are three attributes that have the most influence on the occurrence of floods, namely humidity, length of sunshine, and wind speed. Furthermore, this study used the Support Vector Machine (SVM) algorithm to classify the flood data. By applying feature selection using GA, the test results show an increase in the accuracy of the SVM algorithm by 13.45%. Before the application of feature selection, the accuracy of SVM only reached 52.71%, but after the application of feature selection using GA, the accuracy increased to 66.16%. These results prove that feature selection using GA is effective in improving flood prediction accuracy. The conclusion of this research is that feature selection using GA can identify the attributes that have the most influence on the occurrence of floods in Samarinda. The application of this feature selection resulted in a significant increase in the accuracy of the SVM algorithm for flood prediction.

Keywords: Classification, Support Vector Machine, Genetic Algorithm, Data Mining, Flood

Pendahuluan

Banjir merupakan bencana alam yang dapat diprediksi datangnya dengan memperhatikan curah hujan dan aliran air dan pada umumnya terjadi karena curah hujan tinggi terhadap suatu daerah, namun banjir juga bisa terjadi karena kondisi lingkungan seperti berkurangnya lahan terbuka hijau. Ancaman banjir sering terjadi di beberapa provinsi di Indonesia, khususnya Kalimantan Timur. Di Ibu Kota Kalimantan Timur sendiri yakni Samarinda, banjir menjadi fenomena rutin yang kerap kali melanda setiap tahun.

Banjir di klasifikasikan menurut proses penyebabnya untuk membantu dalam meningkatkan akurasi estimasi frekuensi terjadinya banjir dan besaran banjir (Tarasova et al., 2019). Data Mining dalam pemanfaatannya juga dapat dihubungkan dengan berbagai aspek termasuk bencana alam, khususnya banjir. Data mining mempunyai peran penting dalam menyatukan teknologi dan penelitian. Dapat mengidentifikasi aturan asosiasi dengan klasifikasi serta pengenalan bekerja dengan kategorisasi yang mendapatkan beberapa hasil buruk, rata-rata dan baik (Mian & Ghabban, 2022).

Support Vector Machine (SVM) adalah salah satu algoritma yang paling sering digunakan pada prediksi/klasifikasi dalam data mining. SVM adalah salah satu algoritma yang digunakan untuk klasifikasi data menggunakan hyperplane (Fitriana & Sibaroni, 2020). Namun pada penelitian lain yang dilakukan oleh (Abdullah & Utami, 2018) algoritma SVM mempunyai tingkat akurasi yang rendah. Oleh karena itu akan diterapkan algoritma optimasi untuk membuat performa yang dihasilkan dapat lebih baik dengan menggunakan feature selection.

Genetic Algorithm adalah salah satu algoritma optimasi yang kuat dan dapat digunakan pada berbagai studi kasus dengan pengunaan prinsip teori evolusi. Algoritma ini sering digunakan untuk menemukan solusi optimal pada kasus yang sederhana samapai yang rumit (Elva, 2019).

Dari beberapa uraian diatas maka penelitian ini menggunakan metode Genetic Algorithm untuk menyeleksi fitur mana yang paling berpengaruh pada dataset banjir kota Samarinda dan untuk mencari nilai perubahan yang terjadi terhadap model klasifikasi banjir ketika menggunakan fitur dataset keseluruhan dan ketika menggunakan fitur yang telah diseleksi oleh GA.

Metode

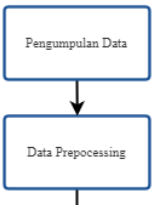
Pada penelitian ini teknik analisis datanya akan disesuaikan dengan metode CRISP-DM (Cross-Industry Standard Process for Data Mining) merupakan proses model standar yang sering digunakan dalam data mining(Schröer et al., 2021).

Penelitian ini menggunakan data sekunder Banjir Kota Samarinda periode tahun 2019-2022 yang diperoleh secara langsung oleh peneliti melalui observasi langsung ke BPBD (Badan Penanggulangan Bencana Daerah) Kota Samarinda dan BMKG Kota Samarinda.

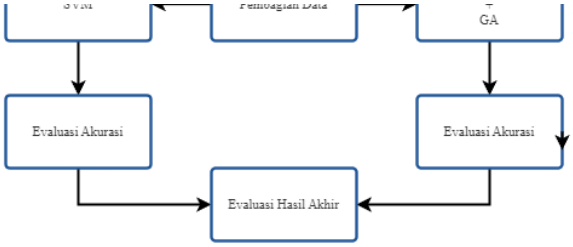
Setelah itu, akan dilakukan penerapan model SVM dan SVM + GA menggunakan python. Hasil dari kedua model tersebut akan dilihat nilai akurasinya dan akan di evaluasi untuk mengetahui apakah terjadi peningkatan akurasi dengan menggunakan optimasi GA. Alur penelitian ditunjukan pada gambar I.

Untuk menghindari pengolahan data yang tidak diperlukan dilakukan data pre-pocessing dengan beberapa tahapan yang harus dilalui seperti data integration, data selection, data transformation, data cleaning, dan data balancing.

1. Data Integration
   Tahapan ini dilakukan dengan penggabungan data BMKG yang memiliki 10 atribut dan data BPBD Kota Samarinda yang memiliki 11 atribut. Proses penggabungan tersebut disesuaikan agar data BMKG dan BPBD dapat saling melengkapi untuk menciptakan dataset yang baik pada klasifikasi banjir.



Gambar 1. Alur Penelitian

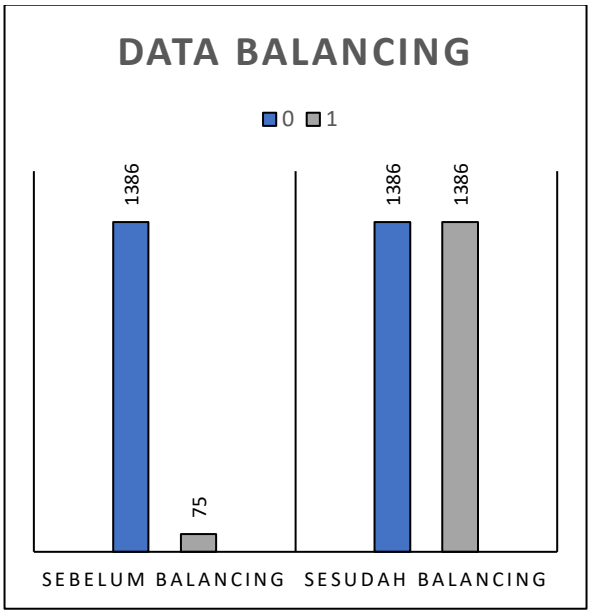


2. Data Selection
   Tahap ini dilakukan untuk memilih atribut apa saja yang akan digunakan dan pada penelitian ini atribut yang akan digunakan berjumlah 10 meliputi Temperatur-maksimum(Tn), Temperatur-minimum(Tx), Temperatur-rata-rata(Tavg), Kelembapan-rata-rata(RH_avg), Curah-hujan(RR), Lamanya-penyinaran-matahari(ss), Arah-angin-maksimum(ff_x), Kecepatan-angin-maksimum(ddd_x), Kecepatan-angin-rata-rata(ff_avg), Arah-angin-terbanyak(ddd_car) dan Terjadi-Banjir.
3. Data Transformation
   Pada tahap ini dilakukan transformasi pada atribut Arah Angin dikarenakan algoritma yang digunakan pada penelitian ini tidak dapat mengolah atribut dalam bentuk string, oleh karena itu atribut yang semula dalam bentuk string akan di tranformasikan menjadi atribut numerik.

ddd_car (Sebelum Transformasi)
E
C
S
NE
NW
N

ddd_car (Sesudah Transformasi)
45
0
225
90
360
135

4. Data Cleaning
   Pembersihan data dilakukan dengan cara membersihkan data yang kosong atau tidak lengkap sehingga pada saat proses pemodelan dapat memberikan hasil akurasi yang maksimal. Dalam penelitian ini salah satu tugas data cleaning adalah menangani nilai kosong pada data. Data yang memiliki nilai kosong akan digantikan dengan nilai rata-rata dari masing-masing kolom atribut.
5. Data Balancing
   Sebelum melakukan data balancing, terdapat ketidakseimbangan yang signifikan antara kelas terjadi banjir dan kelas tidak terjadi banjir. Jumlah sampel pada kelas terjadi banjir hanya 75, sementara jumlah sampel pada kelas tidak terjadi banjir sebanyak 1386. Untuk mengatasi ketidakseimbangan ini maka akan menggunakan teknik oversampling agar jumlah sampel pada kelas terjadi banjir sejajar dengan jumlah sampel pada kelas tidak terjadi banjir, yaitu sebanyak 1386. Setelah melakukan balancing dengan teknik oversampling, jumlah sampel pada kelas terjadi banjir naik menjadi 1386. Dengan demikian distribusi data menjadi lebih seimbang antara kelas terjadi banjir dan kelas tidak terjadi banjir. Setelah melakukan data balancing maka total data final berjumlah 2772 record.



Gambar 2. Hasil Data Balancing

Sebelum dilakukan modelling dataset dibagi menjadi dua bagian yaitu data training dan data testing. Data training digunakan saat melakukan modelling sedangkan data testing digunakan saat mengevaluasi model yang telah dibangun. Pada penelitian ini digunakan teknik K-Fold Cross Validation dengan nilai K=10.

Selanjutnya pemodelan SVM dan SVM + GA akan menggunakan python untuk mengetahui tingkat akurasi tiap-tiap model.

Hasil Dan Pembahasan

Setelah proses pengumpulan data dilakukan akan dilanjutkan dengan data preprocessing yang mana didalamnya terdapat beberapa tahapan yaitu data integration, data selection, data transformation, data cleaning dan data balancing untuk menghasilkan data yang lebih baik dan akurat. Berdasarkan tahapan yang sudah dilakukan maka dilanjutkan penerapan model SVM untuk membandingkan nilai akurasi sebelum dan sesudah dioptimasi dengan GA.

Berdasarkan hasil pengujian, maka diperoleh hasil akurasi dari pengujian pemodelan SVM dengan teknik cross validation K=10 dan confusion matrixnya.

Tabel 1. Hasil Akurasi SVM

Fold	TP	FP	TN	FN	Akurasi
1	124	130	9	15	47.48%
2	115	124	15	24	46.67%
3	118	122	17	20	48.74%
4	106	101	38	32	51.99%
5	79	62	77	59	56.32%
6	110	91	48	28	57.04%

7	119	99	39	20	57.04%
8	118	114	24	21	51.26%
9	113	94	44	26	56.68%
10	115	105	33	24	53.43%
Rata-rata	1117	1042	344	269	52.71%

Pada akhir proses yaitu evaluasi terhadap confusion matrix, hasil yang diperoleh yaitu True Positive adalah 1117 data, True Negative adalah 344 data, False Positive adalah 1042 data, dan False Negative 269 data. Sehingga diperoleh hasil perhitungan sebagai berikut:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

$$= \frac{1117 + 344}{1117 + 344 + 1042 + 269} \times 100\%$$

$$= \frac{1461}{2772} \times 100\% = 52.71\%$$

Selanjutnya pada saat melakukan optimasi seleksi fitur dengan menggunakan GA, berhasil diidentifikasi bahwa terdapat 3 fitur terbaik yang memberikan kontribusi signifikan dalam meningkatkan akurasi model SVM. Fitur-fitur tersebut adalah Kelembapan, Lamanya-penyinaran-matahari, dan Kecepatan-angin. Fitur-fitur tersebut akan dimasukkan ke dalam model SVM untuk mengevaluasi peningkatan akurasi yang dapat dicapai. Dengan memasukkan fitur-fitur terbaik ke dalam model SVM, didapati hasil pada tiap-tiap akurasi pada tiap-tiap fold dan confusion matrixnya sebagai berikut.

Tabel 2. Hasil Akurasi SVM + GA

Fold	TP	FP	TN	FN	Akurasi
1	83	23	116	56	71.58%
2	90	33	106	49	70.50%
3	110	29	110	28	79.42%
4	98	47	92	40	68.59%
5	85	56	83	53	60.65%
6	88	82	57	50	52.35%
7	103	60	78	36	65.34%
8	94	58	80	45	62.82%
9	103	63	75	36	64.26%
10	93	48	90	46	66.06%
Rata-rata	947	499	887	439	66.16%

Setelah melakukan evaluasi dengan menggunakan confusion matrix, diperoleh hasil 947 data yang terklasifikasikan True Positive, 887 data yang terklasifikasikan True Negative, 499 data yang terklasifikasikan sebagai False Positive, dan 439 data yang terklasifikasikan sebagai False Negative. Dengan demikian, hasil perhitungan akhir adalah sebagai berikut.

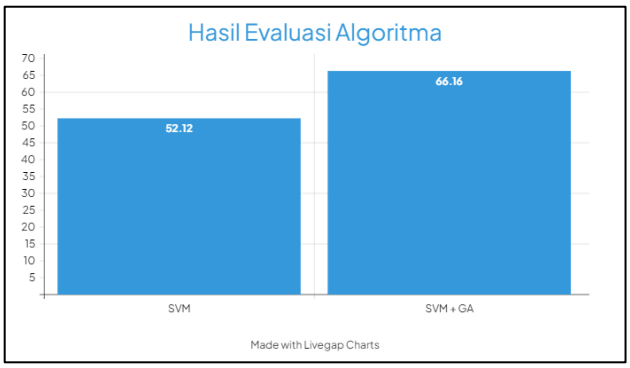
$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%$$

$$= \frac{947 + 887}{947 + 887 + 499 + 439} \times 100\%$$

$$= \frac{1834}{2772} \times 100\% = 66{,}16\%$$

Pembahasan

Pada penerapan algoritma SVM ada dua skema yang dilakukan, dimana skema pertama menggunakan data yang tidak menggunakan seleksi fitur GA dan skema kedua menggunakan data yang telah dilakukan seleksi fitur GA. Sebagai perbandingan dapat dilihat pada grafik berikut.



Gambar 3. Perbandingan Hasil Dua Skema Algoritma

Hasil pengujian yang dilakukan dengan menggunakan teknik confusion matrix didapati bahwa terjadi peningkatan performa pada algoritma SVM setelah menerapkan seleksi fitur dengan GA dimana nilai akurasi meningkat 13,45%.

Simpulan

Berdasarkan penelitian yang telah dilakukan makan dapat diambil beberapa kesimpulan sebagai berikut:

1. Hasil dari penerapan seleksi fitur menggunakan GA pada data banjir kota Samarinda ditemukan bahwa ada 3 atribut yang memiliki pengaruh paling signifikan terhadap terjadinya banjir yaitu kelembapan, lamanya-penyinaran-matahari dan kecepatan-angin.
2. Dari hasil pengujian penerepan selski fitur pada SVM ditemukan bahwa tingkat akurasi algoritma SVM

meningkat sebesar 13.45% dimana sebelum menerapkan seleksi fitur menggunakan GA akurasi yang didapatkan hanya 52,71% dan sesudah menerapkan seleksi fitur GA akurasinya menjadi 66,16%. Hal ini membuktikan bahwa penerapan seleksi fitur terhadap algoritma klasifikasi terbukti meningkatkan akurasi algoritma tersebut.

Dengan demikian, kinerja model algoritma SVM dengan optimasinya GA dapat bekerja dengan baik karena menghasilkan nilai evaluasi sebesar 66.16%. Pada penerapan seleksi fitur menggunakan GA pada data banjir kota Samarinda telah berhasil mengidentifikasi atribut-atribut penting dan meningkatkan akurasi algoritma klasifikasi SVM.

Saran

Berdasarkan kesimpulan yang telah diuraikan terdapat beberapa saran untuk penelitian selanjutnya, diantaranya:

1. Untuk penelitian selanjutnya disarankan menggunakan algoritma klasifikasi lain seperti Naïve Bayes, Random Forest, Decision Tree dan algoritma yang lainnya.
2. Pada peneltian selanjutnya bisa menggunakan fitur-fitur lain yang mempengaruhi banjir selain data dari BMKG. Sebagai contoh yaitu drainase tata kota, dataran daerah dan pencemaran sungai.

Pustaka Acuan

Abdullah, R. K., & Utami, E. (2018). Studi Komparasi Metode SVM dan Naive Bayes pada Data Bencana Banjir di Indonesia pembaca ataupun peneliti bisa melihat pola yang tersembunyi di Indonesia. *Tecnoscienza*, *3*(1), 103–122.

Elva, Y. (2019). Sistem Penjadwalan Mata Pelajaran Menggunakan Algoritma Genetika. *Jurnal Teknologi Informasi*, *3*(1), 49. https://doi.org/10.36294/jurti.v3i1.687

Fitriana, D. N., & Sibaroni, Y. (2020). Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM). *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(5), 846–853. https://doi.org/10.29207/resti.v4i5.2231

Mian, T. S., & Ghabban, F. (2022). Competitive Advantage: A Study of Saudi SMEs to Adopt Data Mining for Effective Decision Making. *Journal of Data Analysis and Information Processing*, *10*(03), 155–169. https://doi.org/10.4236/jdaip.2022.103010

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

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