Universitas Ahmad Dahlan

Jurnal Ilmiah Matematika
Vol. 10, No. 2, Oktober 2023, pp. 75-82
ISSN 2774-3241
http://journal.uad.ac.id/index.php/Konvergensi

Implementasi Metode SVM-PSO Dengan Fitur Selection Variance Threshold Pada Klasifikasi Penyakit Diabetes Mellitus

Pratiwi Kistiya Ningrum [a,1,*], Joko Purwadi [b,2]

[*a] Universitas Ahmad Dahlan, Yogyakarta, Indonesia;
[b] Universitas Ahmad Dahlan, Yogyakarta, Indonesia.
[1] pratiwi2000015037@webmail.uad.ac.id; [2] Joko@math.uad.ac.id.
***Correspondent Author**

Received: *Revised:* *Accepted:*

KATAKUNCI
SVM-PSO
Variance Threshold
Klasifikasi
Diabetes Mellitus
Optimasi

ABSTRAK

Pada penelitian ini membahas tentang kasus klasifikasi pada data penyakit diabetes. Metode yang digunakan dalam penelitian ini adalah metode *Support Vector Machine* yang dioptimalkan dengan algoritma *Particle Swarm Optimization* guna memperoleh parameter terbaik dengan kombinasi seleksi fitur menggunakan *Variance Threshold*. Penelitian ini bertujuan untuk mengetahui cara kerja dan hasil akurasi dari metode *Support Vector Machine* dengan optimasi *Particle Swarm Optimization* menggunakan seleksi fitur *Variance Threshold*. Hasil penelitian menggunakan kombinasi metode tersebut menunjukkan hasil akurasi sebesar 80%. Hasil akurasi tersebut lebih tinggi jika dibandingkan dengan metode *Support Vector Machine* tunggal tanpa optimasi dan seleksi fitur dengan akurasi sebesar 76%. Meningkatkan akurasi sebesar 4% dari 76% menjadi 80%.

KEYWORDS

SVM-PSO
Variance Threshold
Classification
Diabetes Mellitus
Optimization

ABSTRACT

This study discusses classification cases in diabetes data. The method used in this research is the Support Vector Machine method which is optimized with the Particle Swarm Optimization algorithm to obtain the best parameters with a combination of feature selection using Variance Threshold. This research aims to find out how it works and the accuracy results of the Support Vector Machine method with Particle Swarm Optimization using Variance Threshold feature selection. The results of research using a combination of these methods show accuracy results of 80%. The accuracy results are higher when compared to the single Support Vector Machine method without optimization and feature selection with an accuracy of 76%. Increased accuracy by 4% from 76% to 80%.

Pendahuluan

Di era teknologi saat ini, jumlah data yang tersedia di seluruh dunia mengalami pertumbuhan yang sangat pesat. Data dalam skala besar ini akan menjadi sia-sia apabila tidak dimanfaatkan secara optimal. Dataset merupakan kumpulan data dalam jumlah besar yang

 ISSN 2774-3241

seringkali sulit untuk disimpan, dikelola, dianalisis, maupun divisualisasikan [1]. Oleh karena itu, untuk mengolah dan mengekstraksi informasi penting dari dataset, digunakan metode yang dikenal dengan data mining [2].

Data mining adalah metode pembelajaran komputer yang memanfaatkan teknik pengenalan pola seperti metode statistik dan matematika guna menganalisis data secara otomatis serta melatih komputer untuk menemukan pola baru yang berguna. Data mining memungkinkan pengguna untuk menemukan dan menafsirkan pola pengambilan keputusan dari informasi yang telah diolah [3]. Metode ini telah diterapkan secara luas di berbagai bidang industri, seperti telekomunikasi, kesehatan, bioinformatika, perbankan, pemasaran, biologi, asuransi, perencanaan kota, penanggulangan bencana, klasifikasi dokumen, hingga transportasi.

Salah satu pendekatan dalam data mining adalah machine learning, yakni cabang kecerdasan buatan yang berfokus pada pengembangan algoritma agar komputer mampu belajar dari data empiris. Salah satu metode *machine learning* yang efektif untuk kasus prediksi dan klasifikasi adalah *Support Vector Machine* (SVM). Metode SVM bertujuan untuk menemukan *hyperplane* optimal yang mampu memisahkan dua kelas data secara maksimal, sehingga menghasilkan kemampuan generalisasi yang baik. SVM juga mampu menangani data non-linear dengan bantuan fungsi kernel. Karena menggunakan vektor untuk menghitung jarak, SVM memiliki keunggulan dalam hal kecepatan proses dan akurasi yang tinggi, meskipun pemilihan parameter optimal masih menjadi tantangan [4].

Untuk mengatasi permasalahan tersebut, digunakan algoritma *Particle Swarm Optimization* (PSO) sebagai metode optimasi guna menentukan parameter SVM yang optimal. PSO bekerja dengan mengeksplorasi ruang parameter untuk menemukan kombinasi dengan akurasi terbaik. Dalam beberapa studi, PSO terbukti lebih kompetitif dibandingkan algoritma lain seperti algoritma genetika dan C4.5 [5]. Agar performa model semakin optimal, proses seleksi fitur juga diperlukan. Seleksi fitur merupakan bagian dari tahap pra-pemrosesan dalam data mining. Salah satu metode seleksi fitur yang digunakan adalah *Variance Threshold*, yang bertujuan untuk mengeliminasi fitur dengan variansi rendah karena dianggap kurang informatif.

Penelitian terkait kombinasi metode SVM dan PSO telah banyak dilakukan. [6] menerapkan SVM berbasis PSO untuk prediksi penyakit hati dengan pendekatan ELTA (Ekstraksi, Loading, Transformasi, dan Analisis), dan membuktikan bahwa model PSO-SVM memberikan akurasi terbaik dibandingkan metode lainnya. Penelitian lain oleh [7] menunjukkan bahwa kombinasi SVM dan PSO memberikan hasil klasifikasi yang lebih akurat dibandingkan dengan penggunaan SVM tunggal. Hasil serupa juga ditemukan dalam studi [8] pada prediksi penyakit

 ISSN 2774-3241

jantung, di mana SVM dengan PSO menghasilkan nilai recall tertinggi sebesar 96%. Selanjutnya, [9] menggunakan kombinasi SVM-PSO yang dioptimalkan dengan simulated annealing (SA) untuk memprediksi penurunan lubang fondasi, sementara [10] mengembangkan pemilihan fitur optimal dengan *multi-objective* PSO untuk meningkatkan kinerja SVM sekaligus mengurangi kompleksitas perhitungan.

Meskipun telah banyak dilakukan penelitian kombinasi SVM dan PSO, namun penggabungan keduanya dengan metode seleksi fitur *Variance Threshold* khususnya pada kasus klasifikasi penyakit diabetes mellitus masih jarang dilakukan. Oleh karena itu, penelitian ini berfokus untuk meningkatkan akurasi *Support Vector Machine* dengan seleksi fitur *Variance Threshold* berdasarkan algoritma *Particle Swarm Optimization*. Dengan pendekatan ini, diharapkan mampu menghasilkan model klasifikasi penyakit diabetes mellitus yang lebih efisien dan akurat, serta dapat berkontribusi dalam sistem pendukung keputusan di bidang kesehatan.

Metode

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental yang bertujuan untuk meningkatkan akurasi klasifikasi penyakit diabetes mellitus. Data yang digunakan merupakan data sekunder yang diperoleh dari dataset terbuka, kemudian dilakukan proses input dan pra-pemrosesan data. Tahap selanjutnya adalah seleksi fitur menggunakan metode *Variance Threshold* untuk mengeliminasi fitur yang kurang informatif. Setelah fitur terseleksi, dilakukan optimasi parameter model klasifikasi menggunakan algoritma *Particle Swarm Optimization* (PSO) untuk memperoleh parameter optimal pada metode *Support Vector Machine* (SVM). Model SVM dengan parameter terbaik selanjutnya digunakan untuk proses klasifikasi. Evaluasi kinerja model dilakukan menggunakan confusion matrix dengan menghitung metrik seperti akurasi, presisi, recall, dan F1-score. Hasil evaluasi digunakan untuk menarik simpulan mengenai efektivitas kombinasi metode PSO-SVM dan seleksi fitur dalam meningkatkan performa klasifikasi.

Hasil dan Pembahasan

1. Pengolahan Data

Dataset diperoleh dari Kaggle yaitu Pima Indian Diabetes dengan format CSV (*Comma-Separated Values*) salah satu format file yang digunakan untuk menyimpan data tabular dalam bentuk teks yang sederhana.

a. *Cleaning* Data

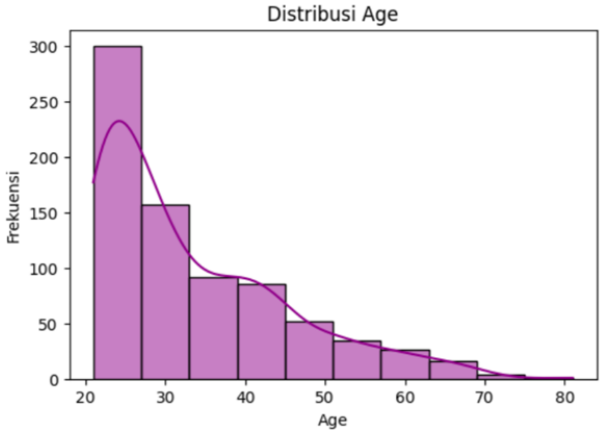
Pada tahap ini bertujuan untuk mengetahui nilai yang hilang untuk melakukan pemrosesan

 ISSN 2774-3241

teks dan mengatasi noise dalam data. Pada tahap ini untuk kasus data yang digunakan tidak terdapat *missing values* maupun data yang terduplikasi maka data tersebut adalah data yang siap untuk diolah.

2. *Eksploratory Data Analysis* (EDA)

Tahapan ini berfungsi untuk mengekslorasi statistik dari data. Visualisasi data menggunakan pustaka matplotlib, seaborn, plotly untuk mendapatkan wawasan yang lebih dalam mengenai karakteristik data tersebut. Pada visualisasi frekuensi distribusi umur menghasilkan plot sebagai berikut :



Gambar 1. Grafik Distribusi Umur

Berdasarkan Gambar 1 terlihat pada plot tersebut frekuensi pada umur 20 tahun hingga sebelum mencapai umur 30 tahun terbilang cukup tinggi hampir 300, sedangkan umur yang paling tinggi yaitu umur 80 tahun frekuensinya tidak melebihi 5 kemungkinan hanya satu atau dua orang saja.

3. *Preparation* Data

Pada tahap ini bertujuan untuk mempersiapkan data pelatihan model dengan beberapa langkah penting. Pertama, variabel x memiliki dimensi (768, 8), yang berarti memiliki 768 sampel data dan masing-masing sampel memiliki 8 fitur. Sementara itu, variabel y memiliki dimensi (768,), menunjukkan bahwa setiap sampel memiliki satu nilai target yang akan diprediksi.

Langkah berikutnya adalah standarisasi data, yang bertujuan untuk menyamakan skala dari fitur-fitur dalam variabel x. Standarisasi ini penting untuk memastikan bahwa setiap fitur berkontribusi secara proporsional dalam pelatihan model, serta untuk mempercepat konvergensi model selama proses pelatihan.

Setelah standarisasi, data dibagi menjadi dua subset: data pelatihan dan data pengujian. Pembagian dilakukan dengan alokasi 80% dari data untuk pelatihan dan 20% untuk pengujian. Oleh karena itu, variabel x dibagi menjadi $xtrain$ dengan dimensi (614, 8) untuk data pelatihan

 ISSN 2774-3241

dan $xtest$ dengan dimensi (154, 8) untuk data pengujian. Demikian pula, variabel y dibagi menjadi $ytrain$ dengan dimensi (614,) untuk data pelatihan dan $ytest$ dengan dimensi (154,) untuk data pengujian. Pembagian ini memastikan bahwa model akan dilatih pada 80% dari data dan diuji pada 20% data yang tidak terlihat selama pelatihan, sehingga memberikan gambaran yang akurat tentang kinerja model pada data baru.

4. Seleksi Fitur Variance Threshold

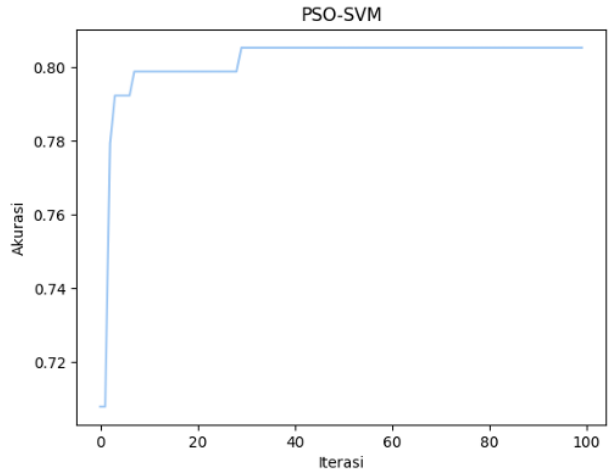
Tahap seleksi fitur *variance threshold* digunakan untuk mengambil fitur yang dianggap penting dan menghilangklan fitur yang nilainya dibawah batas ambang dari Pima Indian Diabetes Dataset. Pemilihan parameter ambang batas dalam metode pemilihan fitur bisa sangat subjektif dan seringkali memerlukan penyesuaian empiris untuk mencapai performa yang diinginkan pada kumpulan data tertentu [11]. Ambang batas 0.02 dipilih sebagai keseimbangan yang baik antara terlalu banyak fitur yang dipertahankan dan terlalu banyak fitur yang dibuang.

Data awal memiliki dimensi (768, 8), yang berarti memiliki 768 sampel data dan masing-masing sampel memiliki 8 fitur. Sementara itu, variabel y memiliki dimensi (768,), menunjukkan bahwa setiap sampel memiliki satu nilai target yang akan diprediksi. Setelah dilakukan selekai fitur, variabel x memiliki 768 sampel data, masing-masing dengan 6 fitur. Variabel y tetap memiliki sampel data sejumlah 768, mencerminkan satu nilai target per sampel data.

5. Penerapan Model PSO-SVM

Pada tahap ini dilakukan proses penerapan model dengan data yang telah dipilih fiturnya menggunakan *Variance Threshold*. Tahap pertama yaitu mengdefinisikan model *Support Vector Machine* (SVM). Model SVM dibentuk dengan parameter C *dan* γ. Parameter C digunakan untuk mengontrol trade-off antara margin yang lebih besar dan kesalahan klasifikasi. Parameter gamma merupakan parameter kernel yang menentukan seberapa besar pengaruh yang dimiliki suatu sampel pelatihan. Nilai terbaik yang dihasilkan parameter C =3.46233216 dan nilai terbaik yang dihasilkan oleh parameter γ = 0.02655854. Kemudian berikut plot akurasi per iterasi yang dihasilkan :

 ISSN 2774-3241



Gambar 2. Grafik PSO-SVM

Berdasarkan Gambar 2 plot tersebut garis berwarna biru menunjukkan bahwa hasil akurasi meningkat tajam pada iterasi awal atau iterasi ke-1 mencapai sekitar 0.72 hingga mendekati 0.79 pada sekitar iterasi ke-5. Kemudian akurasi sedikit mengalami peningkatan dan kestabilan mencapai sekitar 0.80 mulai dari iterasi ke 40 sampai iterasi ke-100. Terdapat peningkatan akurasi yang signifikan pada awal iterasi, yang menunjukkan bahwa PSO mampu menemukan parameter SVM yang lebih sesuai dengan cepat. Setelah 40 kali iterasi, akurasinya cenderung stabil dan tidak menunjukkan peningkatan yang signifikan. Hal ini menunjukkan bahwa algoritma telah konvergen, atau mencapai titik optimal dimana akurasi tidak meningkat secara signifikan dengan iterasi yang berkelanjutan.

6. Mengevaluasi Model dengan Confussion Matriks

Tahap ini dilakukan untuk mengetahui akurasi yang diperoleh dari penelitian yang sudah dilakukan berikut adalah table confusion matrix yang diperoleh :

Tabel 1. Confusion Matrix

Klasifikasi		***Kelas Hasil Prediksi***		
		Ya	***Tidak***	***Jumlah***
Kelas Aktual	Ya	90	9	99
	Tidak	20	33	55

Menghasilkan nilai akurasi,precission,recall, dan f1-score sebagai berikut :

Tabel 2. Confusion Matrix

Confussion Matriks	***Nilai***
Akurasi	80%
Presisi	0.79
Recall	0.60
F1-Score	0.68

 ISSN 2774-3241

7. Hasil Menggunakan Algoritma SVM Tunggal

Penerapan algoritma SVM tunggal untuk data yang sama dengan pembagian data latih dan data uji dengan perbandingan 80:20, hasil akurasi yang dihasilkan sebagai berikut

$$\text{Akurasi} = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100\%$$

$$= \frac{32+85}{154} \times 100\%$$

$$= 75.9\%$$

$$\approx 76\%$$

Dalam hal ini metode SVM tunggal dapat mengklasifikasikan data dengan tingkat akurasi yang cukup baik.

Simpulan

Berdasarkan hasil penelitian, metode *Support Vector Machine* (SVM) yang dikombinasikan dengan algoritma *Particle Swarm Optimization* (PSO) dan seleksi fitur *Variance Threshold* menunjukkan kinerja yang baik dalam mengklasifikasikan data penyakit diabetes. Proses kerja metode ini dimulai dengan menghitung varians pada setiap fitur, menentukan nilai threshold, dan menghapus fitur dengan varians di bawah ambang batas. Hasil seleksi fitur menyisakan 7 dari 9 fitur awal, yang terbukti meningkatkan akurasi model. Selanjutnya dilakukan optimasi parameter SVM menggunakan algoritma PSO, yang mencakup inisialisasi partikel, perhitungan nilai fitness, serta pencarian nilai terbaik (pbest dan gbest) untuk memperoleh parameter optimal berupa nilai C dan γ. Pengujian menggunakan parameter hasil optimasi tersebut menunjukkan bahwa akurasi klasifikasi meningkat hingga mencapai 80%, lebih tinggi dibandingkan metode SVM tanpa optimasi dan seleksi fitur yang hanya menghasilkan akurasi sebesar 76%. Parameter terbaik diperoleh dengan jumlah partikel sebanyak 150, jumlah iterasi 100, rentang nilai C sebesar 0.1–100, dan γ sebesar 0.001–10, serta konstanta akselerasi c_1 dan c_2 masing-masing sebesar 1. Dengan demikian, penerapan kombinasi PSO dan seleksi fitur *Variance Threshold* terbukti mampu meningkatkan performa klasifikasi penyakit diabetes secara signifikan.

Daftar Pustaka

[1] C. Sreedhar, N. Kasiviswanath, and P. Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop," *J. Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0087-2.

[2] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," pp. 203–207. doi: 10.1109/ICMIRA.2013.45.

[3] M. North, *Data Mining for the Masses*. 2012. [Online]. Available: http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/DataMiningForTheMasses.pdf%5Cnhttps://sites.google.com/site/

dataminingforthemasses/

[4] V. N. Vapnik, *The Nature of Statistical Learning Theory*. 2nd ed.: Springer Verlag, 1995.

[5] T. Sousa, A. Silva, and A. Neves, "Particle Swarm based Data Mining Algorithms for classification tasks," *Parallel Comput.*, vol. 30, no. 5–6, pp. 767–783, 2004, doi: 10.1016/j.parco.2003.12.015.

[6] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," *Informatics Med. Unlocked*, vol. 17, no. October, p. 100255, 2019, doi: 10.1016/j.imu.2019.100255.

[7] D. Saputra, W. S. Dharmawan, and W. Irmayani, "Performance Comparison of the SVM and SVM-PSO Algorithms for Heart Disease Prediction," *Int. J. Adv. Data Inf. Syst.*, vol. 3, no. 2, pp. 74–86, 2022, doi: 10.25008/ijadis.v3i2.1243.

[8] D. Saputra, W. Irmayani, D. Purwaningtias, and J. Sidauruk, "A Comparative Analysis of C4.5 Classification Algorithm, Naïve Bayes and Support Vector Machine Based on Particle Swarm Optimization (PSO) for Heart Disease Prediction," *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 2, pp. 84–95, 2021, doi: 10.25008/ijadis.v2i2.1221.

[9] Z. Song, S. Liu, M. Jiang, and S. Yao, "Research on the Settlement Prediction Model of Foundation Pit Based on the Improved PSO-SVM Model," *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/1921378.

[10] I. Behravan, O. Dehghantanha, and S. H. Zahiri, "An optimal SVM with feature selection using multi-objective PSO," *1st Conf. Swarm Intell. Evol. Comput. CSIEC 2016 - Proc.*, vol. 2016, pp. 76–81, 2016, doi: 10.1109/CSIEC.2016.7482135.

[11] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007, doi: 10.1093/bioinformatics/btm344.

[12] H. Bai, "Preparing Teacher Education Students to Integrate Mobile Learning into Elementary Education," *TechTrends*, vol. 63, no. 6, pp. 723–733, Nov. 2019, doi: 10.1007/s11528-019-00424-z.

[13] F. Giannakas, A. Papasalouros, G. Kambourakis, and S. Gritzalis, "A comprehensive cybersecurity learning platform for elementary education," *Inf. Secur. J. A Glob. Perspect.*, vol. 28, no. 3, pp. 81–106, May 2019, doi: 10.1080/19393555.2019.1657527.

[14] R. M. Vink *et al.*, "Self-reported adverse childhood experiences and quality of life among children in the two last grades of Dutch elementary education," *Child Abuse Negl.*, vol. 95, p. 104051, Sep. 2019, doi: 10.1016/j.chiabu.2019.104051.