

Klasifikasi Penyakit Diabetes Retinopati Menggunakan *Support Vector Machine* dengan Algoritma *Grid Search Cross-Validation*

Fatmawati, Nur Azizah Komara Rifai*

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

ARTICLE INFO

Article history :

Received : 19/2/2023

Revised : 25/6/2023

Published : 14/7/2023



Creative Commons Attribution-
NonCommercial-ShareAlike 4.0
International License.

Volume : 3

No. : 1

Halaman : 79 - 86

Terbitan : Juli 2023

ABSTRAK

Ketepatan dalam menentukan objek pada kelas yang sesuai merupakan tantangan pada metode klasifikasi. Support Vector Machine menjadi metode klasifikasi yang mampu melakukan klasifikasi untuk data yang tidak melewati tahap pembelajaran dan mampu menghasilkan model klasifikasi meskipun data training relatif sedikit dengan tetap menghasilkan akurasi yang baik. Peningkatan performa klasifikasi yang dihasilkan oleh metode SVM ditentukan dari pengaturan parameter. Oleh karena itu, algoritma grid search digunakan untuk melihat parameter yang optimal. Masalah diabetes

retinopati menjadi penting karena menjadi salah satu penyebab kebutaan total di beberapa negara. Penelitian ini bertujuan untuk menerapkan metode SVM dengan menerapkan algoritma grid search cross-validation. Grid search merupakan metode penentuan kombinasi model dan hyperparameter dengan melakukan uji coba satu persatu kombinasi model dan melakukan validasi pada setiap kombinasi. Metode grid search dipasangkan dengan cross-validation untuk mendapatkan model terbaik dalam mengklasifikasikan status penyakit pada pasien diabetes retinopati. Dengan menggunakan kernel linear serta 5-fold cross-validation diperoleh hasil sebanyak 30 model. Model terbaik untuk kasus klasifikasi penyakit diabetes retinopati ini didapatkan dari parameter cost 10 sebesar 74,8%.

Kata Kunci : Diabetes Retinopati; Grid Search Cross-validation; Support Vector Machine.

ABSTRACT

Accuracy in determining the object in the appropriate class is a challenge in the classification method. Support Vector Machine is a classification method that is capable of classifying data that does not go through the learning stage and is able to produce a classification model even though there is relatively little data training while still producing good accuracy. The classification performance improvement produced by the SVM method is determined from the parameter settings. Therefore, the grid search algorithm is used to find the optimal parameters. The problem of diabetic retinopathy is important because it is one of the causes of total blindness in several countries. This study aims to apply the SVM method by applying the grid search cross-validation algorithm. Grid search is a method of financing combination models and hyperparameters by testing a single combination of models and validating each combination. The grid search method is paired with cross-validation to obtain the best model in classifying disease status in diabetic retinopathy patients. By using a linear kernel and 5-fold cross-validation, the results obtained for 30 models. The best model for cases of diabetes retinopathy classification is obtained from the cost parameter 10 of 74.8%.

Keywords : Diabetic Retinopathy; Grid Search Cross-validation; Support Vector Machine.

@ 2023 Jurnal Riset Ekonomi Syariah Unisba Press. All rights reserved.

A. Pendahuluan

Berbagai bidang industri menggunakan *data mining* untuk meringankan pekerjaan dengan cara meningkatkan sistem dan layanan data yang lebih efisien dan efektif [1]. Salah satu tugas *data mining* adalah klasifikasi. Masalah klasifikasi seringkali dikaitkan dengan pengambilan keputusan [2]. Tantangan utama dari proses klasifikasi adalah ketepatan dalam menentukan objek pada kelas atau kategori yang sudah ditentukan sebelumnya. Klasifikasi melibatkan tahap pembelajaran yang akan membangun model atau aturan klasifikasi untuk kemudian digunakan pada tahap prediksi [3].

Pada beberapa hasil pengujian, semakin banyak data yang digunakan akan meningkatkan pula akurasi yang didapatkan [4]. Klasifikasi dapat dilakukan dengan menggunakan berbagai jenis metode klasifikasi. Pada tahun 1992, *Support Vector Machine* (SVM) diperkenalkan oleh Vapnik sebagai rangkaian dari beberapa konsep-konsep unggulan dari bidang *pattern recognition*. SVM merupakan metode klasifikasi yang menawarkan kelebihan berupa generalisasi, *Curse of Dimensionality*, dan *Feasibility*. Dengan kelebihan tersebut, SVM mampu melakukan klasifikasi untuk data yang tidak melewati tahap pembelajaran [5] dan mampu menghasilkan model klasifikasi meskipun data *training* relatif sedikit [6].

Peningkatan performa klasifikasi yang dihasilkan oleh metode SVM ditentukan dari pengaturan parameter. Oleh karena itu, membangun model yang akurat membutuhkan pemilihan nilai optimal dari parameter pembelajarannya [7]. *Grid search* adalah metode optimasi parameter yang penggunaannya dikombinasikan dengan *cross-validation* [8].

Diabetes retinopati disebabkan oleh diabetes melitus yang ditampakan melalui retina mata. Penyakit ini menjadi salah satu penyebab kebutaan total di beberapa negara. *Community Eye Health* (2003) menyebutkan, penyakit yang menyerang retina menjadi faktor tertinggi yang menjadikan anak-anak di seluruh dunia mengalami kebutaan. Hal tersebut memberikan informasi bahwa diperlukan suatu sistem yang dapat mengidentifikasi jenis penyakit retina [9].

Berdasarkan latar belakang yang ada, rumusan masalah yang terbentuk adalah sebagai berikut: Bagaimana penerapan serta akurasi yang dihasilkan dari algoritma *Grid search Cross-validation* untuk mencari parameter optimal dari *Support Vector Machine* dalam mengklasifikasikan penyakit diabetes retinopati.

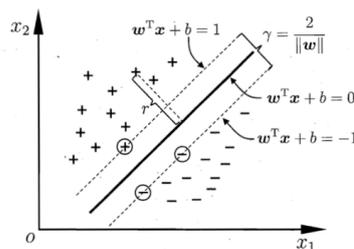
B. Metode Penelitian

Support Vector Machine

SVM berusaha menemukan garis pemisah (*hyperplane*) terbaik dari banyaknya jumlah garis yang tersedia untuk memisahkan data dari dua kelas yang berbeda secara *linear* (*Linearly separable data*). SVM menemukan *hyperplane* ini menggunakan *support vector* dan *margin* [1]. Ide dari metode SVM sebelumnya hanya diimplementasikan pada kasus dimana data pelatihan (*training*) dapat dipisahkan tanpa kesalahan. Akan tetapi dalam kondisi tertentu, beberapa data tidak dapat dipisahkan secara *linear*, sehingga diberikan penalti dengan menambahkan nilai *cost* (C). Penambahan nilai *cost* dalam metode SVM disebut sebagai *soft margin* karena digunakan sebagai solusi dari permasalahan data yang tidak dapat dipisahkan secara *linear* (*Linearly Non-separable Data*).

Linearly Separable Data

Linearly separable data merupakan data yang dapat dipisahkan ke dalam masing-masing kelas secara linear.



Gambar 1. *Linearly Separable Data*

Data terbagi menjadi dua kelas, data dengan tanda negatif sebagai kelas pertama dan data dengan tanda positif untuk kelas kedua. Data-data tersebut dipisahkan secara *linear* oleh garis pemisah. Himpunan vektor baris, $\mathbf{x}_i (i = 1, 2, \dots, m) \in \mathbf{R}^n$ dimana m merupakan banyak data *training* dan n merupakan banyak variabel *independent*, dengan $y_i \in \{1, -1\}$ merupakan kelas untuk data ke- i . Pada kasus data yang dapat dipisahkan secara *linear* $\mathbf{w}_j (j = 1, 2, \dots, n) \in \mathbf{R}^n$ dan $b \in \mathbf{R}$ sehingga fungsi *hyperlane* dari SVM didefinisikan sebagai berikut:

$$\mathbf{w}^T \mathbf{x} + b = 0 \tag{1}$$

Sehingga Vapnik dan Cortes (1995) menurunkan persamaan di atas menjadi [10]:

$$[(\mathbf{w}^T \mathbf{x}_i) + b] \geq 1 \text{ untuk } y_i = 1 \text{ dan } [(\mathbf{w}^T \mathbf{x}_i) + b] \leq -1 \text{ untuk } y_i = -1 \tag{2}$$

Bidang yang ditentukan oleh (1) terletak di antara dua bidang yang didefinisikan oleh $\mathbf{w}^T \mathbf{x} + b = 1$ dan $\mathbf{w}^T \mathbf{x} + b = -1$. Bidang dalam Persamaan (1) memisahkan satu kelas dari kelas yang lain dengan selisih $\frac{1}{\|\mathbf{w}\|}$ untuk setiap sisi. *Problem* optimasi SVM untuk klasifikasi *linear*, akan dirumuskan sebagai berikut [11]:

$$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \tag{3}$$

dengan syarat

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, m \tag{4}$$

dimana \mathbf{w} dan b adalah parameter-parameter yang akan dicari nilainya. Selanjutnya permasalahan tersebut diformulasikan ke dalam *quadratic programming* (QP) *problem* [12].

Solusi optimasi dapat diselesaikan dengan menggunakan fungsi *lagrange* sebagai berikut [10]:

$$L_{(\mathbf{w}, b, \alpha)} = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w}^T \cdot \mathbf{x}_i + b) - 1] \tag{5}$$

α_i adalah *lagrange multipliers*, $i=1, 2, \dots, m$.

Solusi dari masalah optimisasi dengan pembatas ditentukan dengan mencari *saddle point* dari fungsi *lagrangian* [13]. Fungsi ini harus diminimalkan terhadap \mathbf{w} dan b , dan memaksimalkan *lagrange multipliers* terhadap α_i . Dengan melakukan proses tersebut, akan didapatkan dua kondisi optimal seperti berikut [12]:

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i \mathbf{x}_i y_i \tag{6}$$

$$\frac{\partial L_p}{\partial b} = 0 \rightarrow \sum_{i=1}^m \alpha_i y_i = 0 \tag{7}$$

Menurut Santosa (2007) Persamaan (5) dapat disederhanakan menjadi [13]:

$$L_{(\mathbf{w}, b, \alpha)} = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^m \alpha_i y_i (\mathbf{w}^T \cdot \mathbf{x}_i) - b \sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \alpha_i \tag{8}$$

Berdasarkan Persamaan (6) dan (7), maka Persamaan (8) menjadi sebagai berikut:

$$L_d = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \tag{9}$$

Kemudian didapatkan *dual problem*:

$$\max_{\alpha} L_d = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{10}$$

dengan batasan, $\alpha_i \geq 0, i = 1, 2, \dots, m$ dan $\sum_{i=1}^m \alpha_i y_i = 0$.

Data *training* dengan $\alpha_i > 0$ terletak pada *hyperplane* disebut *support vector*. Data *training* yang tidak terletak pada *hyperplane* tersebut mempunyai $\alpha_i = 0$. Setelah solusi permasalahan *quadratic programming* ditemukan (nilai α_i), maka kelas dari data yang akan diprediksi atau data *testing* dapat ditentukan berdasarkan nilai fungsi berikut:

$$f(x_{testing}) = \text{sign}\left(\sum_{i=1}^{m_{sv}} \alpha_i y_i (x_i x_{testing}) + b\right) \tag{11}$$

Keterangan:

x_i : data yang merupakan *support vector* $i=1, 2, \dots, m_{sv}$

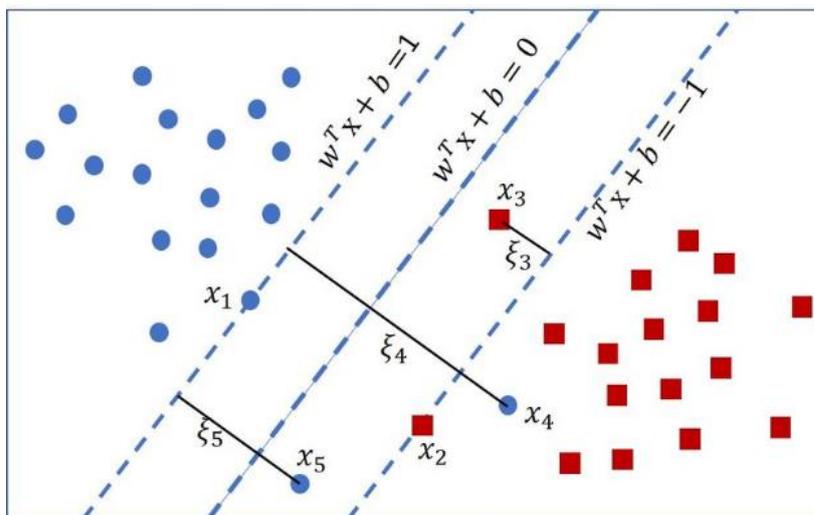
$x_{testing}$: data yang akan diprediksi kelasnya

m_{sv} : banyaknya data yang menjadi *support vector*

b : bias

Linearly Non-separable Data

Beberapa data terletak di luar kelas yang seharusnya, untuk mengklasifikasikan data yang tidak dapat dipisahkan secara *linear*, Persamaan (2) harus dimodifikasi karena tidak akan ada solusi yang ditemukan untuk beberapa data.



Gambar 2. Linearly Non-separable Data

Untuk menangani masalah tersebut, Persamaan (2) dapat dimodifikasi menjadi:

$$[(w^T x_i) + b] \geq 1 - \xi_i \text{ untuk } y_i = 1 \text{ dan } [(w^T x_i) + b] \leq -1 - \xi_i \text{ untuk } y_i = -1 \tag{12}$$

dengan $\xi_i \geq 0, i = 1, 2, \dots, m$. ξ_i merupakan variabel kesalahan dari sampel data ke- i . Metode klasifikasi ini dianggap sebagai *soft margin* karena mengklasifikasikan data dengan beberapa kesalahan. Klasifikasi sampel untuk pengujian diperoleh dengan menentukan tanda $w^T x + b$.

Berikut ini adalah pembatas yang sudah dimodifikasi oleh Gunn (1998) untuk kasus *non-separable* [14]:

$$y_i (w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, m \tag{13}$$

Formulasi SVM dengan *soft margin* dan kernel *linier* diberikan oleh:

$$\min_{w,b,q} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^m \xi_i \tag{14}$$

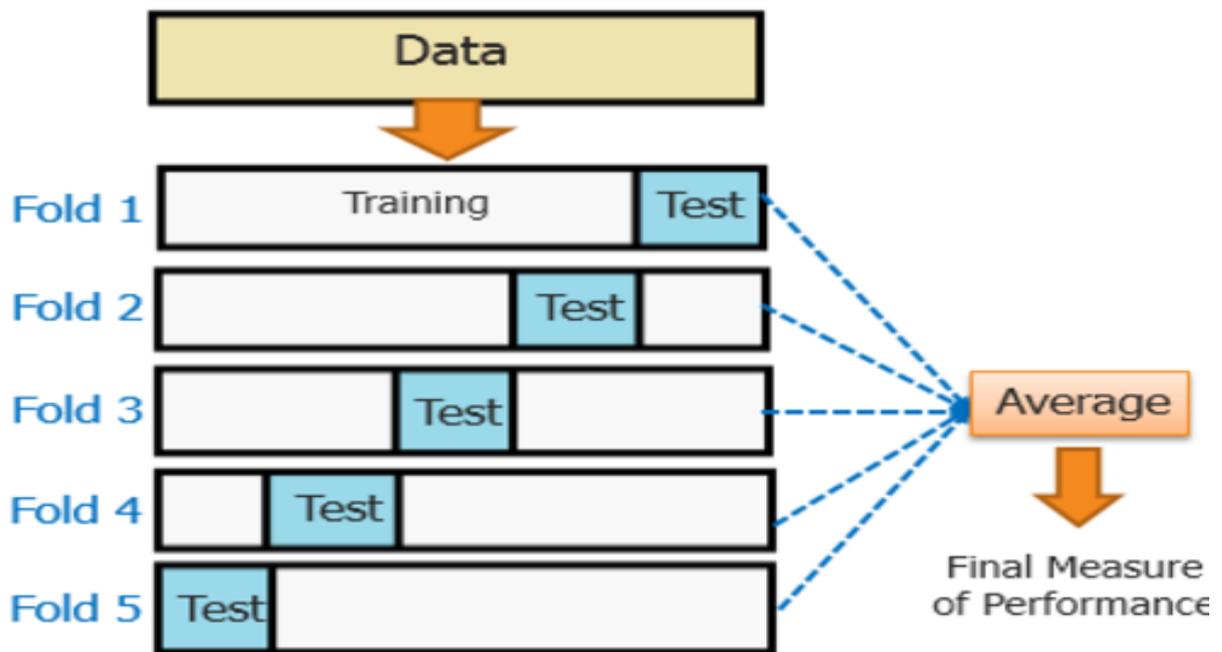
dengan $0 \leq \alpha_i \leq C, i = 1, 2, \dots, m$ (15)

untuk meminimalkan variabel kesalahan, pada rumusan diberikan penalti dengan menambahkan nilai *cost* (C). Nilai C berguna untuk mengontrol pertukaran antara margin dan *error* klasifikasi [12]. Sehingga data *training* dengan $0 < \alpha_i < C$ disebut *support vector* dan data yang memiliki nilai $\alpha_i = C$ disebut *bounded support vector*.

Grid Search Cross-validation

Grid search merupakan metode penentuan kombinasi model dan *hyperparameter* dengan melakukan uji coba satu persatu kombinasi model dan melakukan validasi pada setiap kombinasi [15]. *Hyperparameter* adalah sebuah variabel yang dapat menentukan hasil dari sebuah model dalam data *mining*.

Cross-validation menjadi salah satu metode untuk validasi model. *Cross-validation* digunakan untuk melihat keefektifan dari model yang telah dibentuk. Data dibagi secara acak menjadi 2 bagian yaitu sebagai data *training* dan data *testing*. Setiap kelas pada data set harus diwakili dalam proporsi yang tepat antara data *training* dan *testing*. Pada masing-masing kelas dengan perbandingan yang sama. Data *training* berguna untuk melatih model, kemudian data *testing* digunakan untuk validasi model yang telah dihasilkan. *K-fold cross-validation* merupakan salah satu metode dari *cross-validation* yang umum digunakan, model yang dihasilkan *K-fold cross-validation* bersifat tidak bias [16].



Gambar 3. *K-fold Cross-validation*

Dalam *k-fold cross-validation*, data dibagi menjadi k bagian sama besar. Proses klasifikasi dilakukan sebanyak k kali. Secara berurutan $k-1$ bagian dilatih menggunakan metode klasifikasi dan 1 bagian diuji menggunakan model *training*. Prosedur validasi silang dapat mencegah masalah *overfitting*. Nilai k seringkali berupa bilangan kecil seperti 5 atau 10 [2]. *Grid search cross-validation* merupakan proses uji coba satu persatu kombinasi model yang dihasilkan dari beberapa nilai parameter dengan proses validasi silang.

Confusion Matrix

Suatu proses prediksi berupaya untuk sukses dalam mengklasifikasikan semua dataset sesuai dengan kelas yang sebenarnya, akan tetapi pada praktiknya kinerja klasifikasi tidak mampu untuk dikatakan 100% sukses. Oleh karena itu, dalam proses klasifikasi diperlukan evaluasi. Akurasi merupakan salah satu parameter yang dapat digunakan untuk mengevaluasi proses klasifikasi. Nilai akurasi didapatkan dari *confusion matrix* [17]. *Confusion matrix* merupakan sebuah tabel yang menyajikan informasi dari kinerja pengklasifikasi [2].

Tabel 1. Confusion Matrix

<i>Actual</i>	<i>Predicted</i>	
	<i>Positive</i>	<i>Negative</i>
<i>Positive</i>	TP	FN
<i>Negative</i>	FP	TN

Berikut ini merupakan keterangan untuk Tabel 1: (1) *True positive* (TP), yaitu jumlah data yang berasal dari kelas *positive* dan diklasifikasikan sebagai kelas *positive*. (2) *True negative* (TN), yaitu jumlah data yang berasal dari kelas *negative* dan diklasifikasikan sebagai kelas *negative*. (3) *False positive* (FP), yaitu jumlah data yang berasal dari kelas *negative* dan diklasifikasikan sebagai kelas *positive*. (4) *False negative* (FN), yaitu jumlah data yang berasal dari kelas *positive* dan diklasifikasikan sebagai kelas *negative*.

Dari tabel *confusion matrix*, kinerja klasifikasi dapat dihitung sebagai berikut:

$$Accuracy (\%) = \frac{TP+TN}{TP+FP+FN+TN} \tag{16}$$

Dataset Penelitian

Pada penelitian ini, data yang digunakan merupakan data sekunder yang berasal dari repositori *Universitas California Irvine (UCI) Machine Learning*. Data *Diabetic Retinopathy Debrecen* berisi 1151 sampel dengan 19 fitur dan 2 kelas. Sejumlah 540 pasien *control* dan 611 pasien terdiagnosa diabetes retinopati. Kumpulan fitur tersebut didapatkan dari hasil ekstraksi gambar mesidor untuk kemudian digunakan sebagai indikasi dari setiap pasien apakah terdiagnosa penyakit diabetes atau tidak.

C. Hasil dan Pembahasan

Hasil analisis klasifikasi data Diabetes Retinopati dengan menggunakan metode *grid search cross-validation* untuk melihat parameter optimal dari metode SVM dilihat dari nilai akurasi yang dilakukan melalui bantuan *software R Studio*. Hasil pengujian terlampir pada tabel 2.

Tabel 2. Accuracy Klasifikasi Kernel Linear

Fungsi Kernel	Parameter Hyperplane	Accuracy					Rata-rata
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
<i>Linear</i>	C = 0,01	0,619	0,530	0,530	0,530	0,530	0,5481
	C = 0,1	0,675	0,687	0,583	0,652	0,613	0,6420
	C = 1	0,745	0,687	0,678	0,748	0,713	0,7141
	C = 10	0,758	0,748	0,700	0,787	0,748	0,7480 *
	C = 100	0,753	0,730	0,713	0,787	0,704	0,7376
	C = 1000	0,749	0,704	0,687	0,739	0,696	0,7150

Sumber: Data Penelitian yang Sudah Diolah, 2023.

Dari tabel di atas, akurasi tertinggi dihasilkan oleh model dengan parameter *cost* 10 sebesar 0,748. Pada *Fold* pertama dan kedua, akurasi tertinggi dihasilkan oleh model dengan parameter *cost* sebesar 10. Pada *Fold* ketiga, akurasi tertinggi dihasilkan oleh model dengan parameter *cost* sebesar 100. Pada *Fold* keempat, akurasi

tertinggi dihasilkan oleh model dengan parameter *cost* 10 dan *cost* 100. Pada *Fold* kelima, akurasi tertinggi dihasilkan oleh model dengan parameter *cost* 10.

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan hasil dari penelitian sebagai berikut: Pada penelitian ini, didapatkan beberapa model klasifikasi. Dengan algoritma *grid search* yang dipadukan dengan *5-fold cross-validation*, serta kernel yang digunakan ialah kernel *linear* dan menggunakan 6 buah parameter *cost* yaitu 10^{-2} , 10^{-1} , 10^0 , 10^1 , 10^2 , dan 10^3 menghasilkan 30 model. Model terbaik untuk data *diabetic retinopathy* ini, didapatkan menggunakan parameter *cost* 10 dengan *accuracy* sebesar 0,748 atau 74,8%.

Daftar Pustaka

- [1] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining concepts and techniques*, 3rd ed. San Francisco: Morgan Kaufmann Publishers, 2012.
- [2] M. Bramer, *Principles of Data Mining*. Springer, 2007.
- [3] J. Han and M. Kamber, *Data mining: Data mining concepts and techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.
- [4] Anita, A. Wicaksono, and T. N. Padilah, "Pengaruh Jumlah Record Dataset Terhadap Algoritma Klasifikasi Berdasarkan Data Customer Churn," *Jurnal Ilmiah Informatika*, vol. 6, no. 1, pp. 1–10, Jun. 2021, doi: 10.35316/jimi.v6i1.1223.
- [5] D. Kerami and H. Murfi, "KAJIAN KEMAMPUAN GENERALISASI SUPPORT VECTOR MACHINE DALAM PENGENALAN JENIS SPLICE SITES PADA BARISAN DNA," *MAKARA of Science Series*, vol. 8, no. 3, pp. 89–95, Oct. 2010, doi: 10.7454/mss.v8i3.451.
- [6] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 2, pp. 802–810, 2018, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [7] Irawadi and S. Sunendiari, "Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara," *Jurnal Riset Statistika*, vol. 1, no. 1, pp. 19–27, Jul. 2021, doi: 10.29313/jrs.v1i1.22.
- [8] A. Zakrani, A. Najm, and A. Marzak, "Support Vector Regression Based on Grid-Search Method for Agile Software Effort Prediction," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/CIST.2018.8596370.
- [9] S. Frangky Handono, F. Tri Anggraeny, and B. Rahmat, "IMPLEMENTASI CONVOLUTIONAL NEURAL NETWORK (CNN) UNTUK DETEKSI RETINOPATI DIABETIK," *Jurnal Informatika dan Sistem Informasi (JIFoSI)*, vol. 1, no. 1, pp. 669–678, 2020.
- [10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [11] T. V Rampisela and Z. Rustam, "Classification of Schizophrenia Data Using Support Vector Machine (SVM)," *J Phys Conf Ser*, vol. 1108, Nov. 2018, doi: 10.1088/1742-6596/1108/1/012044.
- [12] M. S. Sari, D. Safitri, and S. Sugito, "KLASIFIKASI WILAYAH DESA-PERDESAAN DAN DESA-PERKOTAAN WILAYAH KABUPATEN SEMARANG DENGAN SUPPORT VECTOR MACHINE (SVM)," *Jurnal Gaussian*, vol. 3, no. 4, pp. 751–760, Oct. 2014.
- [13] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.
- [14] S. R. Gunn, "Support Vector Machines for Classification and Regression," 1998.

- [15] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 4, no. 1, pp. 12–21, May 2022, doi: 10.12928/biste.v4i1.6079.
- [16] Y. Widyaningsih, G. P. Arum, and K. Prawira, "APLIKASI K-FOLD CROSS VALIDATION DALAM PENENTUAN MODEL REGRESI BINOMIAL NEGATIF TERBAIK," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 2, pp. 315–322, Jun. 2021, doi: 10.30598/barekengvol15iss2pp315-322.
- [17] R. D. Septiana, A. B. Susanto, and T. Tukiyyat, "Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic dan Particle Swarm Optimization," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 5, no. 1, pp. 49–56, Sep. 2021, doi: 10.47970/siskom-kb.v5i1.228.