OPTIMISASI PARAMETER SUPPORT VECTOR MACHINE BERBASIS ALGORITMA GENETIKA PADA KLASIFIKASI TEKS PENGADUAN MASYARAKAT

Istiadi[1*)], Aviv Yuniar Rahman[1)]
[1)] Program Studi Teknik Informatika, Universitas Widyagama Malang, Malang
*Email Korespondensi : istiadi@widyagama.ac.id

ABSTRAK

Layanan pesan masyarakat secara online berupa teks yang perlu mendapatkan tanggapan sesegera mungkin oleh pihak terkait. Oleh karena itu dibutuhkan sistem klasifikasi teks dengan akurasi yang baik. Salah satunya adalah dengan metode *Support Vector Machine* (SVM) yang memiliki variasi *kernel*. Akan tetapi kelemahan SVM adalah untuk penentuan parameternya. Salah satu pendekatan dalam optimisasi penentuan parameter adalah menggunakan Algoritma Genetika (AG). Penelitian ini bertujuan melakukan optimisasi parameter pada SVM menggunakan AG pada kasus layanan pesan masyarakat pada Sambat Online Kota Malang. Jumlah sampel data diambil pada layanan tersebut sebanyak 200 pesan dengan tujuh kelas kategori yang dibagi menjadi data latih dan data uji yang bervariasi perbandingannya. Paremater yang dioptimasi pada masing-masing kernel antara lain kernel *linear* (C), *polynomial* (C, γ, derajat), *RBF* (C, γ), dan *sigmoid* (C, γ). Optimisasi dan pengujian dilakukan dengan variasi persentase rasio data latih dan data uji yaitu 20:80, 40:60, 60:40, 80:20. Pengujian dilakukan dengan menerapkan parameter hasil optimisasi dan mengukur tingkat akurasinya. Hasil pengujian terbaik diperoleh pada *kernel linear* dengan tingkat akurasi sebesar 85,37% pada rasio data latih terhadap data uji sebesar 80% : 20%. Hal ini menunjukkan bahwa *kernel linear* lebih sesuai dibandingkan *kernel* lainnya apabila digunakan pada kasus klasifikasi teks pada layanan Sambat Online.

Kata kunci: SVM; Algoritma Genetika; parameter; kernel; klasifikasi teks.

ABSTRACT

Online community message services in the form of text need to get a response as soon as possible by the parties concerned. Therefore this requires a text classification system with good accuracy. One of them is the Support Vector Machine (SVM) method which has kernel variations. However, the weakness of SVM is its parameter determination. One approach in optimizing parameter determination is using a Genetic Algorithm (AG). This study aims to optimize the parameters of SVM using AG in the case of public messaging services at Sambat Online in Malang City. The number of data samples taken at this service was 200 messages with seven class categories which were further divided into training data and test data with varying comparisons. Optimized parameters for each kernel include linear (C), polynomial (C, γ, gedree), RBF (C, γ), and sigmoid (C, γ) kernels. Optimization and testing were carried out by varying the percentage ratio of training data and test data, namely 20:80, 40:60, 60:40, 80:20. The test is carried out by applying the optimization result parameters and measuring the level of accuracy. The best test results were obtained in the Linear kernel with an accuracy rate of 85.37% in the ratio of training data to test data of 80%: 20%. This shows that the Linear kernel is more suitable than other kernels when used in the case of text classification in the Sambat Online service.

Keywords: SVM; Genetic Algorithm; parameter; kernel; text classification.

ISSN Cetak : 2622-1276
ISSN Online : 2622-1284

PENDAHULUAN

Media pengaduan masyarakat pada pemerintahan telah banyak diimplementasikan sebagai layanan berbasis TIK sebagai salah satu bentuk e-government [1]. Seperti Pada layanan Sambat (Sistem Aplikasi Masyarakat Bertanya Terpadu) Online yang dikembangkan oleh Dinas Komunikasi dan Informatika Pemerintah Kota Malang. Teks pengaduan yang masuk dari website maupun dari pesan singkat akan dikategorikan sesuai dengan Organisasi Perangkat Daerah (OPD) yang bertanggung jawab. Namun karena beragamnya pengaduan berbasis teks dan klasifikasinya masih manual, maka pesan tersebut tidak bisa segera tersampaikan pada pihak terkait.

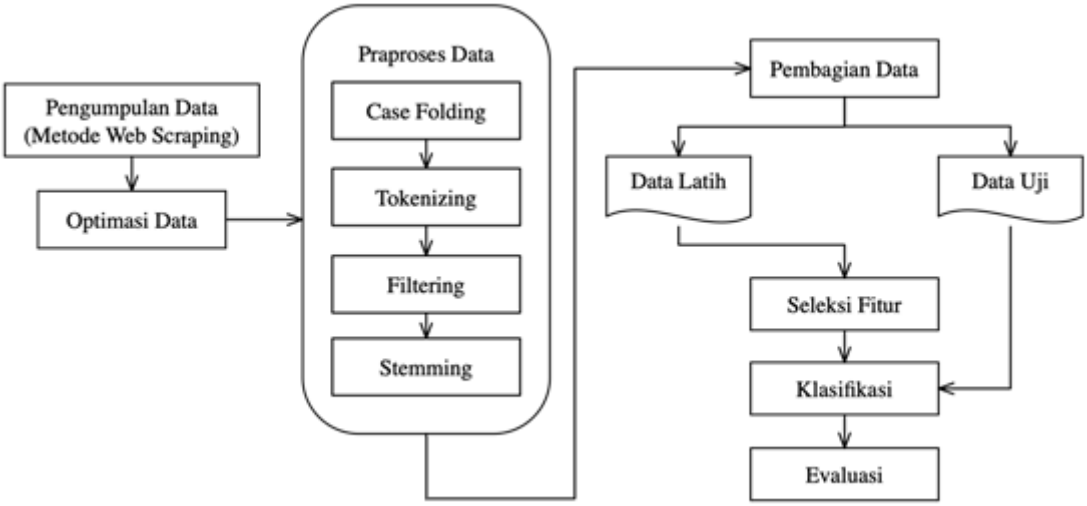
Studi pengembangan pengklasifikasi teks pada layanan Sambat Online telah dilakukan. Suharno dkk dalam [2], telah menerapkan *K-Nearest Neighbor* (KNN) dengan seleksi fitur menggunakan *Chi-Square* untuk klasifikasi dalam tiga kelas diperoleh *f-measure* terbaik dengan nilai 78% pada k = 15 dan k = 5 dengan seleksi fitur sebesar 25%. Prasanti dkk dalam [3], telah melakukan pengujian seleksi fitur *N-Gram* dan *Neighbor Weighted* pada KNN (NW-KNN) untuk klasifikasi dalam tiga kelas yang menunjukkan bahwa dengan nilai tetangga k = 3 dan metode N-Gram dengan *Unigram* memiliki nilai *f-measure* tertinggi sebesar 75.25%. Nugroho dkk dalam [4], melakukan optimasi *Naïve Bayes Classifier* (NBC) dengan menggunakan *Particle Swam Optimization* (PSO) untuk klasifikasi dalam tujuh kelas menghasilkan nilai akurasi sebesar 87,44% yang lebih baik dari k-NN sebesar 75 % dan NBC sebesar 64,38 %.

Upaya perbaikan kinerja klasifikasi tersebut perlu dilakukan agar mampu memberikan hasil akurasi yang lebih baik sehingga layanan Sambat Online potensial untuk dimaksimalkan. Salah satu pendekatan yang dapat dilakukan adalah penggunaan metode *Support Vector Machine* (SVM) yang memiliki kemampuan yang baik untuk klasifikasi teks [5]. Menurut Jumeilah dalam [5], metode SVM mampu memberikan tingkat akurasi 90% pada kasus pengkategorisasian penelitian, selain itu metode SVM memiliki beberapa fungsi kernel yaitu fungsi *Linear*, *Polynomial*, *Radial Basis Function* (RBF), dan *Sigmoid* yang potensial sebagai alternatif dalam menemukan solusi yang terbaik [6]. Namun permasalahan yang muncul adalah ketika menentukan parameter yang optimal pada masing-masing *kernel* tersebut. Harafani dalam [7], telah mengusulkan penggunaan Algoritma Genetik (AG) untuk mengoptimasi parameter VSM pada kasus estimasi penyakit liver. Dengan demikian metode AG potensial digunakan pada VSM pada kasus klasifikasi teks pengaduan masyarakat pada layanan Sambat Online ini.

Penelitian ini bertujuan menerapkan metode SVM dengan optimisasi parameter *kernel* menggunakan AG untuk klasifikasi teks pengaduan masyarakat pada layanan Sambat online. Klasifikasi dilakukan menggunakan dataset dari penelitian [4] yang membagi kelas dalam tujuh kategori. Perbandingan kinerja kernel pada SVM diukur melalui tingkat akurasi sehingga dapat diketahui yang terbaik untuk kasus klasifikasi teks pengaduan masyarakat pada layanan Sambat online.

METODE PENELITIAN

Tahapan penelitian ini memuat empat proses yang mengacu pada Gambar 1.



Gambar 1. Tahapan penelitian

Data penelitian ini diambil dari portal Sambat Online dengan metode *scraping* mengacu pada [4]. Data yang dikoleksi sebanyak 200 data pengaduan yang terdiri dari tujuh label berdasarkan OPD yang bertanggung jawab. Selanjutnya data tersebut disajikan dalam format .xlxs (Microsoft Excel) agar mempermudah praproses selanjutnya. Sebaran data pengaduan berdasarkan kelas OPD disajikan pada Tabel 1.

Tabel 1. Sebaran data kelas berdasarkan OPD.

No	Kategori OPD	Jumlah Data
1	DISPENDUK	29
2	DLH	25
3	DPUPR	30
4	DISPENDIK	30
5	DISHUB	30
6	DPKP	28
7	SATPOL PP	28
	Total Data	200

Praproses data dilakukan agar data siap untuk diolah. Pada umumnya informasi awal yang akan digali berupa format yang tidak terstruktur sehingga diperlukan proses pengubahan menjadi data yang terstruktur dengan tahapan antara lain *Case Folding* yaitu menyamakan semua huruf menjadi huruf kecil) *Tokenizing* yang menyusun daftar kata-kata dari dokumen, *Filtering* yang menyeleksi data token menjadi kata-kata penting, dan *Stemming* yaitu proses untuk mendapatkan kata dasar.

Selanjutnya data dipisahkan atas data latih untuk pembelajaran dan data uji untuk pengujian. Data latih digunakan untuk proses pembelajaran SVM dengan optimasi menggunakan AG. Hasil optimasi parameter SVM digunakan untuk menguji kinerja pengklasifikasi untuk masing-masning *kernel* menggunakan data uji.

ISSN Cetak : 2622-1276
ISSN Online : 2622-1284

Metode klasifikasi menggunakan SVM dapat memanfaatkan *library* SVM (LibSVM) yang telah dikembangkan pada [8]. LibSVM menyediakan jenis pengklasifikasi yaitu *C-Support Vector Clasification* (C-SVC) yang mengacu pada penyelesaian permasalahan utama (1),

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l, \end{aligned} \tag{1}$$

di mana φ (x_i) memetakan x_i ke dalam ruang berdimensi lebih tinggi dan C> 0 adalah parameter regularisasi. Parameter C digunakan sebagai pengatur *trade off* terkait *margin* dengan *error* dalam pengelompokan ξ atau tingkat *error* pada klasifikasi.

Fungsi *kernel* digunakan untuk melindungi titik data ke ruang dimensi yang lebih tinggi untuk meningkatkan kemampuannya dalam menemukan hiperplane terbaik untuk memisahkan titik data dari kelas yang berbeda. *Kernel* adalah fungsi κ untuk semua $x, z \in X$ akan memenuhi syarat

$$\kappa(x, z) = \langle \phi(x). \phi(z) \rangle , \tag{2}$$

dimana ϕ adalah pemetaan hasil kali dalam dari X ke luar area dengan dimensi yang lebih tinggi F

$$\phi: x \mapsto \phi(x) \in F. \tag{3}$$

Beberapa fungsi *kernel* yang umum antara lain:

a. *Linear*

Fungsi kernel linier didefinisikan sebagai:

$$\kappa(\boldsymbol{xi}, \boldsymbol{xj}) = \boldsymbol{xi}^T\boldsymbol{xj} \tag{4}$$

Fungsi kernel linier adalah fungsi *kernel* paling sederhana yang merupakan perkalian titik dari dua vektor.

b. *Radial Basis Function* (RBF)

RBF juga bisa disebut sebagai fungsi *kernel* Gaussian. RBF didefinisikan sebagai:

$$\kappa(\boldsymbol{xi}, \boldsymbol{xj}) = \exp(-\gamma\|\boldsymbol{xi} - \boldsymbol{xj}\|^2), \gamma > 0 \tag{5}$$

dimana γ adalah parameter positif untuk mengatur jarak

c. *Polynomial*

Fungsi *kernel polynomial* dengan memiliki derajat d, di mana r dan d adalah parameter yang didefinisikan sebagai berikut:

$$\kappa(\boldsymbol{xi}, \boldsymbol{xj}) = (\gamma\, \boldsymbol{xi}^T\boldsymbol{xj} + r)^d, \gamma > 0 \tag{6}$$

d. *Sigmoid*

Fungsi kernel sigmoid didefinisikan sebagai:

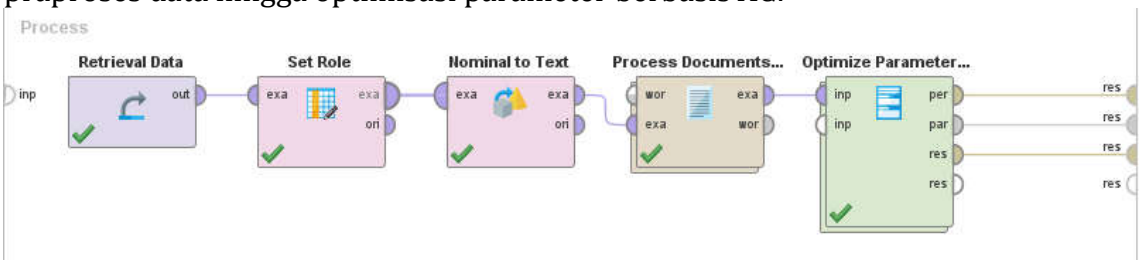
$$\kappa(\boldsymbol{xi}, \boldsymbol{xj}) = \tanh(\gamma\boldsymbol{xi}, \boldsymbol{xj} + \boldsymbol{r}) \tag{7}$$

di mana $\tanh(a) = 2\sigma(a) - 1$, dan $\sigma(a) = 1 / (1 + \exp(a))$

Klasifikasi SVM melibatkan parameter C dan parameter lainnya yang berkaitan dengan jenis kernel-nya. Dengan demikian perlu upaya untuk mendapatkan nilai dari parameter-parameter yang dapat memberikan nilai akurasi

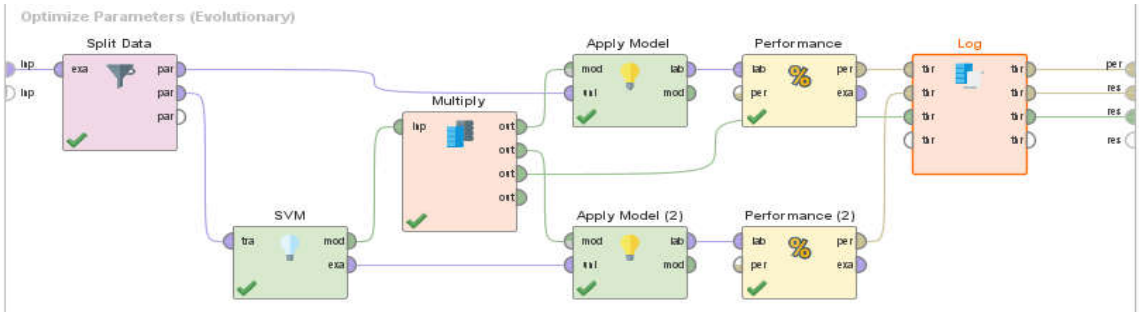
klasifikasi yang optimal. AG mengadopsi konsep evoluasi yang potensial untuk mendapatkan nilai parameter secara optimal [9]. AG berjalan dari generasi ke generasi dalam suatu populasi (kumpulan individu). Suatu individu merepresentasikan sekumpulan parameter yang diharapkan hasilnya untuk dipilih yang terbaik. Individu yang baik adalah yang mampu bertahan melalui rekombinasi atau pindah silang (*cross over*) dan mutasi. Suatu nilai kebugaran (*fitness*) dijadikan ukuran tingkat optimalitas parameter dalam indivitu.

Berdasarkan proses tersebut selanjutnya digunakan perangkat lunak bantu untuk proses komputasinyam yaitu Rapidminer. Perangkat lunak ini yang menyediakan sejumlah fitur untuk simulasi data mining. Gambar 2 menyajikan serangkaian proses dari pemuatan data awal (*raw data*), pengkonversian teks, praproses data hingga optimisasi parameter berbasis AG.



Gambar 2. Model pemrosesan optimisasi paramater.

Bagian optimisasi parameter pada Gambar 2 merupakan modul yang didalamnya terdapat obyek (VSM) yang dioptimasi parameternya. Gambar 3 menyajikan isi dari modul optimisasi parameter.



Gambar 3. Model pemrosesan klasifikasi SVM.

Berdasarkan Gambar 3 dapat diketahui model pemrosesan klasifikasi SVM yang meliputi proses pembelajaran dan uji klasifikasi. Data latih digunakan untuk optimasi dalam pembelajaran SVM dengan menentukan jenis *kernel*-nya. Setelah proses pembelajaran selesai, maka dilakukan uji klasifikasi menggunakan data uji dan dilihat kinerjanya (performance) berdasarkan tingkat akurasi klasifikasi.

HASIL DAN PEMBAHASAN

Proses optimasi parameter pada AG menggunakan jumlah populasi sebanyak 10 individu dan jumlah generasi sebanyak 30. Fungsi *fitness* menggunakan nilai akurasi berdasarkan data latih. Probabilitas pindah silang sebesar 0,9 menggunakan sistem *tournament* dengan nilai fraksi 0,25 sedangkan metode mutasi menggunakan metode Gausian. Rasio data latih dengan data uji dibuat bervariasi untuk pengetahui sejauhmana pengaruhnya. Tabel 2 menyajikan nilai-nilai parameter hasil optimasi.

ISSN Cetak : 2622-1276
ISSN Online : 2622-1284

Tabel 2. Hasil Optimasi parameter VSM

Jumlah data latih	Para-meter	Jenis *Kernel* VSM			
		Linear	*Polynomial*	*RBF*	*Sigmoid*
20%	γ	-	7.748751110169172	96.85688383830217	2.9431484104427947
	C	968.4928217146895	29.36698608959218	182.45502741694176	188.7604816786123
	Derajat	-	2	-	-
40%	γ	-	7.745928557501005	96.84294473371402	2.946641819172087
	C	182.45416895055632	29.351657000085854	182.44742397604898	188.80170157269546
	Derajat	-	2	-	-
60%	γ	-	7.748751110169172	2.936950427817042	2.9555756131963267
	C	623.3716322574757	29.36698608959218	188.74669941536587	564.6347531704695
	Derajat	-	2	-	-
80%	γ	-	7.731710751741236	2.955575613963267	2.914732768698254
	C	623.3806004991275	29.366658535266655	142.33152790398233	188.77177504141284
	Derajat	-	2	-	-

Sesuai dengan Tabel 2 dapat diketahui parameter-paramter yang dihasilkan berdasarkan *kernel* dan jumlah data latih. Parameter yang dioptimasi disesuikan dengan jenis kernelnya yaitu *linear* (C), *polynomial* (γ, C, Derajat), RBF (γ, C), dan *sigmoid* (γ, C). Nilai-nilai parameter masing-masing *kernel* dikelompokkan berdasarkan jumlah data latih yaitu 20%, 40%, 60%, dan 80%. Selanjutnya untuk mengetahui kinerja klasifikasi, nilai-nilai parameter tersebut diterapkan pada SVM dengan menggunakan data uji. Ukuran kinerja didasarkan nilai akurasi sehingga dapat diketahui perbandingan antar *kernel*-nya maupun variasi rasio data latih terhadap data ujinya. Tabel 3 menyajikan data kinerja masing-masing *kernel*.

Tabel 3. Data kinerja masing-masing *kernel*

Rasio Data		Nilai Akurasi pada *Kernel*			
Latih	Uji	*Linear*	*Polynomial*	*RBF*	*Sigmoid*
20%	80%	62.89%	56.60%	59.75%	68.55%
40%	60%	71.67%	55.00%	49.17%	71.67%
60%	40%	80.00%	60.00%	47.50%	75.00%
80%	20%	85.37%	65.85%	53.66%	80.49%

Sesuai dengan Tabel 3 dapat diketahui nilai akurasi dari masing-masing *kernel* dan variasinya terhadap rasio data latih dan data uji. Pada *kernel linear* dan *sigmoid* tampak ada peningkatan seiring dengan peningkatan data latih, namun pada *kernel polynomial* dan RBF tidak tampak perubahan yang sebanding dengan peningkatan jumlah data latih. Kinerja terbaik pada *kernel linear*, *polynomial* dan *sigmoid* terjadi pada rasio data latih terhadap

data uji sebesar 80% : 20% dengan nilai akurasi masing-masing 85.37%, 65.85%, 80.49%, sedangkan pada *kernel* RBF nilai akurasi terbaik dengan nilai 59.75% terjadi pada rasio data latih terhadap data uji sebesar 20% : 80%. Dari kinerja *kernel* secara keseluruhan nilai akurasi tertinggi sebesar 85,37% terdapat pada *kernel linear* pada rasio data latih terhadap data uji sebesar 80% : 20%. Hal ini menunjukkan bahwa *kernel linear* lebih sesuai untuk klasifikasi dengan data teks pengaduan masyarakat pada layanan Sambar Online.

KESIMPULAN

Berdasarkan hasil pengujian menunjukkan AG mampu menghasilkan nilai-nilai parameter untuk SVM berdasarkan masing-masing kernelnya. Dari pengujian menunjukkan variasi jumlah data latih terhadap data uji berpengaruh nilai akurasi pada masing-masing kernel. Hasli kinerja terbaik terjadi pada kernel linear dengan nilai akurasi sebesar 85,37% pada rasio data latih terhadap data uji sebesar 80% : 20%.

Penelitian ini telah melakukan studi optimisasi parameter SVM menggunakan AG untuk klasifikasi teks pada layanan Sambat Online. Upaya optimisasi selanjutnya dapat memanfaatkan metode yang lainnya misalnya menggunakan PSO. Selain itu optimisasi juga potensial pada sisi seleksi fitur setelah praproses dokumen seperti penggunaan metode N-Gram.

UCAPAN TERIMA KASIH

Penelitian ini didanai melalui hibah penelitian internal (Perintis) LPPM Universitas Widyagama Malang tahun 2020.

REFERENSI

[1] Nugroho, L.E., Egaravanda, S. and Achmad, K.A. (2018). *An architecture for facilitating two-way G2C relationships in public service delivery. International Journal on Advanced Science, Engineering and Information Technology*, 8(4), pp.1179-1184.

[2] Suharno, C. F., Fauzi, M. A., & Perdana, R. S. (2017). Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-Square. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[3] Prasanti, A.A., Fauzi, M.A. and Furqon, M.T. (2018). Klasifikasi teks pengaduan pada sambat online menggunakan metode n-gram dan neighbor weighted k-nearest neighbor (NW-KNN). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, p.964X.

[4] Nugroho, K.S., Istiadi, I. and Marisa, F. (2020). Optimalisasi Naive Bayes Classifier untuk Klasifikasi Teks pada E-Government Menggunakan Particle Swarm Optimization. Jurnal Teknologi dan Sistem Komputer, 8(1).

[5] Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 1(1), 19-25.

[6] IntanP. K. (2019). *Comparison of Kernel Function on Support Vector Machine in Classification of Childbirth*. Jurnal Matematika MANTIK, 5(2), 90-99.

[7] Harafani, H. (2020). *Support Vector Machine Parameter Optimization to Improve Liver Disease Estimation with Genetic Algorithm*. SinkrOn, 4(2), 106-114.

[8] Chang, C. C., & Lin, C. J. (2011). *LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST)*, 2(3), 1-27.

ISSN Cetak : 2622-1276
ISSN Online : 2622-1284

[9] Ispandi, I., & Wahono, R. S. (2015). Penerapan Algoritma Genetika Untuk Optimasi Parameter pada Support Vector Machine untuk Meningkatkan Prediksi Pemasaran Langsung. Journal of Intelligent Systems, 1(2), 115-119.