



# Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin

Fadhila Tangguh Admojo<sup>a,1</sup>, Yudha Islami Sulistya<sup>a,2</sup>

<sup>a</sup> Institut Teknologi dan Bisnis Palcomtech, Jl Jend. Basuki Rachmat, Palembang 30151, Indonesia

<sup>b</sup> Universitas Muslim Indonesia, Jl. Urip Sumoharjo Km. 05, Makassar 90231, Indonesia

<sup>1</sup> Fadhila.tangguh@gmail.com; <sup>2</sup> yudhasulistya.labfik@umi.ac.id

INFORMASI ARTIKEL	ABSTRAK
<p>Diajukan : 16 – 01 – 2022 Direvisi : 19 – 02 – 2022 Diterbitkan : 31 – 03 – 2022</p> <hr/> <p><b>Kata Kunci:</b> SGD Machine Learning Klasifikasi Analisis Performa Tahu Berformalin</p>	<p>Tahu berformalin adalah salah satu jenis makanan yang sering mengandung bahan-bahan kimia yang dapat mengawetkan daripada tahu tanpa formalin. Pada tahu berformalin dapat memberikan tekstur lebih kenyal dan berwarna putih bersih. Penelitian ini bertujuan untuk mengklasifikasikan tahu berformalin dan tahu tidak berformalin. Pada paper ini menggunakan algoritma Stochastic Gradient Descent atau dalam penerapannya lebih dikenal dengan SGD Classifier yang merupakan bagian dari algoritma machine learning untuk klasifikasi, regresi maupun jaringan syaraf tiruan serta algoritma ini sangat efisien pada dataset berskala besar. Penelitian ini mencoba menerapkan algoritma SGD pada dataset tahu berformalin dengan jumlah dataset yakni 11000 yang dimana 5500 data tahu berformalin dan 5500 data tahu tidak berformalin. Setelah dilakukan beberapa tahapan dalam pengujian dengan algoritma SGD maka diperoleh hasil akurasi, presisi, recall, f1-score pada model yang masing-masing 82.6% untuk akurasi, 81.7% untuk presisi, 84.1% untuk recall, 83.5% untuk f1-score dan dilakukan pengujian menggunakan 10 data yang tidak termasuk dalam data latih memperoleh performansi rata-rata akurasi sebesar 70%, presisi 71%, recall 70% dan f1-score 70%.</p> <div> </div>

## I. Pendahuluan

Tahu merupakan produk olahan kedelai yang tinggi protein, rendah karbohidrat, serta memiliki nilai gizi dan daya cerna yang sangat baik. Tahu Indonesia telah mengalami banyak perkembangan, dengan berbagai jenis tahu dan makanan berbahan dasar tahu. Tahu memiliki komposisi asam amino yang lengkap dan mengandung protein nabati dengan daya cerna yang tinggi yaitu 85% sampai 95%, sehingga layak dijadikan bahan pangan untuk perbaikan gizi (SNI0131421998). Tahu mengandung berbagai jenis nutrisi seperti protein, lemak, karbohidrat, kalori, mineral, fosfor, vitamin E, vitamin B12, kalium dan kalsium. Pada awalnya hanya ada satu jenis tahu yang disebut tahu putih, namun dengan perkembangan memasak, berbagai perkembangan seperti tahu kuning, tahu susu, tahu sutera, tahu air, dan tahu kulit telah dilakukan. Umur simpan tahu sangat terbatas. Dalam kondisi normal (suhu kamar), umur simpan rata-rata adalah 12 hari. Di luar batas tersebut, rasa tahu menjadi asam dan tidak layak untuk dikonsumsi. Formalin merupakan senyawa antibakteri serbaguna yang dapat membunuh bakteri, jamur dan virus. Selain itu, interaksi formalin dan protein dalam makanan membuat tekstur menjadi rapuh, dan bau yang ditimbulkan oleh formalin tidak hinggap lalat. Perendaman tahu dalam larutan formalin selama 3 menit diketahui dapat memperpanjang umur simpan selama 45 hari pada suhu kamar [1]. Dalam mengklasifikasi tahu berformalin dan tahu yang tidak mengandung formalin dapat menggunakan machine learning dalam penerapannya.

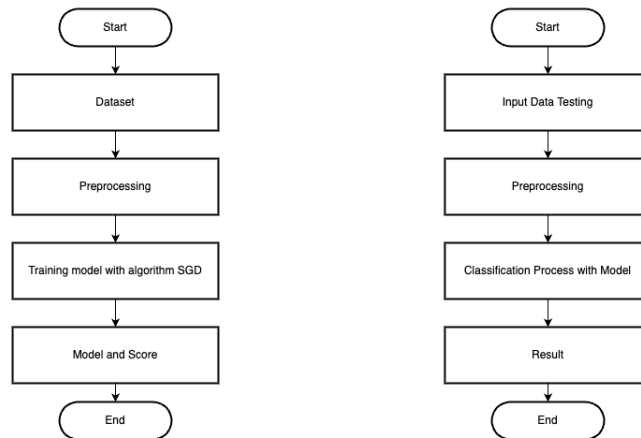
*Machine Learning* adalah bidang ilmu komputer yang memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit [2]. *Machine Learning* (ML) adalah salah satu aplikasi dari *Artificial Intelligent* (AI) yang fokus kepada pengembangan sebuah sistem yang mampu belajar sendiri tanpa harus diprogram berulang kali. ML membutuhkan sebuah data (*data training*) sebagai proses learning sebelum menghasilkan sebuah hasil. Jadi, secara sederhana dapat dijelaskan bahwa ML adalah pemrograman komputer untuk mencapai kriteria/performa tertentu dengan menggunakan sekumpulan data training atau pengalaman di masa lalu (*past experience*) [2]. Beberapa penelitian yang telah dilakukan, menyimpulkan ML dapat digunakan pada bidang medis untuk memprediksi penyakit[3][4][5][6][7].

Salah satu algoritma yang terkenal dalam machine learning atau lebih tepatnya supervised machine learning adalah Stochastic Gradient Descent (SGD). SGD merupakan salah satu variasi dari optimasi gradient decent yang selalu melakukan pembaruan parameter untuk setiap data yang sedang dilatih. Saat melakukan pembaruan parameter, SGDM tidak melakukan perulangan sehingga kinerjanya lebih cepat untuk dataset berjumlah besar. Nilai learning rate standar pada SGD adalah 0,01[8].

Berdasarkan latar belakang diatas, penelitian yang dilakukan adalah analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin. Penelitian ini berfokus bagaimana mengklasifikasikan tahu berformalin dan tahu tidak mengandung formalin. Pada penelitian ini dibagi menjadi lima tahapan. Pertama adalah persiapan dataset tahu berformalin. Kedua adalah preprocessing data tahu berformalin. Ketiga adalah melakukan training model menggunakan algoritma SGD. Keempat adalah menghasilkan skor pada model. Kelima adalah melakukan klasifikasi secara otomatis dengan menggunakan data uji yang diduga tahu berformalin dan tahu tidak mengandung formalin yang dialokasikan menggunakan algoritma SGD.

## II. Metode

Penelitian ini menggunakan *Machine Learning* dengan algoritma Stochastic Gradient Descent untuk mengklasifikasikan tahu berformalin



Gambar 1. (a)Flowchart of make model and score (b) Flowchart of classification process

Berdasarkan **Gambar 1.(a)** tahap pertama adalah menyiapkan dataset yang digunakan sebagai data latih. Tahap selanjutnya adalah preprocessing. Setelah proses dari preprocessing dilakukan, selanjutnya adalah melatih model dengan algoritma SGD, setelah itu bagian terakhir adalah dihasilkannya model yang akan digunakan pada saat pengujian dan skor yang bertujuan untuk seberapa bagus model yang telah dilatih. Serta pada **Gambar 1.(b)**, tahap pertama adalah menginputkan data testing yang sudah disiapkan diluar dari dataset pembuatan model. Tahap selanjutnya adalah preprocessing. Setelah proses dari preprocessing dilakukan, selanjutnya adalah mengklasifikasikan tahu berformalin menggunakan model yang telah dibuat sebelumnya. Setelah itu akan dihasilkan skor dari klasifikasi.

### A. Dataset

Pada tahap ini data yang digunakan berasal dari dataset tahu berformalin berjumlah 11000 data yang masing-masing 5500 berlabel tahu berformalin dan 5500 lainnya berlabel tahu tidak berformalin disertai dengan 8 feature antara lain H2 MQ2(ppm), LPG MQ2(ppm), CO MQ2(ppm), Alcohol MQ2(ppm), Propane MQ2(ppm), CH4 MQ4(ppm), Smoke MQ4(ppm), Temperature(C). Langkah selanjutnya adalah melakukan preprocessing data untuk di proses pada tahapan selanjutnya

### B. Preprocessing

Pada tahap ini dilakukan preprocessing data dengan menggunakan teknik minmaxscaler. Teknik MinMaxScaler[1] merupakan teknik yang digunakan untuk mengubah skala nilai menjadi lebih kecil tanpa melakukan modifikasi isi informasi didalamnya dengan melakukan operasi perubahan skala batas tepi atas 1 dan batas tepi bawah 0 (0,1)[2]. Teknik ini digunakan sebagai solusi dalam penyederhanaan feature yang terlalu jauh serta normalisasi dengan melakukan transformasi linier pada data asli. Sehingga dapat menghasilkan keseimbangan nilai perbandingan antar data[3]. Perhitungan MinMaxScaler dapat menggunakan persamaan sebagai berikut

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Keterangan:

Xsc = Nilai Normalisasi

X = Nilai Value dalam dataset

Xmin = Nilai minimal dalam suatu feature

Xmax = Nilai maksimal dalam suatu feature

### C. Stochastic Gradient Descent Classifier

SGD merupakan sebuah pendekatan sederhana dan efisien dalam melakukan klasifikasi secara linier menggunakan pembelajaran diskriminatif. Metode ini berupa algoritma optimasi iteratif (ulang) yang berguna untuk mencari titik fungsi minimum yang dapat diturunkan. Pada awal proses algoritma dimulai dengan melakukan penebakan. Kesalahan penebakan diperbaiki selama terjadi pengulangan tebakan menggunakan aturan gradien (turunan) dari fungsi yang akan diminimalkan. SGD memiliki kemampuan belajar lebih cepat dalam melakukan pelatihan klasifikasi. Selain itu, berdasarkan ukuran dataset latih tidak terbatas waktu pelaksanaannya[4]

Proses algoritma SGD adalah dengan menemukan nilai  $\theta$  yang dapat meminimalkan fungsi  $J(\theta)$ . Untuk menentukan nilai awal  $\theta$  digunakan algoritma pencarian, kemudian pada setiap iterasi nilai  $\theta$  agar terus diperbaharui sampai menemukan titik minimum atau nilai  $J$  yang paling minimum. Proses pembaharuan nilai  $\theta$  pada setiap iterasi menggunakan Persamaan (2). Pembaharuan dilakukan secara bersamaan untuk semua nilai  $j = 0, \dots, n$ . Variable  $\alpha$  merupakan learning rate yang mengatur seberapa besar pembaharuan nilai. Persamaan nilai  $J(\theta)$  dapat dilihat pada Persamaan (3), di mana  $L$  merupakan loss function yang digunakan pada data pelatihan  $(x_1, y_1), \dots, (x_n, y_n)$ , dan  $R$  merupakan regularisasi atau penalty terhadap kompleksitas model[5].

$$\theta_j = \theta_j - \alpha \frac{\partial J}{\partial \theta_j}(\theta) \quad (2)$$

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \alpha R(W) \quad (3)$$

### D. Model & Score

Pada tahap ini terbagi menjadi dua bagian dapat menentukan skor pada model yang dibuat. Pertama adalah dengan menggunakan cross validation untuk menentukan seberapa bagus model yang dibuat. Kedua adalah menguji model dengan cara mengevaluasi model dengan melakukan perhitungan akurasi, presisi, recall dan f1-score terhadap model yang telah dibuat. Dengan membandingkan hasil prediksi benar-positif dengan hasil positif palsu dalam kelas klasifikasi. Untuk proses evaluasi dengan confusion matrix maka akan diperoleh nilai precision, recall, dan accuracy yang didapat dari rumus:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$precision = \frac{TP}{FP + TP} \quad (5)$$

$$recall = \frac{TP}{FN + TP} \quad (6)$$

$$f1 - score = 2 \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

Keterangan:

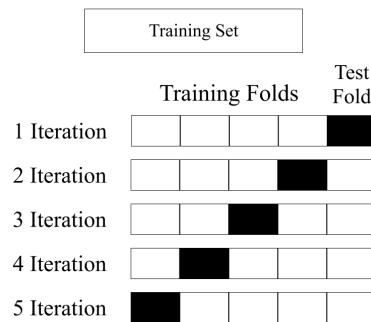
TP = jumlah tahu berformalin yang diklasifikasikan benar bahwa sebagai positif (tahu berformalin)

FP = jumlah tahu tidak berformalin yang diklasifikasikan benar bahwa sebagai positif (tahu berformalin)

TN = jumlah tahu tidak berformalin yang diklasifikasikan salah bahwa sebagai negatif (tahu tidak berformalin)

FN = jumlah tahu berformalin yang diklasifikasikan salah bahwa sebagai negatif (tahu tidak berformalin)

Sedangkan pada tahap ini juga menerapkan cross validation yang merupakan sebuah procedure acak yang membagi nilai atau kumpulan data menjadi  $K$  yang terputus-putus dengan ukuran yang kira-kira sama, dan setiap lipatan digunakan secara bergantian untuk menguji model yang diinduksi dari lipatan  $K-1$  lainnya oleh algoritma klasifikasi[6]. Pada **Gambar 2** merupakan gambaran proses terjadinya cross validation



Gambar 2. Cross validation process

### III. Hasil dan Pembahasan

Bagian ini membahas tentang pembuatan model, klasifikasi dan hasil pengujian dan model klasifikasi yang telah dibangun. Pada penelitian ini menggunakan dataset tahu berformalin sebanyak 12000 yang terbagi menjadi 6000 data untuk tahu berformalin dan 6000 lainnya yang tahu tidak berformalin. Contoh data ditunjukkan pada [Tabel 1](#).

Tabel 1. Example of Dataset Tofu Formalin (6 out of 12000)

No	1	2	3	4	5	6	7	8	9
0	1.6	0.8	3.5	1.3	1.0	1.8	0.2	1	0
1	1.6	0.8	3.5	1.3	1.0	1.8	0.2	1	0
2	1.6	0.8	3.5	1.3	1.0	1.8	0.2	1	0
...	...	...	...	...	...	...	...	...	...
10997	1.2	0.6	2.4	0.9	0.8	1.9	0.2	1	1
10998	1.6	0.8	3.5	1.2	1.0	2.3	0.4	1	1
10999	1.6	0.8	3.5	1.2	1.0	2.3	0.4	1	1

Features : 1 adalah H2\_MQ2(ppm)  
 2 adalah LPG\_MQ2(ppm)  
 3 adalah CO\_MQ2(ppm)  
 4 adalah Alcohol\_MQ2(ppm)  
 5 adalah Propane\_MQ2(ppm)  
 6 adalah CH4\_MQ4(ppm)  
 7 adalah Smoke\_MQ4(ppm)  
 8 adalah Temperature(C)  
 Outcome : 9 adalah label

#### A. Preprocessing

Pada tahap ini melakukan normalisasi pada features menggunakan teknik MinMaxScaler dengan menyederhanakan data pada features menjadi antara 0 hingga 1 seperti terlihat pada [Tabel 2](#) dibawah ini.

Tabel 2. Dataset After Normalisasi With MinMaxScaler

No	1	2	3	4	5	6	7	8	9
0	0.714286	0.75	0.608696	0.714286	0.6	0.0	0.0	0.0	0
1	0.714286	0.75	0.608696	0.714286	0.6	0.0	0.0	0.0	0
2	0.714286	0.75	0.608696	0.714286	0.6	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...
10997	0.142857	0.25	0.130435	0.142857	0.2	0.2	0.0	0.0	1
10998	0.714286	0.75	0.608696	0.571429	0.6	1.0	1.0	0.0	1
10999	0.714286	0.75	0.608696	0.571429	0.6	1.0	1.0	0.0	1

### B. Training Model With Algoritma SGD

Setelah melakukan preprocessing, selanjutnya membuat model dengan algoritma SGD. Sebelum itu dianjurkan untuk membuat sebuah variabel X (features) dan variabel y (outcome/label) seperti **Tabel 3** dibawah ini.

Tabel 3. Defintion X and y in Dataset

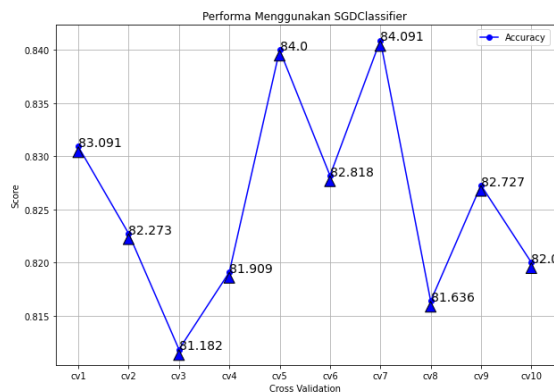
Features	1.	H2_MQ2(ppm)
	2.	LPG_MQ2(ppm)
	3.	CO_MQ2(ppm)
	4.	Alcohol_MQ2(ppm)
	5.	Propane_MQ2(ppm)
	6.	CH4_MQ4(ppm)
	7.	Smoke_MQ4(ppm)
	8.	Temperature(C)
Outcome/Label	Label	

Setelah menentukan variabel X dan y, selanjutnya menerapkan cross validation dengan nilai K yakni 10 atau 10 kali iterasi pada dataset berjumlah 11000 untuk menghasilkan skor pada model dengan perhitungan accuracy, precision, recall, f1-score seperti pada **Gambar 3** dibawah ini

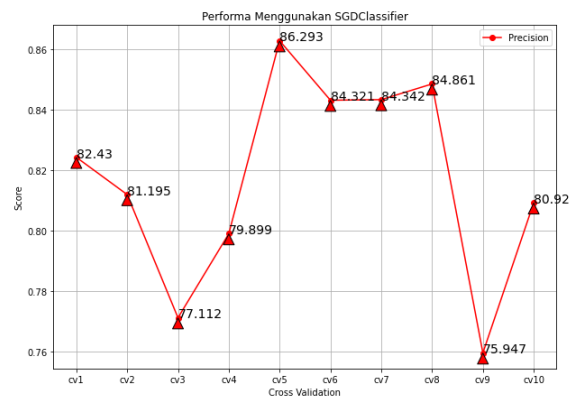
	Training Folds										Test Fold
1 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
2 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
3 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
4 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
5 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
6 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
7 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
8 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
9 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100
10 Iteration	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100	1100

Gambar 3. Process Training Model With Cross Validation

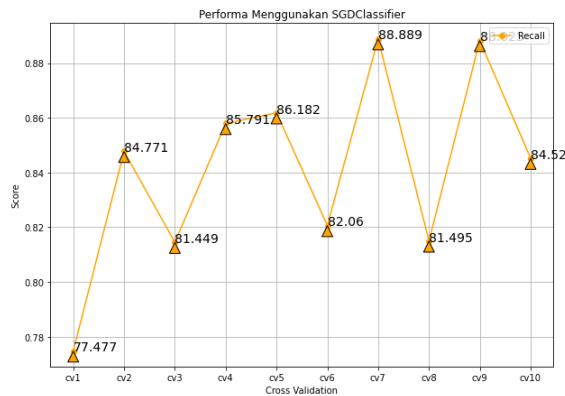
Setelah melakukan training model dengan menerapkan cross validation dengan nilai K yakni 10, maka diperoleh skor pada model dengan perhitungan accuracy, precision, recall, f1-score seperti pada **Gambar 4**, **Gambar 5**, **Gambar 6** dan **Gambar 7** dibawah ini



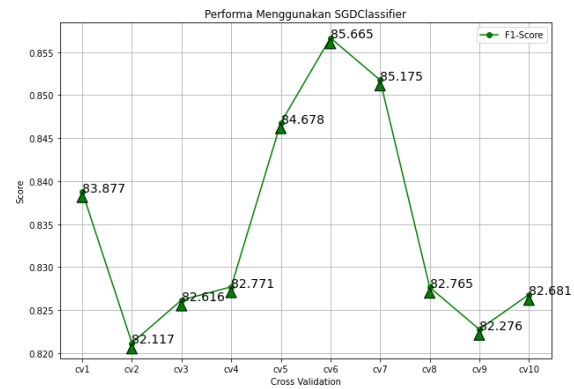
Gambar 4. accuracy score for the tofu model with formalin



Gambar 5. precision score for the tofu model with formalin



Gambar 6. recall score for the tofu model with formalin



Gambar 7. f1-score score for the tofu model with formalin

Perolehan hasil diatas dapat dipastikan kembali dengan mencari nilai rata-rata dari masing-masing skor pada performa diatas yang dapat dilihat pada **Tabel 4** dibawah ini.

Tabel 4. Rata-rata performa

Iteration	1	2	3	4	5	6	7	8	9	10	Result
Accuracy	0.831	0.823	0.812	0.819	0.84	0.828	0.841	0.816	0.827	0.82	0.826
Precision	0.824	0.812	0.771	0.799	0.863	0.843	0.843	0.849	0.759	0.809	0.817
Recall	0.775	0.848	0.814	0.858	0.862	0.821	0.889	0.815	0.888	0.845	0.841
F1-Score	0.839	0.821	0.826	0.828	0.847	0.857	0.852	0.828	0.823	0.827	0.835

Pada **Tabel 4** diperoleh hasil skor atau performa dari dataset tahu berformalin yang berjumlah 11000 dimana hasil akurasi 0.826, presisi 0.817, recall 0.841 dan f1-score 0.835.

### C. Testing

Pengujian klasifikasi dilakukan dengan mengukur nilai akurasi, presisi, recall dan f1-score dari model yang telah dibuat sebelumnya berdasarkan algoritma SGD, serta menggunakan preprocessing untuk normalisasi data dengan teknik MinMaxScaler. Data yang digunakan dalam pengujian model klasifikasi berjumlah 10 data yang sama sekali tidak termasuk dalam 11000 data latih. Data tersebut memiliki 5 data diantaranya ialah berlabel tahu berformalin dan 5 data diantaranya ialah tahu tidak berformalin. Pengujian performansi ini menggunakan metode confusion matrix.

	precision	recall	f1-score	support
0	0.67	0.80	0.73	5
1	0.75	0.60	0.67	5
accuracy			0.70	10
macro avg	0.71	0.70	0.70	10
weighted avg	0.71	0.70	0.70	10

Gambar 8. classification report for data testing



Gambar 9. confusion matrix for data testing

**Gambar 8** dan **Gambar 9** menampilkan classification report dan confusion matrix berdasarkan 10 data testing dengan 5 data tahu berformalin dan 5 data tahu tidak berformalin yang menghasilkan nilai rata-rata nilai performa akurasi sebesar 0.70, presisi 0.71, recall 0.70 dan f1-score 0.70

## IV. Kesimpulan

Berdasarkan hasil dan pembahasan disimpulkan bahwa performansi dari dataset yang berjumlah 11000 dengan 8 features dimana hasil akurasi 82.6%, presisi 81.7, recall 84.1 dan f1-score 83.5 dengan menerapkan

algoritma SGD. Pada pengujian menggunakan 10 data yang tidak termasuk dalam data latih memperoleh performansi rata-rata akurasi sebesar 70%, presisi 71%, recall 70% dan f1-score 70%.

#### Daftar Pustaka

- [1] S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, 2019, doi: 10.1016/j.cmpb.2019.05.005.
- [2] S. Zahara and S. Sugianto, "Prediksi Indeks Harga Konsumen Komoditas Makanan Berbasis Cloud Computing Menggunakan Multilayer Perceptron," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 6, no. 1, p. 21, 2021, doi: 10.31328/jointecs.v6i1.1702.
- [3] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
- [4] Y. Nataliani, S. M. Tambunan, and E. S. Lestari, "Perbandingan Klasifikasi dengan Pendekatan Pembelajaran Mesin untuk Mengidentifikasi Tweet Hoaks di Media Sosial Twitter," *JEPIN (Jurnal Edukasi dan ...)*, vol. 7, no. 2, pp. 112–120, 2021.
- [5] R. Dwiyanaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta, "Deteksi Sms Spam Berbahasa Indonesia Menggunakan Tf-Idf Dan Stochastic Gradient Descent Classifier ( Indonesian Sms Spam Detection Using Tf-Idf And Stochastic Gradient Descent)," *J. Teknol. Informasi, Komput. dan Apl.*, vol. 3, no. 2, pp. 200–207, 2021.
- [6] N. G. Ramadhan and A. Khoirunnisa, "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine," vol. 5, pp. 1580–1584, 2021, doi: 10.30865/mib.v5i4.3347.