KLASIFIKASI DATA DIAGNOSIS COVID-19 MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* (SVM) DAN *GENERALIZED LINEAR MODEL* (GLM)

(*Classification of Covid-19 Diagnosis Data Using Support Vector Machine (SVM) and Generalized Linear Model (GLM) Methods*)

Yeni Rismawati[1)], I Made Tirta[2)], Yuliani Setia Dewi[3)]

[1,2,3)]Jurusan Matematika, Fakultas MIPA, Universitas Jember
Jl. Kalimantan 37, Jember 68121, Indonesia
e-mail: yenirisma27@gmail.com, itirta.fmipa@unej.ac.id, yulidewi.fmipa@unej.ac.id

Abstract. Covid-19 is still a global concern. From the first time, this virus was detected, on December 31, 2019. As of March 20, 2022, there were 460 million positive cases of Covid-19, with 6.06 million deaths worldwide. The high number of Covid-19 cases is due to the rapid spread of this virus. One way to prevent the spread of this virus is by early detection of the disease and mapping the influence factors .The classification method with the support vector machine (SVM) method in machine learning can predict individuals diagnosed as positive for Covid-19 and who do not use the factors considered influential. Traditionally this can also be done with a generalized linear model (GLM). This study aims to compare two methods (SVM and GLM) in predicting individuals diagnosed as positive for Covid-19. In addition, this study also conducted an ensemble between SVM and GLM to determine whether the ensemble performed could produce better accuracy than the single classifier (SVM and GLM). The results showed that the accuracy with SVM and GLM was relatively high. However, SVM is slightly superior with 98.91% accuracy, and GLM with 95.64% accuracy. Meanwhile, the ensemble of both models achieved 98.91% accuracy, as high as SVM.

Keywords: Covid-19, Klasifikasi, *Machine Learning* SVM, GLM

1. Pendahuluan

SARS-CoV-2 (*Severe Acute Respiratory Syndrome Coronavirus 2*) atau Covid-19 merupakan salah satu jenis virus zoonotik yang masih satu keluarga dengan SARS (*Severe Acute Respiratory Syndrome*) dan MERS (*Middle East Respiratory Syndrome*). Saat ini, Covid-19 masih menjadi perhatian dunia. Sejak pertama kali virus ini terdeteksi yaitu pada 31 Desember 2019 hingga 20 Maret 2022, tercatat bahwa ada 460 juta kasus positif Covid-19 dengan kasus kematian sebanyak 6,06 juta di seluruh dunia. Tingginya kasus Covid-19 dikarenakan penyebaran dari virus ini sangat cepat. Salah satu cara untuk mencegah penyebaran virus ini dapat dilakukan dengan deteksi dini dari penyakitnya serta memetakan faktor-faktor yang mempengaruhinya.

Metode klasifikasi dengan *support vector machine* (SVM) pada *machine learning* dapat digunakan untuk memprediksi individu yang terdiagnosa positif Covid-19 dan yang tidak, dengan menggunakan faktor-faktor yang diperkirakan berpengaruh. Pada penelitian

sebelumnya yaitu milik Prahartiwi [4] tentang Komparasi Algoritma *Naive Bayes*, *Decision Tree* dan *Support Vector Machine* untuk Prediksi Penyakit Kanker Payudara menghasilkan metode terbaik yaitu metode *Support Vector Machine*. Menurut Gunn [3], dalam kasus *binary classification* SVM dapat melakukan klasifikasi dengan cara memaksimalkan margin untuk mencari *hyperplane* terbaik sebagai pemisah antara dua buah kelas. Secara tradisional, untuk deteksi dini Covid-19 juga bisa dilakukan dengan metode g*eneralized linear model* (GLM). GLM merupakan metode pengembangan dari model linier klasik dimana proses klasifikasinya dilakukan dengan mengetahui hubungan antara variabel independent dan dependent [6]. Penelitian sebelumnya tentang metode *Generalized Linear Model* milik Souza *et al* [5] yang berjudul *The overlooked potential of Generalized Linear Models in astronomy, I: Binomial regression* menghasilkan metode *Generalized Linear Model* lebih baik dibandingkan dengan metode *Artificial Neural Networks* untuk penyelesaian masalah *binary* klasifikasi.

Tujuan dari penelitian ini adalah untuk mengetahui perbandingan hasil kinerja antara metode SVM dan GLM dalam memprediksi individu terdiagnosis positif Covid-19. Selain membandingan kedua metode tersebut, dalam penelitian ini juga dilakukan *ensemble* antara SVM dan GLM untuk mengetahui apakah *ensemble* dapat menghasilkan model dengan akurasi yang lebih baik dari *single classifier* nya (SVM dan GLM).

2. Metodologi

2.1 Data Penelitian

Data yang digunakan dalam penelitian merupakan data diagnosis Covid-19 dari 2573 individu. Data ini diperoleh dari *website* https://www.kaggle.com/hemanthhari yang terdiri dari 19 variabel yang terbagi menjadi 18 variabel independen dan 1 variabel dependent. Tabel 1 merupakan detail dari 19 variabel tersebut.

2.2 Langkah Penelitian

Adapun langkah-langkah dalam penelitian ini adalah sebagai berikut:

1. Mengambil data dari website https://www.kaggle.com/hemanthhari yang kemudian dibagi menjadi data *training* dan data *testing* dengan proporsi 75% untuk data *training* dan 25% untuk data *testing*.
2. Klasifikasi data menggunakan metode SVM. Langkah pertama dalam proses klasifikasi ini adalah melakukan training 4 buah model yaitu model dengan fungsi *kernel linear*, *polynomial*, RBF (*radial*), dan *sigmoid*. Tabel 2 berikut merupakan persamaan dari keempat fungsi kernel yang digunakan. Setelah dilakukan *training* model, kemudian dipilih model terbaiknya untuk melakukan *tuning parameter*. Proses *tuning parameter* dilakukan dengan tujuan untuk mencari nilai cost dan

gamma terbaiknya. Langkah terakhir yaitu melakukan proses *testing* dengan model terbaik beserta parameter *gamma* dan *costnya.*

Tabel 1. Variabel data

No	Variabel	Kategori	Label
1	Permasalahan pernafasan	Ya	1
		Tidak	0
2	Demam	Ya	1
		Tidak	0
3	Batuk kering	Ya	1
		Tidak	0
4	Sakit tenggorokan	Ya	1
		Tidak	0
5	Flu	Ya	1
		Tidak	0
6	Asma	Ya	1
		Tidak	0
7	Penyakit paru-paru kronis	Ya	1
		Tidak	0
8	Sakit kepala	Ya	1
		Tidak	0
9	Penyakit jantung	Ya	1
		Tidak	0
10	Diabetes	Ya	1
		Tidak	0
11	Darah tinggi	Ya	1
		Tidak	0
12	Kelelahan	Ya	1
		Tidak	0
13	Diare	Ya	1
		Tidak	0
14	Perjalanan keluar negeri	Ya	1
		Tidak	0
15	Riwayat kontak dengan pasien Covid-19	Ya	1
		Tidak	0
16	Menghadiri pertemuan besar	Ya	1
		Tidak	0
17	Mengunjungi tempat umum	Ya	1
		Tidak	0
18	Terdapat keluarga yang bekerja di tempat umum	Ya	1
		Tidak	0
19	Diagnosis Covid-19	Positif	1
		Negatif	0

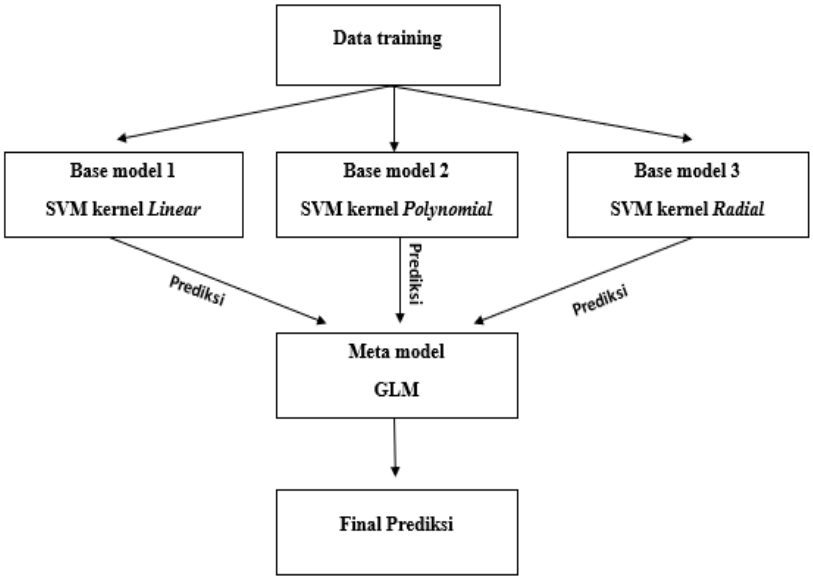
3. Klasifikasi data menggunakan metode GLM. Langkah pertama pada klasifikasi menggunakan GLM adalah melakukan *training* 2 buah model menggunakan *link function logit* dan *probit*. Setelah itu dilakukan pemilihan model terbaik dengan kriteria nilai AIC yang terkecil. Nilai AIC dapat diperoleh dari persamaan (1)

$$-2logL_{fit} + 2k \quad (1)$$

dengan *k* adalah jumlah dari variabel independen. Langkah selanjutnya adalah proses perbaikan model untuk menurunkan nilai AIC dengan metode *stepwise*. Metode *stepwise* merupakan metode gabungan dari *backward selection* dan *forward selection* dalam menentukan variabel yang merupakan prediktor terbaik untuk model [2]. Langkah terakhir yaitu dilakukan *testing* dengan model terbaik yang dihasilkan.

Tabel 2. Fungsi kernel pada SVM [1]

Jenis Kernel	**Persamaan**
Linier	$K(x,y) = x^T.y + c$
Polynomial	$K(x,y) = (\alpha x^T y + c)^d$
RBF (*radial basic function*)	$K(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
Sigmoid	$K(x,y) = tanh\ (\alpha x.y + \beta)$



Gambar 1. Proses pembelajaran *stacking ensemble*

4. Klasifikasi data menggunakan metode *ensemble* dari SVM dan GLM. *Ensemble* merupakan proses pembelajaran dengan penggabungan beberapa metode sekaligus. Pada penelitian ini metode *ensemble* yang digunakan adalah *stacking* ensemble. Proses pembelajaran dengan *stacking ensemble* dilakukan secara paralel. Gambar 1 merupakan proses dari *stacking ensemble* yang digunakan pada penelitian ini. *Base* model yang digunakan yaitu model SVM dengan *kernel linear*, SVM *kernel*

polynomial, dan SVM *kernel radial*. Sedangkan untuk *meta* model yang digunakan adalah model GLM.

5. Evaluasi hasil klasifikasi dengan melihat tabel *confussion matrix. Confussion matrix* merupakan algoritma yang digunakan untuk mengukur kinerja dari sebuah metode atau sistem klasifikasi. *Confussion matrix* dapat menggambarkan kinerja dari metode (sistem) dengan sebuah matriks. Dalam kasus *binary classification*, bentuk matriks keluarannya yaitu seperti pada Tabel 3.

Tabel 3. *Confussion matrix*

Kelas sebenarnya	**Kelas prediksi**	
	Positif	**Negatif**
Positif	TP (True Positif)	FN (False Negatif)
Negatif	FP (False Positif)	TN (True Negatif)

Berdasarkan Tabel 3 nantinya dapat diperoleh nilai akurasi yang dapat menggambarkan kinerja dari metode atau sistem yang digunakan. Untuk memperoleh akurasi dapat dihitung menggunakan persamaan (2)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

3. Hasil dan Pembahasan

Tabel 4 merupakan hasil pembagian data menjadi data *training* dan data *testing* dengan proporsi 75% data *training*, 25% data *testing.*

Tabel 4. Hasil pembagian data *training* dan *testing*

Data *training*	**Data *testing***	**Total**
1931	642	2573

Pada proses *training* model menggunakan SVM, hasil proses *training* nya dapat dilihat pada Tabel 5, dimana model terbaik yang diperoleh adalah model dengan fungsi kernel RBF (*radial*) dengan nilai akurasi sebesar 97,57%.

Tabel 5. Hasil proses *training* dengan SVM

Fungsi kernel	**Kategori**
Linear	95,39%
Polynomial	95,75%
RBF (*radial)*	97,57%
Sigmoid	92,35%

berdasarkan hasil Tabel 5, maka model yang dipilih untuk dilakukan *tuning parameter adalah* model dengan fungsi kernel RBF (*radial*). Dari proses *tuning parameter* diperoleh nilai *cost* terbaik yaitu 1,258925 dengan nilai *gamma* terbaik adalah 0,5. Selanjutnya dilakukan proses *testing* dan menghasilkan sebuah *confusion matrix* seperti pada Tabel 6. Berdasarkan Tabel 6 maka nilai akurasi dari metode SVM adalah 98,91%.

Tabel 6. *Confusion matrix* metode SVM

Kelas sebenarnya	**Kelas prediksi**	
	Positif	**Negatif**
Positif	373	0
Negatif	7	262

Hasil dari proses *training* model dengan GLM, model dengan *link function logit* menghasilkan nilai AIC sebesar 373,01 dan model dengan *link function probit* sebesar 376,49. Oleh karena itu, model yang dipilih untuk dilakukan perbaikan model dengan metode *stepwise* adalah model dengan *link function logit*. Dari proses perbaikan model, nilai AIC model dengan *link function logit* mengalami penurunan yaitu menjadi 366,91. Pada proses *testing* dihasilkan *confusion matriks* untuk metode GLM adalah seperti Tabel 7. Berdasarkan Tabel 7, maka nilai akurasi dari metode GLM dengan model *link function logit* mempunyai akurasi sebesar 95,64%.

Tabel 7. *Confussion matrix* metode GLM

Kelas sebenarnya	**Kelas prediksi**	
	Positif	**Negatif**
Positif	364	16
Negatif	12	250

Tabel 8. *Confusion matrix* metode *ensemble*

Kelas sebenarnya	**Kelas prediksi**	
	Positif	**Negatif**
Positif	373	0
Negatif	7	262

Sedangkan untuk proses *ensemble*, proses *testing* datanya menghasilkan akurasi sebesar 98,91% dengan detail *confusion matrix* seperti pada Tabel 8.

4. Kesimpulan

Berdasarkan hasil penelitian, model yang dibentuk yaitu SVM, GLM, dan *ensemble* menunjukkan bahwa ketiga model tersebut dapat mengklasifikasikan data diagnosis Covid-19 dengan baik. Hal tersebut dapat dilihat dari nilai akurasi ketiga model yang cukup tinggi. Jika dibandingkan dengan metode GLM dengan akurasi sebesar 95,64%, metode SVM sedikit lebih unggul dengan akurasi 98,91%. Sedangkan untuk model *ensemble* dari SVM dan GLM mencapai akurasi yang cukup tinggi pula yaitu 98,91%, akan tetapi masih belum lebih tinggi dibandingkan dengan *single classifier* yaitu SVM.

Daftar Pustaka

[1] Cholissodin, I., Sutrisno, A.A. Soebroto, U. Hasanah, dan Y. I. Febiola. (2020). *AI, Machine Learning dan Deep Learning (Teori & Implementasi)*. Malang: Universitas Brawijaya.

[2] Faulina, R. (2017). Penggunaan Regresi Stepwise Untuk Menentukan Faktor yang Mempengaruhi Motivasi Santri Melanjutkan Studi Ke Perguruan Tinggi (Studi Kasus SMK Ibnu Cholil Bangkalan). *Jurnal Matematika Sains dan Teknologi*, **18(2)**, 68-75, https://doi.org/10.33830/jmst.v18i2.129.2017

[3] Gunn, S. R. 1998. *Support Vector Machine for Classification and Regression*. Southamton: University of Southamton.

[4] Prahartiwi, L. I., dan D. Wulan. (2021). Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Penyakit Kanker Payudara. *Jurnal Teknik Komputer AMIK BSI*, 51-54.

[5] Souza, R. S., E. Cameron., M. Killedar., J. Hilbe., R. Vilalta., U. Maio., V. Biffi., B. Ciardi., dan J. D. Riggs. 2015. The overlooked potential of Generalized Linear Models in astronomy, I: Binomial regression. *Astronomy and Computing*, **12**, 21-32.

[6] Zahro, J., R. E. Caraka, dan R. Herliansyah. 2018. *Aplikasi Generalized Linear Model pada R. Edisi 1*. Yogyakarta: Innosain.