Prediksi Penderita Tuberkulosis Menggunakan Algoritma *Support Vector Regression* (SVR)

Prediction Of Tuberculosis Patients Using The Support Vector Regression (SVR)

1st Ridha Melati N
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
ridhamelati@student.telkomuniversity.ac.id

2nd Tito Waluyo Purboyo
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
titowaluyo@telkomuniversity.ac.id

3rd Meta Kallista
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
metakallista@telkomuniversity.ac.id

Abstrak— Salahsatu penyakit menular yang menjadi topik pembahasan yang ramai di dunia kesehatan adalah Tuberkulosis (TBC). Karena TBC merupakan salahsatu dari 10 penyakit yang menjadi penyebab utama kematian di seluruh dunia dan di Indonesia berada peringkat ketiga dengan kasus tertinggi setelah India dan China. Hal tersebut menjadikan penyakit ini perlu adanya suatu peramalan atau prediksi ke depannya sehingga masyarakat mengantisipasi lebih awal.Dalam penelitian tugas akhir ini penulis akan membuat sistem Prediksi Penderita Tuberkulosis. Hasil dari penelitian ini berupa prediksi jumlah penderita kedepannya. Data yang digunakan berasal dari Dinas Kesehatan Kabupaten Karawang periode 1 Januari 2020 sampai 31 Desember 2021.Sistem Prediksi Penderita Tuberkulosis ini menggunakan metode *Support Vector Regression* dan menggunakan kernel *Radial Basis Function* yang menghasilkan nilai error performansi *Mean Square Error* (MAE) sebesar 0.099448, *Root Mean Square Error* (RMSE) sebesar 0.136204 dan R^2 sebesar 0.220323.

Kata kunci— penyakit tuberkulosis, prediksi, *support vector regression*

Abstrak— One of the infectious diseases that has become a topic of discussion that is crowded in the health world is Tuberculosis (TBC). Because TBC is one of the 10 diseases that are the leading cause of death worldwide and in Indonesia is ranked third with the highest cases after India and China. This makes this disease necessary to have a forecast or prediction in the future so that the public anticipates it early. In this final project research, the author will create a system for Predicting Tuberculosis Patients. The results of this study are in the form of predictions of the number of sufferers in the future. The data used came from the Karawang Regency Health Office for the period January 1, 2020 to December 31, 2021. This Tuberculosis Patient Prediction System uses the Support Vector Regression method and uses the Radial Basis Function kernel which produces a Mean Square Error (MAE) performance error value of 0.099448, Root Mean Square Error (RMSE) of 0.136204 and R^2 of 0.220323.

I. PENDAHULUAN

Tuberkulosis (TBC) merupakan suatu penyakit menular yang disebabkan oleh bakteri *Mycobacterium Tuberculosis*, penyakit ini juga merupakan salah satu dari 10 penyebab utama kematian di seluruh dunia. Indonesia berada pada peringkat ketiga dengan kasus tuberkulosis tertinggi di dunia setelah India dan China dengan jumlah kasus 824.000 dan kematian 93.000 per tahun atau dengan 11 kasus kematian per jam. Pada penyakit ini bila pengobatannya yang tidak tuntas dan tidak optimal maka dapat menimbulkan komplikasi berbahaya hingga berujung pada kematian. Informasi tentang jumlah penderita TBC saat ini dilakukan dengan manual yaitu langsung ke balai kesehatan.

Berdasarkan masalah di atas maka perlu adanya penanganan yaitu adanya suatu peramalan atau prediksi ke depannya yang diimplementasikan ke webiste sehingga masyarakat melakukan tindakan pencegahan lebih awal sebelum ke tingkat yang lebih parah. Salah satu cara untuk mengetahui tingkat penderita penyakit tuberkulosis kedepannya yaitu memprediksi jumlah penderita penyakit tuberkulosis menggunakan *Support Vector Regression* . *Support Vector Regression* adalah pengembangan dari SVM untuk kasus regresi dengan output berupa bilangan rill (nyata) atau data sekuensial. *Support Vector Regression* juga merupakan algoritma yang dapat mengatasi masalah *overfitting*.

Dengan adanya penelitian ini diharapkan dapat menjadi solusi bagi masyarakat dalam menangani penyebaran penyakit tuberkulosis serta dengan metode yang digunakan dapat memberikan performansi yang lebih baik.

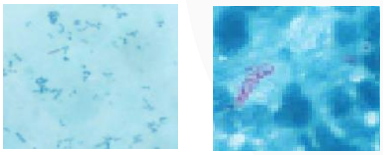
II. KAJIAN TEORI

A. Penyakit Tuberkulosis

Tuberkulosis (TBC) merupakan penyakit yang disebabkan oleh infeksi bakteri *Mycrobacterium Tuberculosis* yang paling sering menyerang organ paru-paru. Menurut laporan data WHO *Global Tuberculosis Report*, jumlah kasus kematian global secara resmi disebabkan oleh penyakit tuberkulosis yang berjumlah 1,3 juta orang tahun 2020 dan jumlah kematian penyakit tuberkulosis berdampak lebih parah oleh COVID-19 pandemi pada tahun 2020 [4]. Dan tahun 2013-2018, jumlah kasus penyakit tuberkulosis di Indonesia tepatnya di Kabupaten Karawang mengalami kenaikan dan penurunan. Pada tahun 2013-2015 kasusnya meningkat dan pada tahun 2016-2017 kasusnya menurun. Dan pada tahun 2018 ada 1.308 kasus menurun dibanding pada tahun 2017 dengan kasus 1.116 [5].

Adapun diagnosis tuberkulosis yang mempunyai berbagai gejala, dan beberapa pasien tidak menunjukkan gejala bahkan jika pasien tersebut didiagnosis dengan tuberkulosis pada saat pemeriksaan. Pada penderita Tuberkulosis, ada beberapa jenis pemeriksaan yang dilakukan yaitu dengan pemeriksaan bakteriologi dan pemeriksaan radiologi. Adapun penjelasan masing-masing pemeriksaan penderita TBC yaitu sebagai berikut [6]:

1. Pemeriksaan bakteriologi merupakan tes yang dilakukan dengan memeriksa dan menghitung jumlah bakteri tuberkulosis dalam tubuh seseorang. Bahan tes biasanya berupa dahak pasien dan urin juga bisa digunakan. Setelah dahak pasien terkumpul, kemudian akan dilakukan pemeriksaan mikroskopis. Tes mikroskopis bertujuan untuk menghitung jumlah bakteri yang terdapat pada dahak. Pada pemeriksaan ini, pertama dilakukan metode pewarnaan pada dahaknya dengan menggunakan perwarnaan *Ziehl-Neelsen*. Hal ini tujuannya untuk mengetahui ada atau tidaknya bakteri Basil Tahan Asam (BTA).



(a) (b)

GAMBAR 2. 1
HASIL PEMERIKSAAN (A) NEGATIF (B) POSITIF

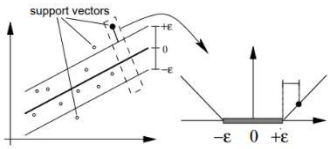
2. Pemeriksaan Radiologi merupakan tes yang dilakukan pada toraks paru-paru pasien. Pemeriksaan ini mempunyai tujuan untuk mengetahui kondisi bagian dalam tubuh pasien.

B. Support Vector Machine

Support Vector Machine (SVM) pertama kali dipublikasikan pada tahun 1992 oleh Vladimir wapnik dan dua rekannya yaitu Bernhard bozer dan Isaberg guyon. Pengertian dari *Support Vector Machine* merupakan salahsatu algoritma machine learning yang efektif dalam membangun pengklasifikasian. Dalam algoritma *Support Vector Machine* dapat juga digunakan di kasus regresi yaitu menggunakan Support Vector Regression. Tetapi tujuan kedua algoritma ini sama yaitu dapat meminimalkan kesalahan *hyperplane* yang memaksimalkan margin yang merupakan bagian dari kesalahan yang ditoleransi [8]. Adapun perbedaan dari *Support Vector Machine* dan *Support Vector Regression* adalah SVM digunakan untuk kasus klasifikasi dengan output berupa diskrit sedangkan SVR digunakan untuk kasus regresi dengan output bersifat kontinu atau bilangan riil.

C. Support Vector Regression

Support Vector Regression merupakan sebuah pengklasifikasian dari metode machine learning yang diterapkan dalam kasus regresi dan dalam output bernilai bilangan bilangan rill atau data sekeunsial [3]. Kelebihan lain dari SVR adalah dapat digunakan dengan relatif baik untuk data berdimensi tinggi karena dapat dikontrol secara eksplisit dengan memilih parameter C dan epsilon yang sesuai [10].



GAMBAR 2. 2
FUNGSI *SUPPORT VECTOR REGRESSION* [11]

Pada gambar 2.2 terlihat bahwa garis tebal diantara dua garis merupakan garis *hyperplane*. *Hyperplane* merupakan garis pemisah antar data. *Hyperplane* diapit oleh $+\varepsilon$ dan garis batas $-\varepsilon$. ε adalah jarak antara *hyperplane* dan dua garis yang berdekatan. *Support vector* adalah data yang paling dekat dengan margin. Margin adalah jarak antara data terdekat dengan *hyperplane*.

Dalam membuat sebuah *hyperplane* perlu adanya parameter yang akan ditentukan dalam mencoba beberapa nilai rentang. Adapun parameter tersebut yaitu [12] :

1. Parameter Complexity (C)
   Parameter ini memberikan *trade-off* pada kompleksitas model serta jumlah maka nilai C yang mempunyai penyimpangan lebih besar dapat ditoleransi.
2. Parameter Epsilon (ε)
   Parameter ini berfungsi untuk menyesuaikan data pelatihan dan untuk mengatur batas kesalahan dalam fungsi f(x).
3. Parameter Gamma (γ)
   Parameter ini berfungsi sebagai parameter fungsi kernel yang digunakan untuk mendapatkan nilai gamma.

1. Regresi

Regresi adalah pengukur hubungan antara dua variabel atau lebih yang diperoleh dalam bentuk hubungan atau fungsi. Hubungan antar variabel bebas dan variabel terkait, disimbolkan dengan x dan y. Untuk memprediksi seberapa bagus nilai model regresi yang digunakan, perlu adanya pengukuran performa. Untuk mengukur keakuratan model regresi yang digunakan, maka perlu menghitung nilai error antara data yang diprediksi dan data yang sebenarnya. Adapun beberapa nilai error yang umum digunakan adalah [14]:

a. MAE (Mean Absolute Error)

Nilai error MAE merupakan nilai error yang mencari kesalahan absolut paling sederhana.
Adapun rumusnya yaitu :

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \quad (2.1)$$

b. MSE (Mean Absolute Error)

Nilai error MAPE merupakan nilai error yang mencari kesalahan absolut di setiap periode yang dibagi dengan nilai observasi pada periode tertentu.
Adapun rumusnya yaitu:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (2.2)$$

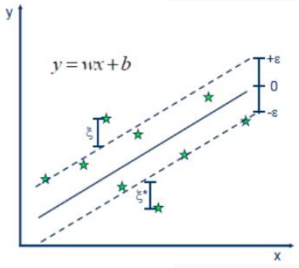
c. RMSE (Root Mean Absolute Error)

Nilai error RMSE merupakan nilai error yang digunakan untuk menghitung seberapa bedanya nilai-nilai. Adapun rumusnya yaitu :

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (2.3)$$

2. Support Vector Regression Linear

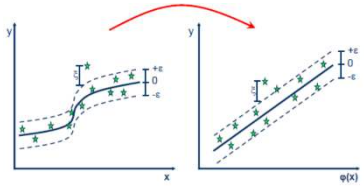
Support Vector Regression dan *Support Vector Machine* menggunakan fungsi kernel linier yang sama untuk regresi. Tetapi tidak seperti *Support Vector Machine*, *Support Vector Regression* menentukan batas toleransi dengan (ε) untuk prediksi [15].



GAMBAR 2. 3
MODEL *SUPPORT VECTOR REGRESSION LINIER* [8]

3. Support Vector Regression Non- Linear

Support Vector Regression non-linier ini menggunakan fungsi kernel non-linier untuk memproses data training di ruang fitur.



GAMBAR 2. 4
MODEL SUPPORT VECTOR REGRESSION NON-LINIER [8]

4. Kernel

Kernel merupakan fungsi yang dapat memetakan data ke ruang fitur dengan dimensi tinggi agar data dapat diproses menjadi terstruktur dengan efisien. Beberapa fungsi kernel yang dapat digunakan dalam menyelesaikan masalah SVR yaitu sebagai berikut [17]:

a. Linear

Linear kernel adalah fungsi kernel yang sangat sederhana. Adapun persamaanya yaitu :

$$k(x_1, x_2) = x_1^T x_2 + c \quad (2.7)$$

Ket :
$k(x_1, x_2)$ = Fungsi kernel yang menunjukkan pemetaan linier pada *feature space*
x_1 = Variabel input
x_2 = Fungsi basis
c = Konstanta
T = Parameter kernel

b. Polynomial

Polynomial kernel adalah kernel yang mewakili kesamaan vektor (data pelatihan) dalam ruang fitur polinomial dari variabel asli, yang memungkinkan pelatihan model nonlinier. Adapun Persamaannya yaitu:

$$k(x_1, x_2) = (\beta x_1^T x_2 + c)^d \quad (2.8)$$

Ket :
$k(x_1, x_2)$ = Fungsi kernel
x_1 = variabel input
x_2 = Fungsi basis
β = Parameter untuk kemiringan
T = Parameter kernel
c = Konstanta
d = Tingkat Polinomial

c. Radial Basis Function (RBF)

Radial Basis Function (RBF) kernel adalah fungsi kernel yang paling umum digunakan. RBF ini digunakan dalam analisis ketika data dipisahkan secara linear. Adapun persamaannya yaitu:

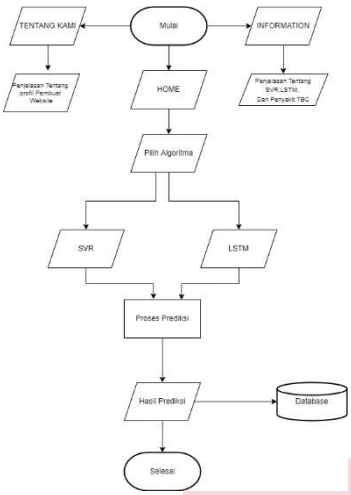
$$k(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2) \quad (2.9)$$

Ket :
$k(x_1, x_2)$ = Fungsi kernel
x_1 = variabel input
x_2 = inti yang dipilih dari data pelatihan
σ = spread
$\gamma = ½\sigma^2 > 0$ adalah parameter lebar

III. METODE

A. Gambaran Umum Sistem

Gambaran umum sistem dalam penelitian ini mencakup sistem Prediksi Tuberkulosis menggunakan algoritma *Support Vector Regression (SVR)*. Data yang digunakan dari Dinas Kesehatan Kabupaten Karawang periode Januari 2020 sampai Desember 2021 yang kemudian akan diimplementasikan di website. Output yang dihasilkan berupa gambar grafik dan tabel prediksi selama 30 hari kedepan dari tanggal terakhir data dikumpulkan. Adapun gambaran sistem ini adalah sebagai berikut:



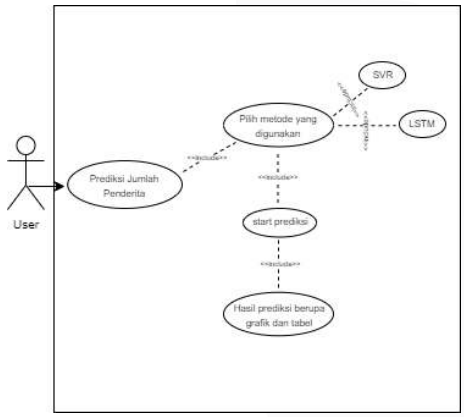
GAMBAR 3. 1
GAMBARAN UMUM SISTEM

B. Perancangan Perangkat Lunak Web

Adapun perancangan perangkat lunak web pada penelitian ini adalah :

1. Use Case Diagram

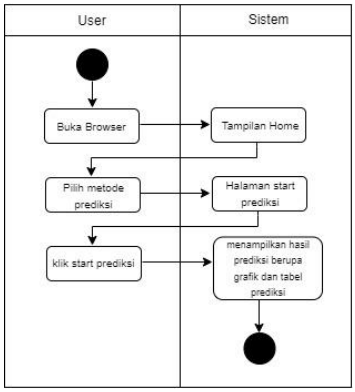
Use diagram digunakan untuk menggambarkan interaksi sistem dengan user. Berikut adalah *Use Case* Diagram dari prediksi penderita penyakit tuberkulosis.



GAMBAR 3. 2
USE CASE DIAGRAM

2. Activity Diagram

Pada activity diagram menjelaskan aktifitas dalam perancangan perangkat lunak yang dilakukan pengguna selama mengakses website.



GAMBAR 3. 3
ACTIVITY DIAGRAM

C. Perancangan Data

Sebelum data diolah yang akan dimasukkan ke model *machine learning*, terlebih dahulu kita melakukan persiapan dan perancangan data.

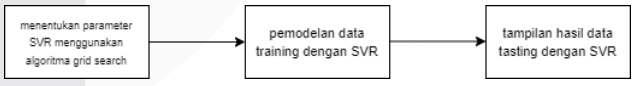
1. Data

Data yang akan menjadi bahan penelitian didapatkan dari Data *raw* Dinas Kesehatan Kabupaten Karawang dengan periode 1 Januari 2020 sampai 31 Desember 2022. Berikut tampilan data raw dari Dinas Kesehatan Kabupaten Karawang.

2. *Preprocessing Data*

Sebelum data diolah perlu persiapan data agar bisa diolah oleh algoritma, sehingga didapatkan hasil prediksi. Berikut langkah-langkahnya :

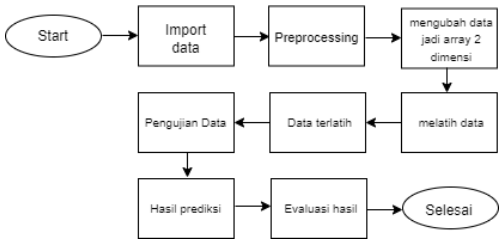
a. Setelah data terkumpul yaitu Data *raw* dari Dinas Kesehatan Kabupaten Karawang kemudian di proses dan data nya digabung sehingga hanya dua parameter saja yang ditampilkan. Kemudian data tersebut dijadikan kedalam format csv.
b. Mengubah data kosong jumlah penderita menjadi Nan. Tujuan dari proses ini agar data yang kosong dapat diisi dengan interpolasi di proses selanjutnya.
c. Kemudian data yang bernilai NaN akan diisi. Ada banyak cara untuk memasukkan data NaN, salah satunya adalah interpolasi linier. Interpolasi linier merupakan cara untuk menentukan nilai yang berada di antara dua nilai diketahui berdasarkan persamaan linier.
d. Data hasil yang sudah diinterpolasi akan menyambung. Data ini digunakan untuk mengganti nilai Nan agar menjadi nilai real. Berikut grafik hasil interpolasi data



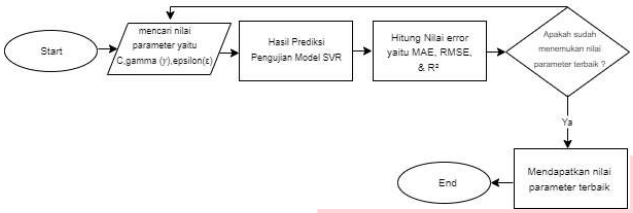
Gambar 3. 3 Diagram Blok Perancangan Algoritma

D. Perancangan Algoritma Support Vector Regression

Perancangan algoritma sistem ini menggunakan algoritma *Support Vector Regression* (SVR) dan juga menggunakan fungsi kernel RBF dalam mencariparameter terbaik. Berikut adalah diagram blok perancangan algoritma.



GAMBAR 3. 4
DIAGRAM ALIR METODE SVR



GAMBAR 3. 5
DIAGRAM ALIR MENDAPATKAN ANGKA PARAMETER TERBAIK DENGAN ALGORITMA GRID SEARCH

IV. HASIL DAN PEMBAHASAN

A. Pengujian Partisi Data

Dalam pengujian partisi data menggunakan algoritma grid search dan menggunakan parameter default Support Vector Regression yaitu C =1, epsilon (ε) =0.1, dan gamma (γ) = 1. Pada pengujian menggunakan algoritma grid searh, parameter yang akan dicari yaitu gabungan nilai C , epsilon (ε) , dan gamma (γ) yang menghasilkan nilai error paling kecil. Kemudian kernel yang digunakan yaitu kernel RBF. Adapun hasil pengujian dapat dilihat dibawah ini.

TABEL 4. 1
HASIL PENGUJIAN PARTISI DATA MENGGUNAKAN ALGORITMA GRID SEARCH

Training	Testing	MAE	RMSE	R²
50%	50%	0.099448	0.136204	0.220323
60%	40%	0.107986	0.147269	0.197817
70%	30%	0.108025	0.147434	0.196024
80%	20%	0.135180	0.183921	0.039142
90%	10%	0.176314	0.230920	-0.229830

TABEL 4. 2
HASIL PENGUJIAN PARTISI DATA MENGGUNAKAN PARAMETER DEFAULT SVR

Partisi Data	MAE	RMSE	R²
50% 50%	0.103035	0.139096	0.186860
60% 40%	0.108743	0.148173	0.187938
70% 30%	0.115321	0.159945	0.180456
80% 20%	0.136944	0.186410	0.012968
90% 10%	0.179453	0.234660	-0.269992

B. Pengujian Parameter SVR

Dalam pengujian parameter *Support Vector Regression* dilakukan dengan pencarian nilai *C*, *epsilon* dan *gamma*. Pengujian parameter ini bertujuan untuk mencari parameter terbaik dengan memakai kernel RBF.

TABEL 4. 3
HASIL PENGUJIAN PARAMETER SVR

No	C	Epsilon	Gamma	MAE	RMSE	R²
1	100	0.01	0.01	0.106284	0.147694	0.083224
2	100	0.01	0.05	0.104732	0.144088	0.127442
3	100	0.01	0.2	0.106465	0.155304	-0.013682
4	100	0.01	0.5	0.100980	0.137380	0.206797
5	100	0.01	1	0.099448	0.136204	0.220323
6	100	0.1	0.01	0.105990	0.144174	0.126399
7	100	0.1	0.1	0.105126	0.144050	0.127900
8	10	0.1	1	0.101938	0.137905	0.200721
9	1	0.1	1	0.103035	0.139096	0.186860

V. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan pada tugas akhir ini dapat diambil kesimpulan adalah sebagai berikut:

1. Sistem prediksi penderita tuberkulosis berbasis website ini dapat berjalan dengan baik. Sistem yang buat berhasil melakukan prediksi selama tiga puluh hari ke depan dari tanggal 01 Januari 2022 - 30 Januari 2022. Dari hasil pengujian alpha didapatkan sebesar 100% dan hasil pengujian beta didapatkan dari 42 responden sebesar 83,6 % dimana responden memilih sangat baik dan baik penggunaan websitenya. Oleh karena itu, website ini sudah cukup digunakan untuk melakukan prediksi
2. Metode yang digunakan pada sistem prediksi penderita tuberkulosis berbasis website ini adalah *Support Vector Regression* dengan hasil pengujian data *training* dan data *testing* yang diuji dengan nilai berbeda, telah didapat hasil terbaik dengan nilai data *training* 50% dan data *testing* 50%. Dari hasil tersebut mendapatkan nilai error MAE = 0.099448, nilai RMSE = 0.136204 dan nilai R² = 0.220323.

REFERENSI

[1] M. drg.Oscar Primadi, “Profil Kesehatan Indonesia 2020,” Jakarta, Kementerian Kesehatan Republik Indonesia, 2020.
[2] M. Indah, “Infodatin Pusat Data dan Informasi Kementerian Kesehatan RI,” Jakarta, ISSN , 2018.
[3] R. K. Mariette Awad, “Efficient Learning Machines,” dalam *Support Vector Regression*, Berkeley,CA, IEEE, 2015.
[4] WHO, “Global Tuberculosis Report,” WHO, 2021.
[5] D. H. N. Hidayat, “Profil Kesehatan Kabupaten Karawang Tahun 2018,” dalam *Pengendalian Penyakit*, Karawang, Dinas Kesehatan Pemerintah Kabupaten Karawang , 2019.
[6] Majdawati, "Uji Diagnostik Gambaran Lesi Foto Thorax pada penderita dengan klinis tuberkulosis paru" mutiara medika, 2010

[7] J. Y. Wu, "Housing Price Prediction Using Support Vector Regression," *Computer Science,* 2017.
[8] S. Rutgers, "Support Vector Machine-Regression (SVR),"*https://www.saedsayad.com/support_vector_machine_reg.htm,* 2019.
[9] J. P. Nikmatun Khasanah, "Identifikasi Skor Kebahagiaan, Metode Support Vector Regression," Jurnal Ilmiah Matematika, 2022.
[10] S. P. Shom Prasad Das, "Support Vector Machines for Prediction of Futures Prices in Indian Stock Market," *International Journal of Computer Applications,* Vol.1, 2012.
[11] P. Cortez, "A Data Mining Approach to Predict Forest Fires Using Meteorological Data," *IEEE,* 2007.
[12] S. Shataee, S. K, A. Fallah, & D. Pelz, "Forest Attribute Imputation using Machine Learning methods and aster data: Comparison of k-NN, SVR and Random Forest Regression Algorithms," Int. J. Remote Sens, vol. 33, 2012
[13] B. Yuniarto & Robert Kurniawan, Analisis Regresi : Dasar dan penerapannya dengan R, Jakarta: Kencana, 2016.
[14] M. J. W. G. J. Davide Chicco, "The Coefficient of determination R-Squared is more informative than SMAPE,MAE,MAPE,MSE,RMSE in regression analysis evaluation," *PeerJ Computer Science*, 2021.
[15] S. V. d. R. S.Kavitha, "A comparative analysis on linear regression and support vector regression," dalam *International Conferance*, 2016.
[16] M. H. Hekmatyar, "Peramalan Jumlah Kasus Demam berdarah di Kabupaten Malang menggunakan metode Support Vector Regression," *Sistem Informasi,* 2019.
[17] M. G. &. M. F. Lorenzi L, "Support vector regression with kernel combination for missing data reconstruction," *IEEE Geoscience and Remote Sensing Letters,* 2013.
[18] F. Yusup, "Uji Validitas dan Reliabilitas Instrumen Penelitian Kuantitatif," *Jurnal Tarbiyah:Jurnal Ilmiah Kependidikan*, 2018.