JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

PENERAPAN REGRESI LINEAR BERGANDA UNTUK MEMPREDIKSI DIABETES SECARA DINI

Ayunita Sari[1*], Dian Utami Putri[2]
[1,2]Sistem Informasi, STMIK Royal Kisaran
email: [*]yunitasari77@gmail.com

Abstract: *Diabetes is a long-lasting disease characterized by high or above normal blood sugar (glucose) levels. Lack of public information about diabetes and constraints on the problem of hospital costs make people reluctant to check themselves at the health center, therefore this can help people find information early on to better maintain their health so they don't suffer from diabetes. Predictive efforts are needed to find out the approximate outcome of diagnosing diabetes in someone early on with Pregnancies, Glucose (sugar), BloodPressure (blood pressure), SkinThickness, Insulin, BMI (Weight), DiabetesPedigreeFunction, Age (Age). Furthermore, predictions use the regression method where there are 7 more methods including (1) Linear Regression, (2) Support Vector Regression - Linear, (3) Support Vector Regression - RBF, (4) Decision Tree Regression, (5) Random Forest Regressor, (6) Gradient Boosting Regression, (7) NLP Regressor applied in this study. The aim of the research is to be able to determine the best regression method based on 7 regression methods with the best accuracy value that will be used in the deploy process to predict the outcome of diagnosing diabetes in someone from an early age. Decision tree regression as the best method among other regression methods in 4 accuracy tests with ratios of 90:10, 80:20, 70:30 and 60:40.*

Keywords: *diabetes, decision tree, regression*

Abstrak: Diabetes merupakan penyakit yang berlangsung lama ditandai dengan kadar gula (glukosa) darah yang tinggi atau diatas normal. Kurangnya informasi masyarakat mengenai penyakit diabetes dan kendala masalah biaya kerumah sakit membuat masyarakat enggan untuk memeriksakan dirinya ke puskesmas, oleh karena itu ini dapat membantu masyarakat mengetahui informasi sejak dini untuk lebih menjaga kesehatan agar tidak menderita diabetes. Diperlukannya upaya prediksi untuk mengetahui perkiraan hasil diagnose diabetes pada seseorang sejak dini dengan Pregnancies, *Glucose* (gula), BloodPressure (tekanan darah), SkinThickness , Insulin, BMI (Berat Badan), DiabetesPedigreeFunction, *Age* (Umur). Selanjutnya prediksi menggunakan metode regresi dimana terdapat 7 metode lagi meliputi (1) *Linear Regression*, (2) *Support Vector Regression – Linear*, (3) *Support Vector Regression – RBF*, (4) *Decision Tree Regression*, (5) *Random Forest Regressor*, (6) *Gradient Boosting Regression*, (7) NLP *Regressor* yang diterapkan dalam penelitian ini. Tujuan penelitian untuk dapat menentukan metode regresi terbaik berdasarkan 7 metode regresi dengan nilai akurasi paling terbaik yang akan digunakan dalam proses pengolahan deploy untuk prediksi hasil diagnose diabetes pada seseorang sejak dini. *Decision tree regression* sebagai metode terbaik diantara metode regresi lainnya dalam 4 pengujian akurasi dengan rasio 90:10, 80:20, 70:30 dan 60:40.

Kata Kunci: diabetes, decision tree, regresi

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

PENDAHULUAN

Pada zaman yang semakin modern saat ini, banyak manusia yang membutuhkan suatu alat bantu yang praktis, untuk mempermudah manusia melakukan berbagai kegiatannya. Teknologi mempunyai peranan yang sangat penting untuk menunjang kemudahan itu. Sudah banyak teknologi yang diciptakan oleh manusia untuk mewujudkan keinginan manusia itu sendiri. Upaya yang dilakukan ini agar kita tidak perlu repot-repot untuk melakukan aktifitas yang melelahkan atau untuk mendapatkan hasil yang lebih maksimal. Begitu pesatnya perkembangan teknologi didunia, juga berimbas kepada kemajuan teknologi pada bidang kesehatan khususnya dalam memprediksi diabetes.

Diabetes merupakan penyakit yang berlangsung lama atau kronis yang ditandai dengan kadar gula (glukosa) darah yang tinggi atau diatas normal. Glukosa yang menumpuk didalam darah akibat tidak diserap sel tubuh dengan baik dapat menimbulkan berbagai gangguan organ tubuh yang menyebabkan timbulnya berbagai komplikasi yang membahayakan nyawa penderitanya. Menurut Federasi Diabetes Internasional, ada 285 juta orang diabetes diseluruh dunia. Total ini diperkiraksan akan meningkat menjadi 380 juta dalam 20 tahun. Untuk mengetahui seseorang terkena diabetes atau tidak, penerapan teknologi membantu untuk memberikan solusi untuk mendiagnosis seseorang apakah kemungkinan terkena diabetes.

Dalam mendiagnosis seseorang apakah terkena diabetes, tentunya dokter akan melakukan pemeriksaan pada pasien. Namun saat ini masyarakat masih belum sepenuhnya paham dan menyadari apakah ia terkena diabetes, karena untuk mengecek apakah ia terkena diabetes tentunya membutuhkan biaya dalam pemeriksaan nya, kemudian juga jauhnya rumah sakit ataupun puskesmas terdekat, ada juga yang takut melakukan pemeriksaan apakah ia terkena diabetes dikarenakan takut akan hasil pemeriksaan yang dilakukan nantinya.

Dari permasalahan diatas maka para peneliti tertarik untuk melakukan proses penelitian mengenai prediksi diabetes. Ini dilakukan guna membantu masyarakat mendapat peringatan dini untuk lebih menjaga kesehatan agar tidak menderita diabetes dan memberitahu masyarakat betapa pentingnya untuk melakukan pemeriksaan kesehatan dan juga guna membantu ahli medis dalam menganalisis diagnosis diabetes melalui penerapan teknologi informasi.

Maka dari itu diperlukannya teknologi komputer pada bidang kesehatan dalam menentukan klasifikasi penyakit demam berdarah dan menghasilkan suatu informasi yang lebih akurat. Metode regresi linear berganda yaitu model regresi yang melibatkan lebih dari satu variable independen. Melalui penerapan metode ini dalam menentukan Penyakit demam berdarah terdapat 8 kriteria yang diperlukan sebagai inputan dalam perhitungan metode regresi linier berganda, yaitu *Pregnancies*, *Glucose* (gula), *BloodPressure* (tekanan darah), *SkinThickness*, Insulin, BMI (Berat Badan), *DiabetesPedigreeFunction*, *Age* (Umur). Sehingga akan menghasilkan sebuah informasi tentang penyakit diabetes berdasarkan semua kriteria yang ada. Dengan adanya teknologi ini diharapkan nantinya dapat menentukan penyakit diabetes, yang bisa dilakukan

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

dengan lebih cepat dan dapat mengurangi tingkat akurasi penentuan diabetes.Regresi linier berganda merupakan model persamaan yang menjelaskan hubungan satu variabel tak bebas/ response (Y) dengan dua atau lebih variabel bebas/ predictor (X1, X2,…Xn)[1].

Pada penerapan penelitian di bidang kesehatan, metode regresi linear berganda juga telah banyak digunakan salah satunya oleh Fitri Kurniyawati (2020) dengan judul "Pemodelan Faktor-Faktor Penyakit Diare Dengan Metode Regresi Linear Beganda" pada penelitian tersebut regresi linear berganda digunakan sebagai metode untuk memprediksi factor-faktor penyakit diare.

Pada penelitian ini juga bertujuan mendapatkan hasil prediksi dan mengetahui metode manakah yang menghasilkan nilai akurasi yang lebih akurat dengan membandingkan 7 model yang terdapat dalam regresi diantaranya sebagai berikut : (1) *Linear Regression*, (2) *Support Vector Regression – Linear*, (3) *Support Vector Regression – RBF*, (4) *Decision Tree Regression*, (5) *Random Forest Regressor*, (6) *Gradient Boosting Regression*, (7) NLP *Regressor*.

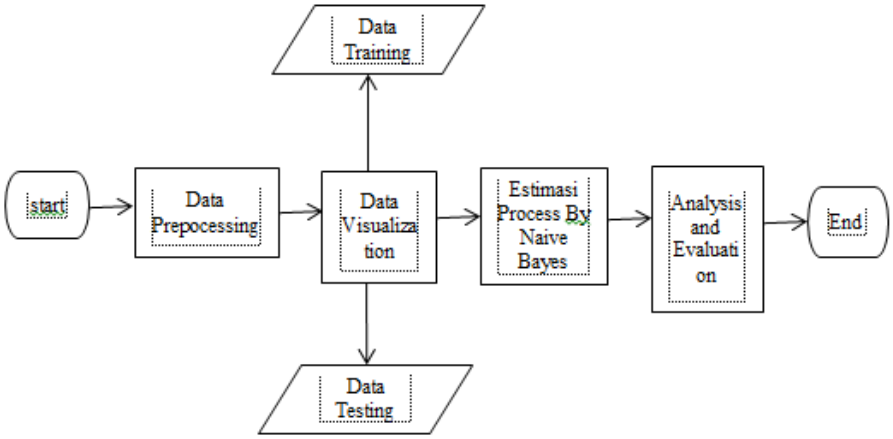
Dalam hasil pengujiannya menggunakan model pengukuran nilai prediksi yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Square Error* (RMSE), dan R2-Score.

METODE

Pada *machine learning*, penelitian ini juga dikategorikan sebagai penelitian *supervised learning* yaitu sebuah teknik model dimana data yang akan diproses memiliki label/target/*class* dengan tujuan agar mengetahui hubungan kausalitas antara variabel bebas (variabel independen) dan variabel yang menjadi target/labelnya (variabel y). Berikut variabel yang terdapat dalam penelitian ini terdiri dari variabel independen atau variabel x meliputi Pregnancies, *Glucose* (gula), *BloodPressure* (tekanan darah), *SkinThickness*, Insulin, BMI (Berat Badan), *DiabetesPedigreeFunction*, *Age* (Umur), Serta yang menjadi variabel dependennya adalah *Outcome*.

1. Tahapan Data Mining

Berikut ini Tahapan *data mining* yang digunakan penelitian ini yang terdapat pada gambar berikut ini:



Gambar 1. Flowchart Tahapan Sistem Prediksi

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

Pada Gambar 1. dapat dijelaskan bahwa tahapan pertama yang akan dilakukan adalah *Data Prepocessing* dimana dalam proses tersebut terdapat *data cleaning* untuk melakukan *handling missing* pada data, *data selection* untuk menyeleksi data berdasarkan kolom yang akan diteliti dan *data transformation.* Pada tahapan kedua yaitu melakukan *data visualization* yaitu untuk menyajikan representasi dari sebaran data yang digunakan. Tahapan berikutnya melakukan pemisahan data set menjadi 2 bagian yaitu *data training* dan *data testing*. Selanjutnya, memasuki tahapan proses prediksi dengan menggunakan regresi yang meliputi 7 model regresi dan menguji data tersebut dengan data testing. Pada tahapan yang terakhir adalah tahapan analisis dan evaluasi data dengan menggunakan 4 metode evaluasi yaitu MAE, MSE, RMSE dan R2-Score, tujuannya yaitu untuk mendapatkan hasil akurasi dari proses prediksi sehingga diperoleh metode terbaik dengan hasil evaluasi yang terbaik pula.

2. Data Riset

Data riset yang akan dipakai yaitu dataset *diabetes* yang berasal dari *kaggle.com* pada tahun 2022. Data tersebut berjumlah sebanyak 769 baris data. Terdapat 8 kolom variabel X yang akan digunakan untuk pengujian dan 1 kolom (Outcome) sebagai variabel target/label. Sampel dataset yang akan digambarkan dalam penelitian ini menggunakan 5 data teratas seperti yang terdapat dalam Tabel 1 berikut ini.

Tabel 1. Sampel Dataset

pregnancies	glucose	blood pressure	Skin thickne	insulin	bmi	Diabetes pedigree function	age	outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

3. Pembagian Data

Dalam pembagian data penelitian ini menggunakan *data training* dan *data testing* dengan rasio sebesar 90:10 Dimana *data training* memiliki kisaran data sebanyak 90 dari dataset serta *data testing* sebanyak 10 dari dataset. Pada Algoritma yang menggunakan rasio 90:10 memiliki performa yang lebih baik dibandingkan dengan algoritma lainnya dimana tingkat akurasinya dapat diperoleh sebesar 99,40% [2].

Adapun rasio 80:20, 70:30, dan 60:40 yang digunakan dalam penelitian ini sehingga rasio pembagian data lebih bervariatif dimana tujuan yang diharapkan adalah agar memperoleh model regresi manakah yang memiliki hasil prediksi terbaik.

4. Regresi Berganda

Regresi adalah suatu alat yang paling populer digunakan untuk

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

memperoleh hasil evaluasi pengatuh sebuah variabel x (variabel bebas) dan variabel y (variabel terikat). *Linear regression* atau regresi linier terbagi atas 2 yaitu *simple linear regression* dan *multiple linear regression.* Menurut penelitian bahwa regresi linear sederhana adalah model yang menganalisis hubungan antara 1 variabel prediktor dan 1 variabel respon.

Dalam penelitian tersebut juga menyatakan bahwa prediksi menggunakan regresi linear sederhana memperoleh hasil prediksi dengan kategori baik karena telah menunjukan hasil prediksi terbaik selama tahun 2021 pada periode bulan januari [3]. Adapun persamaan dalam regresi linear sederhana beriku ini :

$$Y = a + bx$$

Dimana :
Y : variabel dependen
a : titik perpotongan garis terhadap koordinat y
b : koefisien variabel independen
x : variabel independen

Kemudian *multiple linear regression* atau regresi linear berganda menurut [4] merupakan suatu model analisis yang menggambarkan hubungan variabel respon (Y) dengan variabel yang mempengaruhinya (x) dimana variabel (x) tersebut berjumlah lebih dari satu. Persamaan yang digunakan dalam regresi linear berganda adalah

$$Y = b0 + b1X1 + bnXn$$

Dimana :
Y : variabel respon
X : variabel prediktor b0 : intercept
b : koefisien

5. Support Vector Regression Linear

Merupakan salah satu model regresi yang digunakan untuk memperbaiki *overfitting* dan memiliki nilai akurasi yang baik [5]. *Overfitting* adalah kondisi dimana data saat diolah/*training* mendekati hasil prediksi hampir sempurna [6]. Persamaan yang terdapat dalam model *support vector regression* sebagai berikut ini :

$$f(x) = w^t (p(x) + b$$

Dimana :
f(x) : fungsi regresi
W^t : vektor bobot yang memiliki dimensi *l*
(p(x) : titik dalam *space* F, hasil dari pemetaan x pada input *space*
b : merupakan bias

6. RBF (Radial Basis Function)

Adapun persamaan yang terdapat dalam fungsi kernel Gaussian-RBF :

$$K(x_i, x) = \exp\left(-\frac{1}{2a^2}||x - x_i^T||^2\right) \quad (2)$$

7. Decision Tree Regression

Persamaan *decision tree regression* terlebih dahulu mencari nilai entrophy seperti dalam persamaan berikut ini :

$$entro(S) = \sum_{i=1}^{m} -p(w_i|S) \; . \; log_2(w_i|S)$$

Dimana :
S : Himpunan Kasus
M : Total kelas data
$(w_i|S)$: Proporsi kelas ke-*i* dalam semua data latih yang diproses di node S.

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

Kemudian, setelah diperoleh nilai entrophy maka selanjutnya kita mencari nilai gain nya sebagai berikut

$$Gain(S, J) = Entropy(S) - \sum_{i=j}^{n} p(v_i|S) * E(S_i)$$

Dimana :
S : Himpunan kasus
J : Fitur
n : Banyak kelas dalam *node* (akar)
$(v_i|S)$: Proporsi nilai v muncul pada kelas dalam *node* (akar)
(S_i) : Entropi komposisi nilai v dari kelas ke-j dalam data ke-i node

8. Random Forest Regressor

Menurut penelitian [7] *random forest* merupakan model algoritma pada regresi dengan tehnik mengkombinasi prediksi dari beberapa pemecahan dalam *machine learning* guna memdapatkan prediksi dengan akurasi lebih baik daripada mode tunggal. Dalam penelitian tersebut juga menyimpulkan bahwa *random forest regression* lebih akurat dibandingkan dengan *linear regression* dimana rating akurasi pada *random forest* diperoleh sebesar 97.7% dengan menggunakan perhitungan nilai RMSE dan MAPE. Persamaan *random forest regressor* seperti yang terdapat dibawah ini :

$$\hat{\gamma}_i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{\gamma}_n$$

Dimana :
$\hat{\gamma}_i$: hasil prediksi
N_{tree} : total jumlah pohon
$\hat{\gamma}_n$: hasil prediksi pohon ke-n

9. Gradient Boosting Regression

Merupakan bagian dari algoritma ensemble yang menggunakan peningkatan akurasi sebuah nilai. Model *gradient boosting regression* mampu mengatasi pola yang kompleks. Struktur data dari *gradient boosting regression* adalah pohon keputusan (*decision tree).* Adapun persamaan *gradient boosting regression* adalah sebagai berikut ini :

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \text{ for } i=1,\ldots, n$$

10. Model Pengukuran Akurasi

Data yang telah diolah menggunakan 7 model regresi kemudian diukur nilai akurasinya dengan memasukkan nilai eror. Nilai *Error* adalah selisih antara nilai pengamatan yang sebenarnya dengan nilai prediksi. Untuk prediksi yang diukur menggunakan MAE, MSE, dan RMSE nilai yang paling baik adalah nilai yang paling kecil. Berbeda dengan R2-*Score* nilai koefisien determinan yang mendekati 1 mengartikan bahwa variabel independennya memberikan semua informasi yang diperlukan dalam mengprediksi variabel dependennya [1]. nilai *error* yang akan digunakan adalah 4 model pengukuran akurasi nilai prediksi diantaranya :

***Mean Absolute Error* (MAE)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |A_i - F_i|$$

Dimana :
n : ukuran sampel
A_i : nilai data aktual ke-i
F_i : nilai data prediksi ke-i

JURNAL TEKNISI **E-ISSN: 2775-0965**
(Jurnal Teknologi Komputer dan Sistem Informasi)
Februari 2023, Vol 3, No 1, Hlm. 32 - 39
Available online at http://jurnal.goretanpena.com/index.php/teknisi

***Mean Square Error* (MSE)**

$$MSE = \frac{I}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Dimana :
n : jumlah data
Y_i : nilai yang diamati
$\hat{Y}_i$: nilai prediksi

***Root Mean Square Error* (RMSE)**

$$RMSE = \left(\frac{\sum (y_i - \hat{y}_i)}{n}\right)^{1/2}$$

Dimana :
y_i : nilai hasil pengamatan
$\hat{y}_i$: nilai hasil prediksi
n : jumlah data

7.1 **R2-*Score***

$$R^2 = 1 - \frac{RSS}{TSS}$$

Dimana :
R^2 : koefisien determinan
RSS : jumlah kuadrat
TSS : total kuadrat

HASIL DAN PEMBAHASAN

Dalam pembahasan penelitian ini juga dilakukan visualisasi data agar data dapat lebih mudah untuk diamati dan dapat mengetahui hubungan antara tiap variabel x terhadap variabel y. Setelah, data akan diuji dengan 4 bentuk pembagian data dengan rasio yang berbeda–beda dimana didalamnya akan dilakukan pengujian data menggunakan 7 model regresi dan diukur dengan 4 model nilai evaluasi. Hal ini bertujuan untuk menemukan model regresi manakah yang paling baik dengan menghasilkan nilai akurasi yang paling akurat untuk digunakan dalam prediksi hasil diagnosa diabetes sejak dini oleh seseorang.

Pembagian data untuk diuji sebanyak 4 bentuk diantaranya pembagian rasio pertama yaitu 90:10, kedua 80:20, ketiga 70:30 dan rasio keempat yaitu 60:40. Pengujian data dengan rasio yang bervariatif untuk mencari serta menentukan model regresi manakah yang paling sering muncul dalam menghasilkan nilai prediksi yang terbaik. Sehingga, nantinya akan diketahui model regresi yang terbaik dalam melakukan prediksi terhadap dataset untuk mencari *Outcome* diabetes.

Setelah melakukan pengujian akurasi dengan 4 bentuk rasio data, kemudian diperoleh metode terbaik yang paling banyak menghasilkan nilai prediksi terbaik. Metode tersebut adalah *Decision Tree Regression*. Pada metode ini memiliki tingkat akurasi yang tinggi pada dalam rasio 90:10 , 80:20, 70:30 *dan* 60:40 menjadi metode terbaik pada rasio tersebut menghasilkan akurasi yang baik yaitu sebesar 99%. Berikut ini hasil perbandingan metode terbaik dengan tingkat akurasi yang dihasilkan :

Tabel 2. Metode Regresi Terbaik

Metode Regresi Terbaik	90:10	80:20	70:30	60:40
Decision Tree Regression	99%	99%	99%	99%

KESIMPULAN

Prediksi menggunakan 7 metode regresi untuk memperoleh metode manakah yang menghasilkan nilai akurasi yang paling akurat.Dari hasil pengolahan data menggunakan 7 metode regresi menggunakan serta pengujiannya menggunakan 4 rasio pengukuran nilai akurasi, maka diperoleh metode terbaik yang menghasilkan nilai akurasi paling bagus yaitu metode *decision tree regression.* Maka metode tersebut yang akan di implementasikan dalam pembuatan *deploy* aplikasi untuk mengprediksi *Outcome* untuk mendiagnose diabetes sejak dini oleh masyarakat.

DAFTAR PUSTAKA

[1] "Memahami Koefisien Determinasi Dalam Regresi Linear – Accounting."

[2] A. I. Sang, E. Sutoyo, and I. Darmawan, "Analisis Data Mining Untuk Klasifikasi Data Kualitas Udara Dki Jakarta Menggunakan Algoritma Decision Tree Dan Support Vector Machine Data Minning Analysis For Classification OF Air Quality Data Dki Jakarta Using Decision Tree Algorthm And Support Vector," Vol. 8, No. 5, Pp. 8954–8963, 2021.

[3] M. M. Syaifulloh, "Prediksi Indeks Standar Pencemaran Udara Di Kota Surabaya Berdasarkan Konsentrasi Gas Karbon Monoksida," vol. 2, no. November, 2021.

[4] A. Prasetyo, "Prediksi Produksi Kelapa Sawit Menggunakan Metode Regresi Linier Berganda," vol. 6, no. 2, pp. 76–80, 2021.

[5] A. Aulia *et al.*, "Prediksi Harga EmasdenganMenggunakan Algoritma Support Vector Regression (Svr) danLinearRegression (LR)," vol. 8, no. 5, 2022, doi: 10.5281/zenodo.6408864.

[6] N. D. Maulana, B. D. Setiawan, and C. Dewi, "Implementasi Metode Support Vector Regression (SVR) Dalam Peramalan Penjualan Roti (Studi Kasus : Harum Bakery)," vol. 3, no. 3, pp. 2986–2995, 2019.

[7] S. Fachid and A. Triayudi, "Perbandingan Algoritma Regresi Linier dan Regresi Random Forest dalam Memprediksi Kasus Positif Covid-19," *J. Media Inform. Budidarma*, vol. 6, no. 1, pp. 68–73, 2022, doi: 10.30865/mib.v6i1.3492.