

PREDIKSI JUMLAH KASUS COVID-19 MENGGUNAKAN METODE LINEAR REGRESSION DAN POLYNOMIAL REGRESSION

Devid Sumarlie, Valentinus, Paulus Samotana Zalukhu

Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara

Jl. Letjen S. Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410

e-mail: devid.535180008@stu.untar.ac.id, valentinus.535180140@stu.untar.ac.id,

paulus.535180150@stu.untar.ac.id

ABSTRAK

Coronavirus Disease 2019 (Covid-19) adalah penyakit menular yang disebabkan oleh virus *corona* jenis baru yang disebut *SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus-2)*. Penyakit ini utamanya menyerang sistem pernapasan manusia sistem pernapasan manusia dan dapat menyebabkan gejala mulai dari ringan seperti demam dan batuk hingga berat seperti *pneumonia* dan sindrom pernapasan akut. Penularan awal terjadi dari hewan ke manusia dan menular antar manusia melalui *droplet*, kontak langsung, atau permukaan yang terkontaminasi. Pada penelitian ini, fokus utamanya adalah menggunakan algoritma *Linear Regression* dan *Polynomial Regression* untuk memprediksi jumlah kasus *Covid-19* yang terjadi di Indonesia. Data yang digunakan merupakan historikal persentase dari jumlah penyebaran kasus *Covid-19* di beberapa daerah di Indonesia dan bersifat *time series*. Hasil yang diperoleh dari penelitian ini adalah bahwa metode *Linear Regression* menjadi metode terbaik dibandingkan dengan *Polynomial Regression* karena hasil evaluasi dengan *R2-square* menunjukkan metode ini memiliki persentase 95.56% lebih besar dari metode lainnya yang digunakan yang hanya mencapai 93.76%. Selain itu, dapat disimpulkan seperti itu dikarenakan nilai dari hasil evaluasi dengan metode *Root Mean Square Error (RMSE)*, metode *Polynomial Regression* memperoleh nilai sebesar 1373.25 dan metode *Linear Regression* hanya sebesar 1043.94, sehingga performa terbaik dalam memprediksi jumlah kasus *Covid-19* adalah algoritma *Linear Regression*.

Kata kunci: Covid-19, Linear Regression, Polynomial Regression, R2-squared, RMSE.

ABSTRACT

Coronavirus Disease 2019 (Covid-19) is an infectious disease caused by a new type of coronavirus called *SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus-2)*. This disease primarily attacks the human respiratory system and can cause symptoms ranging from mild, such as fever and cough, to severe, such as *pneumonia* and acute respiratory syndrome. Initial transmission occurred from animals to humans and is transmitted between humans through *droplets*, direct contact, or contaminated surfaces. In this study, the main focus is on using *Linear Regression* and *Polynomial Regression* algorithms to predict the number of *Covid-19* cases occurring in Indonesia. The data used is historical data on the percentage of the spread of *Covid-19* cases in several regions in Indonesia and is of a *time series* nature. The results obtained from this study are that the *Linear Regression* method is the best method compared to *Polynomial Regression* because the evaluation results with *R2-square* show that this method has a percentage of 95.56% greater than the other methods used, which only reached 93.76%. In addition, it can be concluded that this is because the value of the evaluation results using the *Root Mean Square Error (RMSE)* method, the *Polynomial Regression* method obtained a value of 1373.25 and the *Linear Regression* method only 1043.94, so that the best performance in predicting the number of *Covid-19* cases is the *Linear Regression* algorithm.

Keywords: Covid-19, Linear Regression, Polynomial Regression, R2-squared, RMSE.

1. PENDAHULUAN

Coronavirus menjadi bagian dari keluarga besar virus yang menyebabkan penyakit yang terjadi pada hewan ataupun manusia. Manusia yang terjangkit virus tersebut akan menunjukkan tanda-tanda penyakit infeksi saluran pernapasan mulai dari flu sampai yang lebih serius, seperti *Middle East Respiratory Syndrome (MERS)* dan *Severe Acute Respiratory Syndrome (SARS)* atau sindrom pernapasan akut berat [1]. Tahun 2020 merupakan tahun yang berat bagi dunia ketika mendadak

terjadi wabah *Covid-19* yang awalnya muncul secara lokal di Wuhan, China, lalu menyebar dan menjadi pandemi global yang mengguncang seluruh dunia [2]. Indonesia menjadi salah satu negara yang terdampak wabah ini. Oleh karena itu, perlu tindakan segera dari pemerintah dan kesadaran penuh dari masyarakat agar angka penyebaran virus ini dapat ditekan [3].

Virus ini menyebar lewat percikan cairan dari mulut (*droplet*) yang menyebabkan infeksi pernafasan ringan seperti flu dan mampu menyebabkan infeksi pernafasan berat pada paru-paru (*pneumonia*) [4]. Pada penelitian ini, akan diprediksi jumlah kasus *Covid-19* di Indonesia dengan dua pendekatan *machine learning* yaitu metode *Linear Regression* dan *Polynomial Regression*. Dengan mengetahui jumlah kasus covid-19 di Indonesia, masyarakat diharapkan agar dapat membantu menanggulangi penyebaran kasus *Covid-19*.

2. TINJAUAN LITERATUR

Penelitian ini didukung dengan beberapa studi literatur yang relevan yang membahas mengenai permasalahan yang diangkat atau menggunakan metode yang sama. Berdasarkan hasil eksplorasi, berikut telah ditemukan beberapa penelitian yang relevan:

Penelitian dengan judul “Prediksi ITB tentang Akhir Pandemi *Corona RI*” membahas model representasi yang dapat memperkirakan jumlah kasus melalui pendekatan dari model *Logistic* dengan *Richard’s Curve* yang secara matematik, ditemukan bahwa model ini memiliki jumlah *error* yang cenderung rendah dalam tren untuk data kasus terlapor saat ini [4]. Penelitian berikutnya dengan judul “*Covid Symptom Severity using Decision Tree*” yang menyimpulkan bahwa gejala yang ditimbulkan oleh virus ini ternyata memiliki karakteristik yang sama dengan gejala batuk atau pilek biasa [5]. Penelitian relevan yang terakhir berjudul “Prediksi Kasus *Covid* di Indonesia dengan menggunakan metode *Backpropagation*” menunjukkan bahwa metode ini cukup ampuh dalam memprediksi secara akurat bagaimana penyebaran virus ini terjadi di berbagai daerah [6].

3. METODE PENELITIAN

3.1 Pengolahan Data

Dataset yang digunakan bersumber dari platform *Kaggle* (pada link berikut ini: <https://www.kaggle.com/hendratno/covid19-indonesia>) dengan jumlah sampel sebanyak 16.284 data dan 37. Data ini bersifat *time series* yang dapat dilihat pada Gambar 1 yang telah disesuaikan ke dalam format file *Comma Separated Value* (CSV).

Date	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Location Len
01/08/2020	Jawa Timur	9,3,1,5,9,5,10,-6	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
01/09/2020	Jawa Timur	0,1,23,-24,9,6,33,-30	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
01/10/2020	Jawa Timur	0,1,14,-15,9,7,47,-45	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
01/11/2020	Jawa Timur	0,3,8,-11,9,10,55,-56	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
01/12/2020	Jawa Timur	0,3,1,-4,9,13,56,-60	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
1/13/2020	Jawa Timur	0,0,0,0,9,13,56,-60	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
1/14/2020	Jawa Timur	0,0,0,0,9,13,56,-60	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
1/15/2020	Jawa Timur	0,0,0,0,9,13,56,-60	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			
1/16/2020	Jawa Timur	0,0,0,0,9,13,56,-60	Province	Jawa Timur	Jawa	UTC+07:00	29,47803,40479023,9,1.127.329.414,-7.723.345.579			

Gambar 1. Dataset Penelitian

Data tersebut kemudian akan dicek *missing value* seperti yang ditunjukkan pada gambar 2 dan dilakukan penanganan dengan fungsi *fillna*. Setelah itu, beberapa kolom yang tidak diperlukan akan dihilangkan sesuai dengan kepentingan pengolahan lebih lanjut seperti yang dapat dilihat pada gambar 3.

```
In [9]: df.isna().sum()

out[9]: Date 0
Location ISO Code 0
Location 0
New Cases 0
New Deaths 0
New Recovered 0
New Active Cases 0
Total Cases 0
Total Deaths 0
Total Recovered 0
Total Active Cases 0
Location Level 0
City or Regency 16283
Province 495
Country 0
Continent 0
Island 495
Time Zone 495
Special Status 13890
Total Regencies 0
Total Cities 469
Total Districts 0
Total Urban Villages 471
Total Rural Villages 496
Area (km2) 0
Population 0
Population Density 0
Longitude 0
Latitude 0
New Cases per Million 0
Total Cases per Million 0
New Deaths per Million 0
Total Deaths per Million 0
Case Fatality Rate 0
Case Recovered Rate 985
Growth Factor of New Cases 1939
Growth Factor of New Deaths
dtype: int64
```

Gambar 2. Cek Missing Value

date	Location	New Cases	New Deaths	New Recovered	New Active Cases	Total Cases	Total Deaths	Total Recovered	Total Active Cases	Location Level	Area (km2)	Population	Population Density	Longitude	Latitude
2020-08-08	Jawa Timur	9	3	1	5	9	3	10	-6	Province	47903	40479023	846.78	1.127.329.414	-7.723.345.579
2020-08-09	Jawa Timur	0	1	23	-24	9	6	30	-30	Province	47903	40479023	846.78	1.127.329.414	-7.723.345.579
2020-08-10	Jawa Timur	0	1	14	-15	9	7	47	-45	Province	47903	40479023	846.78	1.127.329.414	-7.723.345.579
2020-08-11	Jawa Timur	0	3	0	-11	9	10	55	-56	Province	47903	40479023	846.78	1.127.329.414	-7.723.345.579
2020-08-12	Jawa Timur	0	3	1	-4	9	13	55	-60	Province	47903	40479023	846.78	1.127.329.414	-7.723.345.579

Gambar 3. Menghapus Data yang Tidak Diperlukan

Data tersebut dibagi menjadi 2 bagian dengan 80% sebagai data latih dan 20% akan menjadi data uji secara berurutan sesuai tanggal yang dapat dilihat pada potongan kode pada Gambar 4 berikut.

```
# X untuk variabel input
# Y untuk variabel target prediksi
X = data[['New Cases', 'New Recovered', 'Total Cases', 'Total Deaths', 'Total Recovered', 'Total Active Cases']]
y = data[['New Deaths']]

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, shuffle = False)
```

Gambar 4. Pembagian Data

Fitur akhir yang digunakan pada dataset setelah dilakukan operasi *drop feature* yang tidak diperlukan dan pembagian data menjadi bentuk dataset final yang akan diteliti adalah *Date* (waktu dan tanggal kasus Covid-19), *Location* (lokasi kejadian), *New Cases*, *New Deaths*, dan *New Recovered* yang masing-masing menunjukkan jumlah kasus baru, kematian baru, dan kesembuhan baru per hari. *New Active Cases* merepresentasikan perubahan jumlah kasus aktif harian, sedangkan *Total Cases*, *Total Deaths*, *Total Recovered*, dan *Total Active Cases* menunjukkan akumulasi totalnya. Selain itu, terdapat *Location Level* yang menunjukkan tingkatan wilayah (misalnya provinsi), *Area* (km²) sebagai luas wilayah, *Population* dan *Population Density* yang menggambarkan jumlah dan kepadatan penduduk, serta *Longitude* dan *Latitude* sebagai koordinat geografis lokasi tersebut.

3.2 Algoritma

Seperti yang sudah disebutkan sebelumnya bahwa algoritma yang digunakan dalam penelitian ini adalah *Linear Regression* dan *Polynomial Regression*. *Linear Regression* dapat dinyatakan seperti dalam persamaan dibawah ini, dimana Y dan X masing-masing adalah variabel dependen dan independen, sedangkan f adalah *intercept* dengan simbol α dan β mewakili parameter regresi dan ϵ sebagai *slope* untuk kesalahan acara masing-masing.

$$Y = \alpha + \beta X + \epsilon$$

Polynomial Regression adalah jenis regresi khusus yang bekerja pada kurva linear hubungan antara nilai-nilai dependen dan nilai-nilai independen, dimana X adalah variabel bebas dan 0 adalah bias juga intersep $\theta_0, \theta_1, \dots, \theta_n$ adalah bobot atau koefisien parsial yang diberikan pada prediktor dan n adalah derajat polinomial yang persamaannya dapat didefinisikan sebagai berikut [7][8][9]:.

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 \dots + \theta_n X_n$$

3.3 Rancangan Alur Eksperimen

Pada penelitian terdapat beberapa rancangan alur eksperimen yang dilakukan dalam melakukan prediksi terhadap jumlah kasus *Covid-19* di Indonesia dengan menggunakan pendekatan *Linear Regression* dan *Polynomial Regression*. Pada Gambar 5 berikut ditunjukkan diagram sebagai bentuk alur dari eksperimen yang dilakukan.



Gambar 5. Tahapan Penelitian

3.4 Metode Evaluasi

Pada penelitian ini juga terdapat beberapa metode evaluasi yang digunakan untuk mengevaluasi hasil prediksi dari masing-masing algoritma. Oleh karena itu, untuk mengevaluasi efektivitas metode regresi linear dan polinomial regresi terhadap prediksi jumlah kasus *Covid-19* di Indonesia maka digunakan *Root Mean Square Error* relatif (RMSE) dan *Coefficient of Determination* atau yang biasa lebih dikenal dengan nilai koefisien R^2 (R^2 -squared).

MSE dan RMSE sebenarnya terikat secara monoton melalui akar kuadrat. Pengurutan model regresi yang didasarkan pada MSE akan identik dengan pengurutan model dengan RMSE. Persamaannya dapat didefinisikan sebagai berikut [10]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{n}}$$

Coefficient of Determination menggambarkan seberapa besar variasi variabel terikat dapat dijelaskan oleh variabel bebas. Pada dasarnya ditujukan untuk mengukur seberapa jauh variabel independen mampu untuk menggambarkan variabel-variabel dependen. Persamaannya dapat didefinisikan sebagai berikut [11].

$$R^2 = 1 - \left(\frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (\hat{y}_i - y_i)^2} \right)$$

4. HASIL DAN PEMBAHASAN

Fitur yang ditentukan sebagai variabel X dan Y berasal dari variabel akumulatif yaitu *Total Cases* dan *Total Deaths* untuk memprediksi jumlah total kematian saat mencapai jumlah kasus tertentu berdasarkan input yang diinginkan serta memprediksi sekaligus keduanya pada saat mencapai minggu tertentu, misalnya dalam 5 minggu ke depan akan diprediksi berapa total kasus dan total kematian yang terjadi. Format pada kolom tanggal juga diubah ke dalam bentuk *Week Number* agar dapat lebih mudah memanipulasi data pada rentang waktu seminggu dari tanggal X menuju Y , sehingga dapat diketahui hasil prediksi sesuai jumlah minggu yang ingin dituju.

Pada proses pelatihan dan pengujian dibagi dari total sebanyak 16284 sampel 80% adalah data latih dan 20 % adalah data uji. Tahap pengujian inilah yang menjadi inti dari hasil yang ingin dicari dengan metode *Linear Regression* dan *Polynomial Regression*. Didapat total akurasi sebesar 95% dengan metode evaluasi R^2 -squared serta nilai RMSE mencapai 1043.94 untuk *Linear Regression*. Sedangkan, untuk *Polynomial Regression* memperoleh akurasi sebesar 93% dengan metode R^2 -squared serta nilai RMSE sebanyak 1373.25.

Selain itu, diuji juga dengan permisalan prediksi pada saat total kasus dan total kematian mencapai angka-angka tertentu yang dapat dilihat pada Gambar 6. Misalnya, prediksi dilakukan pada saat angka kematian mencapai 40.000 dan 50.000 total kasus, Terakhir, prediksi yang dilakukan adalah dengan input total kasus dan total kematian berdasarkan beberapa minggu setelah bulan yang sekarang (dalam hal ini 5 minggu ke depan) yang dapat dilihat pada gambar 7. Dapat diketahui bahwa pada rentang waktu antara tanggal 10 – 16 September total kasus yang diprediksi mencapai 32.132 total kasus dengan sebanyak 974 total kematian, sehingga menjadi minggu dengan total kasus sekaligus total kematian terbanyak. Sedangkan, minggu dengan pencapaian total kasus dan total kasus kematian paling sedikit terjadi pada tanggal 9 – 15 Oktober dengan hanya ada 23.505 total kasus dan jumlah total kematian yang cukup besar mencapai 773.

```

prediksi kematian saat x kasus tercapai

#prediksi kematian saat 40000 kasus
x_new1 = [[40000]] #total cases
y1 = lm.predict(x_new1) #total deaths
print(y1)

[[1172.33115612]]

#prediksi kematian saat 50000 kasus
x_new2 = [[50000]] #total cases
y1 = lm.predict(x_new2) #total deaths
print(y1)

[[1442.90487558]]

```

Gambar 6. Prediksi Total Kasus dan Total Kematian

```

d = ('total predicted cases', 'total cases predicted', 'total predicted deaths', 'total deaths predicted', 'week number', 'week number', 'date')
dataset = pd.DataFrame(data=d)

```

	total predicted cases	total predicted deaths	week number	date
0	3212	914	28	10 Sept
1	2983	830	29	17 Sept-06 Sept
2	2762	868	30	25 Oct-1 Oct
3	2549	817	31	2 Oct-8 Oct
4	2305	773	32	9 Oct-15 Oct

Gambar 7. Total Prediksi Jumlah Kasus dan Jumlah Kematian dalam 5 Minggu

5. KESIMPULAN

Dari penelitian ini dapat diketahui bahwa algoritma *Linear Regression* lebih efektif dibandingkan *Polynomial Regression* dalam memprediksi jumlah kasus Covid-19 di Indonesia. Hal ini dibuktikan dengan nilai evaluasi R^2 -score sebesar 95,56% dan RMSE sebesar 1043,94 yang menunjukkan tingkat akurasi dan ketepatan prediksi yang lebih tinggi dibandingkan *Polynomial Regression* yang hanya memiliki nilai R^2 -score sebesar 93,76% dan RMSE 1373,25. Data yang digunakan bersifat time series dengan fokus pada akumulasi persentase penyebaran kasus Covid-19 di berbagai daerah.

Algoritma yang dibangun dengan menggunakan variabel Total Cases dan Total Deaths untuk memprediksi perkembangan jumlah kematian berdasarkan jumlah kasus atau minggu tertentu. Hasil pengujian menunjukkan performa model *Linear Regression* sangat baik, terutama pada prediksi lima minggu ke depan. Prediksi tersebut memperlihatkan bahwa puncak kasus dan kematian diperkirakan terjadi pada minggu 10–16 September, sedangkan nilai terendah pada minggu 9–15 Oktober, menunjukkan kemampuan model dalam menggambarkan pola peningkatan dan penurunan kasus secara akurat.

DAFTAR PUSAKA

- [1] Nasution, N. H., & Hidayah, A. (2021). Gambaran Pengetahuan Masyarakat Tentang Pencegahan Covid-19 Di Kecamatan Padangsidempuan Batunadua, Kota Padangsidempuan. *Jurnal Kesehatan Ilmiah Indonesia (Indonesian Health Scientific Journal)*, 6(1), 107-114.
- [2] Muhyiddin, M. (2020). Covid-19, New Normal, dan Perencanaan Pembangunan di Indonesia. *Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning*, 4(2), 240-252.
- [3] Wahidah, I., Athallah, R., Hartono, N. F. S., Rafqie, M. C. A., & Septiadi, M. A. (2020). Pandemi COVID-19: Analisis Perencanaan Pemerintah dan Masyarakat dalam Berbagai Upaya Pencegahan. *Jurnal Manajemen dan Organisasi*, 11(3), 179-188.
- [4] ITB Prediksi Akhir Pandemi Corona RI Bergeser Hingga Lebaran. Diakses pada 16 November 2021, dari <https://www.cnnindonesia.com/teknologi/20200326144636-185-487127/itb-prediksi-akhir-pandemi-corona-ri-bergeser-hingga-lebaran>.
- [5] Rochmawati, N., Hidayati, H. B., Yamasari, Y., Yustanti, W., Rakhmawati, L., Tjahyaningtijas, H. P., & Anistyasari, Y. (2020, October). Covid Symptom Severity Using Decision Tree. In *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)* (pp. 1-5). IEEE.
- [6] SABIQ, H., Palupi Rini, D., & Arsalan, O. (2021). Prediksi Kasus Positif Covid-19 Di Indonesia Dengan Menggunakan Metode Backpropagation (Doctoral dissertation, Sriwijaya University).
- [7] Hendratno, (2021). Kaggle, COVID-19 Indonesia Dataset. Diakses pada 19 Oktober 2021, dari <https://www.kaggle.com/hendratno/covid19-indonesia/metadata>.
- [8] Rath, S., Tripathy, A., & Tripathy, A. R. (2020). *Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model*. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5), 1467-1474.
- [9] Neve, D., Patel, H., & Dhiman, H. S. (2020). On Modeling of COVID-19 for the Indian Subcontinent using Polynomial and Supervised Learning Regression. *medRxiv*.
- [10] Chicco, D., Warrens, M. J., & Jurman, G. (2021). *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation*. *PeerJ Computer Science*, 7, e623.
- [11] Achmad, I. N., & Witastuti, R. S. (2018). *Underpricing, Institutional Ownership and Liquidity Stock of IPO Companies in Indonesia*. *Management Analysis Journal*, 7(3), 280-291.