Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 6 No. 1 (2025) 9 – 16 | ISSN: 2723-1453 (Media Online)

Konsistensi Model Regresi Empat Variabel Pada Populasi dan Sampel untuk Prediksi Temperatur

Nurjannah Syakrani[1], Naufal Athaya S. R [2]
[1]Jurusan Teknik Komputer dan Informatika, Politeknik Negeri Bandung, Bandung, Indonesia
[1]nurjannahsy@jtk.polban.ac.id, [2]naufal.athaya.tif415@polban.ac.id

Abstract

The ability to predict future events or trends has become very important today. One method that can be used to predict the future is to use linear regression. Accurate regression modeling requires sampling representative data, especially when working with large datasets. This research takes a relatively large volume as a data set by looking at the accuracy and consistency of the coefficients of a multi-variable linear regression model for temperature prediction which is built based on all the data, and looks at the differences in the regression model built from the sample data. The number of sample data (n) is determined based on the Slovin formula which depends on the number of population data (N) and the level of confidence (σ), so that as many as (N/n) new regression models can be built. Each group of sample data is divided into 75% for modeling and 25% testing data. The dataset used is weather information in the Szeged area which was measured in 2006 - 2016. So the regression model is in the form of Y (temperature value) which is influenced by Xi (weather factors), namely humidity, wind speed, wind direction and visibility. Using 96,453 data records and a 1% significance level in Slovin's formula, 10 samples were generated. Nine out of ten sample regression models agree with the population model, with positive coefficients for visibility and wind direction and negative values for humidity and wind speed. There is an abnormality in sample #4. While the other nine sample regression models are consistent with positive R^2 values, Sample #1 displays an oddity with negative values. The RMSE interval values for each regression model in this study fall between 4.334 and 9.582.

Keywords: *Multi Variable Linear Regression, Population, Sample, Slovin, RMSE*

Abstrak

Kemampuan untuk memprediksi peristiwa atau tren masa depan menjadi sangat penting pada masa kini. Salah satu metode yang dapat digunakan untuk memprediksi masa dapan adalah dengan menggunakan regresi linear. Pemodelan regresi yang akurat memerlukan pengambilan sampel data yang representatif, terutama ketika bekerja dengan *dataset* yang berukuran besar. Penelitian ini mengambil volumenya yang relative besar sebagai data set dengan mencermati akurasi dan konsistensi koefisien model regresi linear banyak variabel untuk prediksi temperatur yang dibangun berdasarkan seluruh data, dan melihat perbedaan terhadap model regresi yang dibangun dari data sampelnya. Banyak data sampel (n) dihitung berdasarkan rumus Slovin yang bergantung pada banyak data populasi (N) dan tingkat kepercayaan (σ), sehingga bisa dibangun sebanyak (N/n) model regresi baru. Setiap kelompok data sampel dibagi atas 75% data untuk pembuatan model dan 25% data pengujian. *Dataset* yang digunakan berupa informasi cuaca pada daerah Szeged yang diukur pada tahun 2006 - 2016. Sehingga model regresinya berupa Y (nilai temperatur) yang dipengaruhi Xi (faktor-faktor cuaca) yaitu kelembaban, kecepatan angin, arah angin dan jarak pandang. Dari 96.453 record data dengan taraf signifikan 1% menurut formula Slovin dihasilkan 10 sampel. Sembilan dari sepuluh model regresi sampel konsisten terhadap model populasi dengan koefisien negatif pada kelembapan dan kecepatan angin, serta koefisien positif pada arah angin dan jarak pandang. Anomali terjadi pada sampel ke #4. Sembilan dari sepuluh model regresi sampel konsisten dengan nilai R^2 yang positif, anomali terjadi pada sampel #1 dengan nilai R^2 negatif. Nilai interval RMSE dari semua model regresi pada penelitian ini adalah 4,334 sampai 9,582

Kata kunci: Regresi Linear Banyak Variabel, Populasi, Sampel, Slovin, RMSE

1. Pendahuluan

Kemampuan untuk memprediksi peristiwa atau tren di masa depan menjadi semakin penting pada masa kini. Analisis prediktif memainkan peran yang signifikan di berbagai bidang, mulai dari bisnis, ekonomi, cuaca, hingga kesehatan. Prediksi atau ramalan atau prakiraan, adalah hasil dari kegiatan memprediksi atau meramal atau memperkirakan nilai pada masa yang akan datang dengan menggunakan data masa lalu [1]. Salah satu prediksi yang dapat berguna dalam berbagai bidang adalah prediksi suhu atau temperatur. Prediksi mengenai temperatur bermanfaat mulai dari pertanian, energi, hingga kesehatan. Kemampuan untuk memprediksi temperatur menjadi sangat penting, contohnya pada bidang pertanian. Informasi tentang temperatur atau cuaca yang tepat dapat membantu petani untuk

merencanakan waktu tanam dan panen, yang sejalan dengan meningkatkan hasil dari pertanian tersebut [2]. Berdasarkan daftar istilah BMKG [3], cuaca adalah keadaan udara pada saat tertentu di wilayah tertentu yang relatif tidak luas pada jangka waktu yang singkat dan menyatakan keadaan yang berlangsung pada saat atau selama waktu kesatuan, dalam hal ini direpresentasikan dengan temperature (Celsius). Kelembaban udara didapatkan dari perhitungan suhu udara yang diamati dari *thermometer* bola basah dan bola kering dengan satuan " % ". Angin memiliki dua parameter pengukuran, yaitu arah angin dan kecepatan angin. Arah angin merupakan arah dari mana datangnya angin. Standar penentuan arah angin adalah dengan menggunakan satuan derajat melingkar dari 0 hingga 360, sedangkan kecepatan angin menggunakan satuan knot atau km/jam dengan kesetaraan 1 knot = 1,852 km/jam [3]. Sedangkan *visibility* atau jarak pandang adalah kemampuan melihat jarak horizontal terjauh dimana sebuah objek yang jelas dapat terlihat dengan mata telanjang dan diungkapkan dalam satuan jarak [4]. Berdasarkan peraturan Menteri Perhubungan KM 18 tahun 2010 jarak pandang aman dalam penerbangan adalah lima kilometer ke atas [5], jarak pandang kurang dari 5 KM berdampak penundaan penerbangan (*delay*). Keterkaitan unsur cuaca ini penting untuk dicermati terutama bagi penerbangan, maritim, dan pertanian.

Salah satu metode untuk melakukan prediksi adalah dengan menggunakan model regresi. Model regresi merupakan salah satu metode statistika yang umum digunakan untuk prediksi tentang karakteristik kualitas maupun kuantitas. Terdapat dua tipe regresi yang paling esensial yaitu regresi linear dan regresi non-linear. Regresi linear ketika parameter dari suatu model mempunyai ketergantungan (*dependent*) terhadap variabel lainnya secara linear [6]. Pentingnya model regresi linear terletak pada kemampuannya untuk memberikan prediksi yang jelas dan interpretasi yang mudah dipahami dari data, terutama ketika ingin memahami bagaimana satu atau lebih variabel independen (bebas) memengaruhi variabel dependen (terikat).

Salah satu tantangan utama dalam pemodelan regresi adalah ketika ukuran populasi data sangat besar, mengolah seluruh data bisa menjadi sangat tidak efisien dan memakan banyak sumber daya, baik dari segi waktu maupun kapasitas komputasi. Selain itu, dalam situasi tertentu, data populasi mungkin tidak selalu tersedia atau sulit untuk diakses maupun dicatat. Oleh karena itu, pengambilan sampel yang representatif menjadi solusi yang efektif. Dengan menggunakan sampel, waktu dan beban komputasi dapat dikurangi analisis dilakukan lebih cepat tanpa mengorbankan validitas hasil. Melalui sampel, masih dapat diuji apakah model yang dihasilkan cukup akurat dan robust. Perbandingan antara model yang dibangun dari sampel dan seluruh populasi menjadi penting untuk memastikan bahwa hasil yang diperoleh tidak bergantung hanya pada ukuran data, tetapi juga pada kualitas pemilihan sampel dan keakuratan prediksi yang dihasilkan dari model tersebut.

Dari pemaparan beberapa masalah diatas memunculkan pertanyaan, seperti apakah data besar harus diolah seluruhnya untuk mendapatkan gambaran secara kuantitatif? atau cukup dilakukan dengan sampel saja, Bagaimana dengan akurasi model populasi terhadap sampel? Apakah terdapat perbedaan yang signifikan antara model populasi dan model sampelnya? Apakah pengolan sampel yang efisien sudah merepresentasikan model dari seluruh data? Apakah model yang dibangun dari sampel bisa memberikan hasil prediksi yang konsisten?

Beberapa penelitian terdahulu menggunakan model regresi linear banyak variabel untuk memprediksi temperatur. Diantaranya adalah penelitian yang dilakukan oleh Karna dan kawan-kawan [7]. Penelitian ini melakukan prediksi temperatur dengan menggunakan linear regresi pada *time series* data. Penelitian ini menggunakan *mean absolute difference* (MAD) dalam melakukan analisis, rata-rata dan standar deviasi digunakan untuk menganalisis korelasi antar variabel. Tahir dan kawan-kawan [8] melakukan penelitian mengenai prediksi temperatur harian di kota Karachi. Penelitian ini menggunakan dua model regresi (linear dan non-linear) dengan menggunakan data kelembaban dan titik embun. Regresi linear banyak variabel juga digunakan oleh Ardytha Luthfiarta dan kawan-kawan [9] untuk prediksi cuaca dalam hal ini curah hujan berdasarkan suhu, kelembaban, tekanan udara, dan kecepatan angin dan menggunakan metrik nilai koefisien determinasi (R^2). Evita Fitri, menggunakan perbandingan Metode Regresi Linier, *Random Forest Regression* dan *Gradient Boosted Trees Regression Method* untuk prediksi harga [10]. Sebagai pendukung perhitungan banyak sampel dari populasi Zhicheng Liu, Aoqian Zhang [11] mendeskripsikan beberapa teknik *sampling* dan *profiling* terhadap *Big Data*, juga Ganesh [12] yang menguraikan beberapa cara perhitungan *sampling* data kuantitatif. Penelitian yang dilakukan oleh Gupta dan kawan-kawan [13] menggunakan metode *multiple linear regression* dengan *mean absolute error* sebagai metode evaluasi dari model. Penelitian yang dilakukan oleh Tedja dan kawan-kawan [14] menggunakan metode yang sama dengan tambahan parameter level polutan. Pada penelitian lain, yang dilakukan oleh Ahmed dan kawan-kawan [15] menggunakan metode *multiple linear regression* dengan dipadukan *artifical neural network* (ANN) untuk memprediksi curah hujan. Parameter yang digunakan diantaranya adalah suhu, kecepatan angin, dan titik embun. Hasil penelitian menunjukan bahwa distribusi data pada saat *training* dan *testing* sangat berpengaruh.

Penelitian ini mencoba menjawab pertanyaan yang telah didefiniskan sebelumnya melalu perbandingan hasil model setiap kelompok sampel terhadap keseluruhan

DOI : https://doi.org/10.52158/jacost.v6i1.971

data atau populasi dengan menggunakan model prediksi regresi linear banyak variabel.

2. Metode Penelitian

Penelitian ini dilakukan dengan koleksi data, normalisasi data (*pre- processing*) dan perhitungan banyak sampel, membuat model regresi dan menghitung nilai akurasi serta analisis hasil. Terdapat lima formula pendukung untuk perhitungan menggunakan data set, yang diantaranya sebagai berikut.

2.1 Persamaan Linear Banyak Variabel

Model yang digunakan pada penelitian ini mengikuti persamaan 1a-1b [2][4].

$$Y = A + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + \varepsilon \quad (1a)$$

Diringkas

$$Y = \mathrm{A} + \sum_{i=1}^{k} B_i X_i + \varepsilon \quad (1b)$$

Dengan Yadalah variabel terikat (*dependent*, akibat), A adalah konstanta atau nilai titik potong dengan sumbu Y pada saat setiap Xi bernilai nol, Xi adalah variabel bebas (independent, explanatory, faktor penyebab), Bi adalah koefisien regresi untuk variabel bebas yg menunjukkan kenaikan (+) atau penurunan (-) nilai Y, ketika Xi berubah satu satuan dan ε adalah *factor error* (ketidaktepatan model).

Sebagai contoh [16], pengolahan dua belas himpunan data pada Tabel 1, dengan variabel terikat Y harga apartemen dalam ($1000), variabel bebas X_1 jarak (km) masing-masing apartemen dari pusat kota, X_2 ukuran (luas) apartemen dalam *feet*^{2}, menjadi persamaan linear banyak variable.

Tabel 1. Data harga, jarak, dan luas apartemen.

Y ($1000)	X_1 (km)	X_2 (*feet*^{2})
55	1.5	350
51	3	450
60	1.75	300
75	1	450
55.5	3.1	385
49	1.6	210
65	2.3	380
61.5	2	600
55	4	450
45	5	325
75	0.65	424
65	2	285

Diolah dengan bantuan Excel [11], menghasilkan

$$Yp = 60.041 - 5.393 * X1 + .03 * X2 \quad (2)$$

Persamaan (2) menunjukkan harga (Y) akan berkurang (lebih murah) sebesar 5.393 x 1000 dolar untuk setiap penambahan 1 km jarak apartemen dari pusat kota dan harga akan bertambah sebesar 0,03 x 1000 dolar untuk setiap penambahan luas area 1 *feet*^{2}.

2.2 Normalisasi Data

Normalisasi data adalah tahap penting dalam melakukan sebuah pemrosesan data (pre-processing), tahap ini dilakukan untuk menstandarisasi rentang nilai pada atribut dalam *dataset*. Tujuan utama dari normalisasi ini adalah untuk memastikan bahwa semua atribut memiliki skala yang seragam, sehingga model analisis data, seperti regresi atau klasifikasi, tidak terpengaruh secara tidak proporsional oleh atribut dengan rentang nilai yang lebih besar [17].

Salah satu metode normalisasi yang umum digunakan adalah Z-Score Normalization. Metode ini melakukan transformasi pada data dengan menghitung nilai z-score untuk setiap data yaitu mengurangi rata-rata data dan membaginya dengan standar deviasi data tersebut. Z-score adalah ukuran standar deviasi dari data, yang menggambarkan seberapa banyak deviasi data dari rata-rata data tersebut. Sesuai hasil riset Dimas dan Wiwit [18] Z-Score Normalization mampu meningkatkan akurasi model regresi dibandingkan data original yang diukur dengan metrik R^2 dan MSE. Formula Z-Score Normalization dinyatakan dengan persamaan 3.

$$Z = \frac{X - \mu}{\sigma} \quad (3)$$

Dengan X adalah Nilai asli dari data, μ adalah Rata-rata dari data (*mean*), dan σ adalah Standar deviasi dari data.

2.3 Ukuran Sampel

Ukuran sampel ditentukan dengan menggunakan formula Slovin [11], [12] dengan persamaan 4.

$$n = \frac{N}{1 + N\,\epsilon^2} \quad (4)$$

Dengan n adalah ukuran sampel, N adalah banyak populasi, dan ϵ adalah margin error atau (1- taraf signifikansi).

Sebagai contoh $N = 5000$, dengan taraf signifikansi $\alpha = 0{,}97$ atau margin error $e = 3\%$ maka ukuran sampelnya dapat didapatkan sebagai berikut.

$$n = \frac{5000}{1 + 5000\,(0.03)^2} = 903$$

Ini berarti banyak sampel tanpa irisan k = 5000/903. Dengan besar sampel yang sama $N = 5000$ dan margin error dengan nilai 5%, 7%, dan 10% maka ukuran sampel berturut-turut 370, 176, dan 98.

2.4 Evaluasi Keakuratan

Evaluasi akurasi metode prediksi pada riset ini menggunakan koefisien determinasi atau R^2 dan *root mean square error* (RMSE) serta RSR, rasio RMSE dengan standar deviasi.

Menurut Chicco, Warrens dan Jurman [19], koefisien determinasi dapat diartikan sebagai proporsi varians

DOI : https://doi.org/10.52158/jacost.v6i1.971

variabel terikat *(dependen*) yang dapat diprediksi dari variabel bebas *(independen*). Hasil riset mereka menunjukkan koefisien determinasi *R-square* (R^2) lebih informatif dibanding SMAPE, MAE, MAPE, MSE dan RMSE dalam evaluasi analisis regresi. Formula R^2 memiliki nilai terburuk (*worst value*) = -∞ dan nilai terbaik (*best value*) = +1 [19], Formula R^2 dinyatakan pada persamaan 5.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2} \quad (5)$$

Dengan, 1 dikurang dari rasio jumlah kuadrat selisih antara nilai aktual dan prediksi dengan jumlah kuadrat dari selisih antara nilai aktual dan rata-rata nilai aktual. Koeifisien determinasi R^2 menunjukan seberapa baik model regresi menjelaskan variasi dalam data. Nilai R^2 menunjukkan semakin baik jika mendekati 1 [19].

Root mean square error (RMSE) yaitu akar dari rata-rata selisih nilai aktual dan nilai prediksi. Performa dan prediksi model meningkat seiring dengan menurunnya nilai RMSE, dan RMSE yang lebih besar menunjukkan penyimpangan yang signifikan dari data aktual dan prediksi [20], nilai terbaik (*best value*) = 0 dan nilai terburuk (*worst value*) = +∞ [19]. Formula RMSE sebagai berikut.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{M}(y_t - \hat{y}_t)^2}{M}} \quad (6)$$

Dengan
Madalah banyak data, y_t adalah nilai aktual ke t, dan $\hat{y}_t$ adalah nilai prediksi ke t.

Sebagai contoh, berdasarkan Tabel 1 dengan menerapkan perhitungan formula (2) untuk Y taksiran atau Yp dan formula (4) untuk RMSE, diperoleh hasil pada Tabel 2.

Tabel 2. Data harga real, taksiran harga, dan delta kuadrat.

Y ($1000)	Yp (2)	$(Y - Yp)$^2
55	62,45	55,52
51	57,36	40,48
60	59,60	0,16
75	68,15	46,95
55.5	54,87	0,39
49	57,71	75,90
65	59,04	35,56
61.5	67,26	33,12
55	62,45	55,52
45	57,36	40,48
75	59,60	0,16
65	68,15	46,95

Diperoleh nilai dari RMSE sebesar

$$RMSE = \sqrt{\frac{386{,}76}{12}} = 5{,}68\,.$$

Efektifitas model yang digunakan dapat didemonstrasikan menggunakan empat indeks statistic yaitu *the coefficient of determination* (R2, koefisien determinasi), *the root mean square error* (RMSE, akar rata-rata error kuadrat), *the mean absolute error* (MAE rata-rata eror absolut), *the RMSE-Standard Deviation Ratio* (RSR, rasio RMSE dengan standar deviasi) [21].

Standard-deviation Ratio (RSR) digunakan unuk menstandarkan *Root Mean Square Error* (RMSE) yaitu dengan membagi RMSE terhadap standar deviasi. RSR seperti bernilai positif dan semakin kecil menuju nol (nilai ideal) untuk menunjukkan perfomansi simulasi atau prediksi model semakin baik [22]. RSR dikatagorikan menjadi empat peringkat performansi sebagaimana Tabel 3.

Tabel 3. Rentang Nilai RSR dan Kategori Performansi.

Rentang Nilai	Kategori
(0.00 < RSR ≤ 0.50)	*Very Good*/ Sangat Baik
(0.50 < RSR ≤ 0.60)	*Good*/ Baik
(0.60 < RSR ≤ 0.70)	*Satisfactory*/ Cukup
(RSR > 0.70)	*Unsatisfactory*/ Kurang

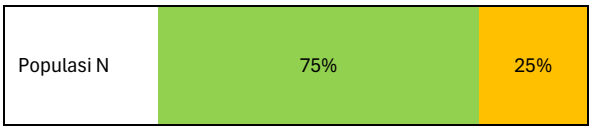
Sehingga pada contoh di atas, dengan RMSE bernilai 5,68 dan standar deviasi Y sebesar 9,51 maka RSR = 0.597. Ini berarti model regresi (2) berkategori baik.

2.5 Tahapan Penelitian

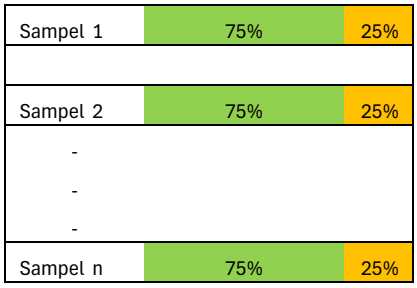
Penelitian ini dilaksanakan secara bertahap, sebagai berikut. (1). Kumpulan data (*dataset*) yang digunakan untuk studi kasus prediksi temperatur adalah data sekunder yaitu cuaca daerah Szeged, Hungaria dari Kagel. *Dataset* tersebut berupa informasi cuaca yang diukur pada tahun 2006 – 2016, sebanyak 96.453 *record* (himpunan data) dalam file format csv [23]. (2). Dari *dataset* yang sudah didapatkan pada poin 1 terdapat dua belas variabel yang tersedia, namun dalam penelitian ini dipilih 4 variabel yang berpengaruh yaitu temperatur (Y) sebagai variabel terikat dan empat variabel bebas yang berkaitan dengan faktor cuaca yaitu X_1 sebagai variabel Kelembaban (%), X_2 sebagai variabel kecepatan angin (km/jam), X_3 sebagai variabel arah angin (derajat), dan X_4 sebagai variabel Jarak pandang (km). (3). Keempat variabel bebas tersebut akan dinormalisasi menggunakan metode *Z-Score Normalization*. Metode ini mengubah data menjadi skala dengan rata-rata 0 dan deviasi standar 1, memungkinkan perbandingan yang lebih baik antar variabel. (4). Setelah pemilihan fitur dan normalisasi variabel bebas dilakukan, selanjutnya dihitung jumlah sampel n yang diperlukan dari populasi data menggunakan formulasi Slovin untuk menentukan ukuran sampel dengan margin kesalahan 1%. (5). Dari perhitungan pada poin 4, ditentukan jumlah sampel (k) yang akan digunakan dalam analisis, (6). Setiap sampel maupun populasi akan dipilah menjadi 2 bagian yaitu 75% untuk pembuatan model (*training*) dan 25% untuk pengujian akurasi (*testing*) sesuai dengan Gambar 1. Dalam hal ini tidak ada irisan antara sampel yang satu dengan yang lain dan tidak ada irisan antara data untuk pembuatan model maupun untuk pengujian baik pada data sampel maupun populasi sesuai dengan Gambar 2,

DOI : https://doi.org/10.52158/jacost.v6i1.971

(7). Perhitungan regresi linear banyak variabel untuk 75% data populasi sebagai persamaan utama dan untuk masing-masing 75% data sampel akan terbentuk sebanyak k persamaan regresi sampel. (8). Setiap model regresi linear banyak variabel yang diperoleh dilakukan uji terhadap 25% data dan digunakan untuk perhitungan RMSE seperti pada Tabel 2, dan (9). Analisis dilakukan dengan membandingkan setiap koefisien X_1, X_2, X_3, dan X_4 model regresi linear banyak variabel masing-masing n sampel terhadap koefisien bersesuaian di model populasi. Juga memperhatikan nilai RMSE masing-masing.



Gambar 1. Pemilahan untuk data populasi



Gambar 2. Pemilahan untuk data sampel

Langkah-langkah diatas dapat digambarkan sesuai diagram alir pada Gambar 3.

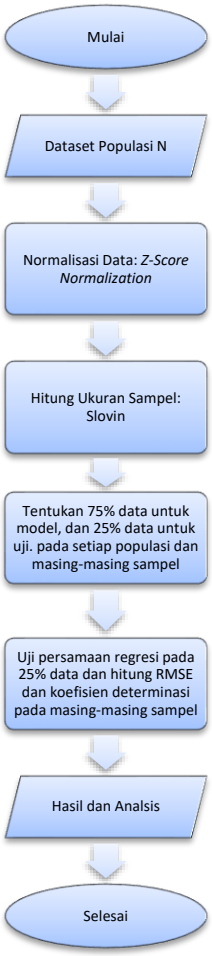
3. Hasil dan Pembahasan

Sesuai dengan langkah yang telah diberikan pada metode diatas akan dihitung persamaan regresi linear untuk 1 variabel terikat dan 4 variabel bebas. Tabel 3 berisi 10 contoh sampel data yang diambil pada *dataset*.

Tabel 4. Sampel dataset.

Temp-eratur (Y)	Kelemb-aban (X_1)	Kecepatan Angin (X_2)	Arah Angin (X_3)	Jarak Pandang (X_4)
9.47	0.89	14.11	251.0	15.82
9.35	0.86	14.26	259.0	15.82
9.37	0.89	3.92	204.0	14.95
8.28	0.83	14.10	269.0	15.82
8.75	0.83	11.04	259.0	15.82
9.22	0.85	13.95	258.0	14.95
7.73	0.95	12.36	259.0	9.98
8.77	0.89	14.15	260.0	9.98
10.82	0.82	11.31	259.0	9.98
13.77	0.72	12.52	279.0	9.98
9.47	0.89	14.11	251.0	15.82
9.35	0.86	14.26	259.0	15.82

Total data yang terdapat pada *dataset*, berjumlah 96.453 set. Dengan margin error 1% dihitung ukuran sampel (n) dan banyak sampel yang akan dibuat model regresi *k* sebagai berikut.



Gambar 3. Diagram Alir proses populasi dan sampel

$$n = \frac{96453}{1 + 96453\,(0.01)^2} = 9061$$

$$k = \frac{96453}{9061} = 10{,}64$$

Berdasarkan hasil perhitungan diatas, sampel k dibulatkan menjadi 10 sampel terurut yang tidak saling beririsan dengan setiap sampel terdapat kurang lebih ± 9000 data. Dari pemilihan sampel tersebut, maka didapat hasil perhitungan 10 persamaan regresi dari masing-masing sampel dan RMSE tercantum pada Tabel 4.

Berdasarkan sepuluh persamaan regresi sampel dan persamaan regresi populasi yang disajikan pada Tabel 4, terdapat perbedaan dalam nilai koefisien variabel bebas di antara sampel-sampel tersebut, serta dibandingkan dengan koefisien pada model populasi. Hasil menunjukkan koefisien untuk masing-masing variabel bervariasi dan bahwa model regresi untuk setiap sampel memiliki sedikit perbedaan dalam pengaruh variabel-variabel independen terhadap variabel dependen. Perbedaan ini dapat mencerminkan variasi data di antara sampel serta perbedaan antara sampel dan populasi secara keseluruhan.

DOI : https://doi.org/10.52158/jacost.v6i1.971

Tabel 5. Persamaan Regresi untuk 10 sampel.

No	Persamaan Regresi	RMSE
1	$Y = 12.02 - 5.98\,X1 - 1.94\,X2 + 0.416\,X3 + 2.62\,X4$	6.801
2	$Y = 10.713 - 7.49\,X1 - 1.403\,X2 + 0.173\,X3 + 1.247\,X4$	5.016
3	$Y = 10.839 - 6.087\,X1 - 1.611\,X2 + 0.127\,X3 + 1.47\,X4$	7.32
4	$Y = 12.65 - 5.834\,X1 - 1.382\,X2 - 0.014\,X3 + 1.502\,X4$	9.129
5	$Y = 14.648 - 4.597\,X1 - 1.624\,X2 + 0.149\,X3 + 1.345\,X4$	9.419
6	$Y = 10.438 - 6.052\,X1 - 1.051\,X2 + 0.54\,X3 + 2.847\,X4$	9.582
7	$Y = 10.865 - 5.411\,X1 - 0.725\,X2 + 0.137\,X3 + 1.309\,X4$	8.548
8	$Y = 12.417 - 5.215\,X1 - 1.869\,X2 + 0.665\,X3 + 1.21\,X4$	5.366
9	$Y = 12.301 - 4.821\,X1 - 1.871\,X2 + 0.095\,X3 + 1.922\,X4$	5.896
10	$Y = 12.939 - 5.758\,X1 - 1.616\,X2 + 0.493\,X3 + 1.563\,X4$	4.334
Populasi	$Y = 11.827 - 5.883\,X1 - 1.426\,X2 + 0.35\,X3 + 1.949\,X4$	6,743

Tabel 6. Nilai Koefisien

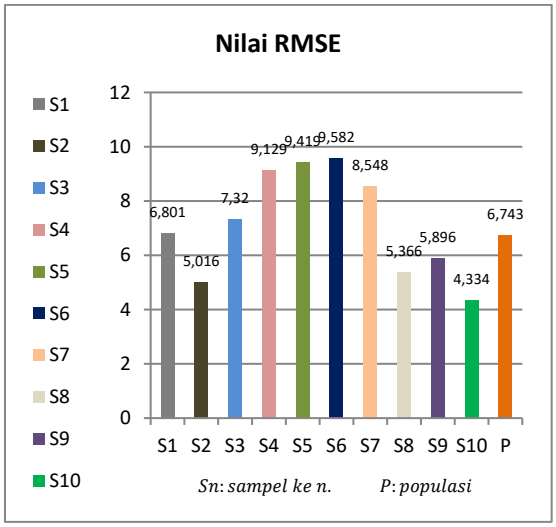
Koefisien	Y	X_1	X_2	X_3	X_4
Min10-Sampel	10.438	-7.49	-1.94	-0.014	1.21
Max10-Sampel	14.648	-4.597	-0.725	0.665	2.847
Populasi	11.827	-5.883	-1.426	0.35	1.949

Berdasarkan nilai minimum dan maksimum koefisien variabel regresi pada Tabel 5, koefisien-koefisien dari persamaan regresi untuk masing-masing sampel tidak menunjukkan perbedaan yang signifikan. Rentang nilai untuk koefisien peramaan regresi sampel, bernilai minimum sampai maksimum relatif berdekatan, dan semua rentangnya memuat nilai koefisien populasi.

Perbandingan nilai RMSE bisa dilihat pada Gambar 4. Koefisien-koefisien dari persamaan regresi untuk masing-masing sampel tidak menunjukkan perbedaan yang signifikan walaupun delta antar RMSE lebih besar.

Rentang nilai RMSE dari sampel 4.334 hingga 9.582 menunjukkan variasi yang cukup besar. Ada beberapa sampel dengan RMSE yang lebih rendah dibandingkan RMSE populasi yaitu 4.334 dan 5.016\, dan ada juga beberapa sampel dengan RMSE yang lebih tinggi dibandingkan dengan RMSE populasi yaitu 9.129, 9.419, dan 9.582. Secara keseluruhan, ini menunjukkan bahwa model regresi pada beberapa sampel memiliki akurasi prediksi yang lebih baik dibandingkan dengan model regresi populasi, sedangkan yang lainnya memiliki akurasi yang lebih rendah.

Pada Gambar 5 dan Gambar 6 dapat dilihat hasil perbandingan prediksi dengan nilai R^2 terendah yang terdapat pada sampel 1 dan tertinggi pada sampel 10.



Gambar 4. Perbandingan nilai RMSE setiap sampel dan populasi

Untuk mengetahui efektivitas model pada data set sepuluh sampel dibandingkan dengan model populasi, ditunjukkan R^2, RMSE, dan RSR pada Tabel 6. Khususnya nilai R^2 berada disekitar nol.

Tabel 7. Nilai RMSE, R^2, RSR sampel dan populasi

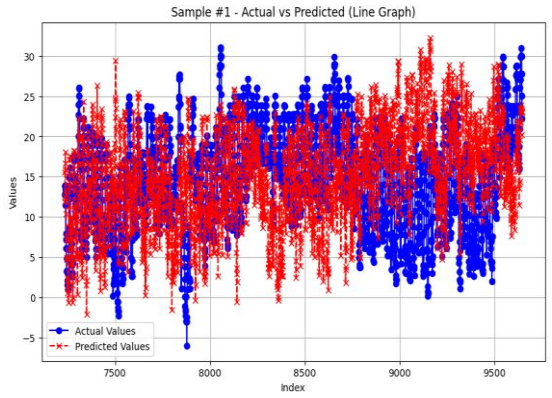
No	RMSE	R^2	RSR
1	6.801	-0.13	1.063
2	5.016	0.562	0.662
3	7.32	0.548	0.672
4	9.129	0.04	0.98
5	9.419	0.263	0.859
6	9.582	0.405	0.771
7	8.548	0.209	0.89
8	5.366	0.36	0.8
9	5.896	0.196	0.896
10	4.334	0.628	0.61
Populasi	6,743	0.427	0.82

Nilai RMSE dan R^2 dari 10 sampel dan populasi menunjukkan bahwa beberapa model regresi pada sampel (#2, #8, #9, #10) memiliki akurasi prediksi yang lebih baik dibandingkan dengan model populasi, meskipun variasi data yang berbeda antara sampel mengindikasikan bahwa tidak semua model dapat diandalkan secara konsisten. Sampel #10 dengan RMSE terendah (4.334) dan R-squared tertinggi (0.628) serta RSR terendah (0,61) adalah contoh model regresi yang sangat baik dalam menjelaskan variasi data dan memberikan prediksi yang akurat. Sementara itu, Sampel #1 dengan RMSE 6.801 (bukan nilai terbesar) tetapi R^2 (-0.13) serta RSR lebih dari satu, menunjukan model tersebut sangat kurang baik diantara sepuluh sampel. Jika merujuk pada Tabel 3 untuk kategori nilai RSR, model #2, model #3, dan model #10 merupakan model yang dikategorikan cukup representatid.

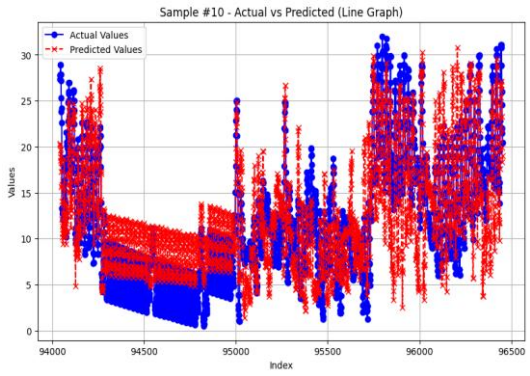
Secara khusus untuk nilai R^2 yang bernilai negatif pada sampel #1, kasus ini hanya mungkin terjadi pada regresi linier ketika titik potong atau kemiringan garis regresi dibatasi sedemikian rupa sehingga garis "*best-fit*" lebih

DOI : https://doi.org/10.52158/jacost.v6i1.971

buruk daripada garis horizontal [19]. Misalnya, jika garis regresi (*hyperplane*) tidak mengikuti data. Terakhir, nilai R^2 negatif juga bisa terjadi ketika konstanta dihilangkan dari persamaan, yaitu memaksa garis regresi melewati titik (0,0) [24].



Gambar 5. Hasil prediksi pada sampel #1



Gambar 6. Hasil prediksi pada sampel #10

4. Kesimpulan

Studi kasus ini mengkaji model regresi linear banyak variabel (*linear multiple regression*) yang digunakan untuk memprediksi temperatur dengan variabel terikat berupa suhu dan variabel bebas meliputi kelembaban (%), kecepatan angin (km/jam), arah angin (derajat), dan jarak pandang (km) menggunakan data cuaca di Szeged, Hungaria, selama periode 2006–2016. Dari populasi data sebesar 96.453, dilakukan pemilihan sepuluh sampel yang diurutkan tanpa beririsan, masing-masing sampel berukuran sekitar ±9,061, dengan proporsi 75% digunakan untuk *training* atau pembuatan model regresi linear berganda dan 25% untuk data uji akurasi. Dengan terlebih dahulu menormalisasi data set , hasil pemodelan menunjukkan bahwa semua persamaan regresi sampel dibandingkan regresi populasi memiliki koefisien yang konsisten dengan tanda positif, negatif, kecuali koefisien X_3 pada sampel 4. Rentang nilai R^2 untuk sampel berkisar antara -0,13 hingga 0,628, dan RMSE sampel antara 4,334 hingga 9,419, hasil temuan ini menandakan bahwa penggunaan sampel dengan ukuran yang sesuai dapat menghasilkan model regresi yang representatif dan bahkan bisa lebih baik terhadap populasi.

Dari hasil penelitian ini, terdapat beberapa poin penting yang berkaitan dengan rumusan masalah. Pertama, data populasi tidak perlu diolah seluruhnya untuk mendapatkan gambaran kuantitatif yang akurat, melainkan sampel yang representatif dapat memberikan informasi yang cukup. Kedua, akurasi model dengan menggunakan data dari sampel umumnya menunjukkan konsistensi dan kemiripan dengan model yang menggunakan data populasi, meskipun terdapat variasi dalam nilai R^2 dan RMSE maupun status RSR.

Penelitian selanjutnya disarankan untuk melakukan percobaan serupa dengan beragam kasus, baik menggunakan regresi linear maupun non-linear, serta memperluas volume data dan variabel bebas untuk evaluasi akurasi dan konsistensi model prediksi yang dihasilkan.

Ucapan Terimakasih

Terima kasih kami ucapkan kepada UPPM Polban yang telah memberikan kesempatan dan dana DIPA Politeknik Negeri Bandung sehingga penelitian ini dapat berlangsung.

Daftar Rujukan

[1] KBBI, "KBBI." Accessed: Jul. 01, 2024. [Online]. Available: https://kbbi.web.id/prediksi

[2] D. Bai, L. Ye, Z. Yang, and G. Wang, "Impact of climate change on agricultural productivity: a combination of spatial Durbin model and entropy approaches," *Int J Clim Chang Strateg Manag*, vol. 16, no. 4, pp. 26–48, Jan. 2024, doi: 10.1108/IJCCSM-02-2022-0016.

[3] BMKG Kotawaringin Timur, "Buletin Meteorologi," Jun. 2024. Accessed: Jul. 01, 2024. [Online]. Available: https://stamet-kotim.bmkg.go.id/wp-content/uploads/2024/06/Buletin-Juni-2024.pdf

[4] BMKG, "Visibility." Accessed: Jul. 02, 2024. [Online]. Available: https://maritim.bmkg.go.id/glossaries/55/Visibility

[5] Kementerian Perhubungan, "Peraturan Menteri Perhubungan Nomor 18 Tahun 2010." Accessed: Jul. 02, 2024. [Online]. Available: https://peraturan.bpk.go.id/Details/104835/permenhub-no-18-tahun-2010

[6] K. Qu, "Research on linear regression algorithm," *MATEC Web of Conferences*, vol. 395, p. 01046, May 2024, doi: 10.1051/matecconf/202439501046.

[7] N. Karna, P. Roy, and S. Shakya, *Temperature Prediction using Regression Model*. 2021.

[8] "New Regression models for Estimation daily temperature of Karachi and its Neural Network analysis," *Global NEST Journal*, Oct. 2021, doi: 10.30955/gnj.003953.

[9] A. Luthfiarta, A. Febriyanto, H. Lestiawan, and W. Wicaksono, "Analisa Prakiraan Cuaca dengan Parameter Suhu, Kelembaban, Tekanan Udara, dan Kecepatan Angin Menggunakan Regresi Linear Berganda," *JOINS (Journal of Information System)*, vol. 5, pp. 10–17, May 2020, doi: 10.33633/joins.v5i1.2760.

[10] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *Journal of Applied Computer Science and Technology*, vol. 4, pp. 58–64, Jul. 2023, doi: 10.52158/jacost.v4i1.491.

[11] Z. Liu and A. Zhang, *A Survey on Sampling and Profiling over Big Data (Technical Report)*. 2020.

[12] G. Adhikari, "Calculating the Sample Size in Quantitative Studies," *Scholars' Journal*, pp. 14–29, Dec. 2021, doi: 10.3126/scholars.v4i1.42458.

[13] I. Gupta, H. Mittal, D. Rikhari, and A. K. Singh, "MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day," Mar. 2022.

[14] Tedja Diah Rani Octavia, Neny Rosmawarni, Ati Zaidiah, and Nunik Destria Arianti, "Implementation of Multiple Linear Regression Algorithm to Predict Air Temperature Based on Pollutant Levels in South Tangerang City," *Jurnal Inotera*, vol. 9, no. 2, pp. 378–390, Aug. 2024, doi: 10.31572/inotera.Vol9.Iss2.2024.ID383.

[15] H. A. Y. Ahmed and S. W. A. Mohamed, "Rainfall Prediction using Multiple Linear Regressions Model," in *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, IEEE, Feb. 2021, pp. 1–5. doi: 10.1109/ICCCEEE49695.2021.9429650.

[16] Mohammad Mahbobi and Thomas K.Tiemann, *Introductory Business Statistics with Interactive Spreadsheets - 1st Canadian Edition*, 1st ed. BCcampus, 2010.

[17] S. G. Patro, P. Sahoo, I. Panda, and D.-K. K. Sahu, "Technical Analysis on Financial Forecasting," Mar. 2015.

[18] D. Anggoro and W. Supriyanti, "Improving Accuracy by applying Z-Score Normalization in Linear Regression and Polynomial Regression Model for Real Estate Data," *International Journal of Emerging Trends & Technology in Computer Science*, pp. 549–555, Nov. 2019, doi: 10.30534/ijeter/2019/247112019.

[19] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[20] A. Jadon, A. Patil, and S. Jadon, "A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting," Nov. 2022.

[21] S. D. Latif, "Developing a boosted decision tree regression prediction model as a sustainable tool for compressive strength of environmentally friendly concrete," *Environmental Science and Pollution Research*, vol. 28, no. 46, pp. 65935–65944, Dec. 2021, doi: 10.1007/s11356-021-15662-z.

[22] A. Mekoya, "Estimation of Evaporation at Tharandt, using Daily and Ten-Minute Class-A Pan Data from Automatic Measuring Pressure Sensor Instrument," *International Journal of Environmental Sciences & Natural Resources*, vol. 19, no. 1, Apr. 2019, doi: 10.19080/IJESNR.2019.19.556003.

[23] Norbert Budincsevity, "Weather in Szeged 2006-2016," https://www.kaggle.com/datasets/budincsevity/szeged-weather. Accessed: Jul. 01, 2024. [Online]. Available: https://www.kaggle.com/datasets/budincsevity/szeged-weather

[24] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

DOI : https://doi.org/10.52158/jacost.v6i1.971