Jurnal Sains Matematika dan Statistika
p-ISSN: 2460-4542
e-ISSN: 2615-8663

Prediksi Angka Harapan Hidup Menggunakan Regresi Linear Berganda, Lasso, Ridge, Elastic Net, dan Kuantil Lasso

Muhammad Daryl Fauzan[1], Mohamad Khoirun Najib[2]*, Sri Nurdiati[3], Nazwa Khoerunnisa[4], Syammira Dhifa Maulia[5], Raden Roro Carissa Triwulandari[6], Muhammad Farhan Aziz[7]

Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, IPB University, Bogor, Indonesia
Jl. Raya Dramaga, Babakan, Kec. Dramaga, Kabupaten Bogor, Jawa Barat, 16680
Email: ryls93daryl@apps.ipb.ac.id, mkhoirun_najib@apps.ipb.ac.id, nurdiati@apps.ipb.ac.id, nazwaakhnazwa@apps.ipb.ac.id, syammiradhifa@apps.ipb.ac.id, carissatriwulandari@apps.ipb.ac.id, farhan_aziz@apps.ipb.ac.id
*Korespondensi penulis: mkhoirun_najib@apps.ipb.ac.id

Abstrak

Angka harapan hidup mejadi salah satu indikator penting dalam mengevaluasi kesejahteraan dan kualitas hidup suatu populasi atau negara. Metode yang biasa digunakan untuk memprediksi adalah regresi linear berganda. Terdapat banyak perkembangan model regresi linear berganda, seperti regresi lasso, *ridge*, *elastic net*, kuantil, serta kuantil lasso. Untuk melihat kontribusi setiap variabel independen pada model, digunakan metode *Mean Absolute Shapley* Values (MASV). Oleh karena itu, tujuan dari penelitian ini adalah membandingkan model regresi linear berganda, lasso, *ridge*, *elastic net*, kuantil, serta kuantil lasso dalam memprediksi nilai angka harapan hidup. Penelitian diawali dengan melakukan eksplorasi data. Selanjutnya, model-model regresi tersebut dilatih. Pelatihan model tersebut juga dilakukan berulang kali dengan mengacak data pada pembagian data latih dan data uji. Terakhir, kontribusi setiap variabel independen diukur. Performa model regresi linear berganda pada iterasi pertama cukup baik dengan nilai *r-square* lebih besar dari 85% baik pada data latih dan data uji. Namun, Performa model lasso, *ridge*, *elastic net*, kuantil, dan kuantil lasso tidak jauh berbeda dengan performa model regresi linear berganda. Ketika dilakukan pengacakan data latih dan data uji. Model regresi kuantil lasso memiliki performa yang lebih konsisten dalam memprediksi nilai angka harapan hidup dibandingkan model lainnya. Pada setiap model regresi, tingkat kelahiran dan tingkat kematian bayi merupakan variabel yang memiliki kontribusi terbesar dalam memprediksi nilai angka harapan hidup, sedangkan persentase orang yang mengikuti sekolah formal dan persentase populasi yang tinggal di perkotaan bukan variabel independen yang cukup baik untuk memprediksi angka harapan hidup.

Kata Kunci: angka harapan hidup, model regresi, data latih, data uji.

Abstract

Life expectancy is an important indicator in evaluating the welfare and quality of life of a population or country. A commonly used method for prediction is multiple linear regression. There are many developments of multiple linear regression models, such as lasso, ridge, elastic net, quantile, and quantile lasso regression. To see the contribution of each independent variable to the model, the Mean Absolute Shapley Values (MASV) method is used. Therefore, the purpose of this study is to compare multiple linear regression, lasso, ridge, elastic net, quantile, and lasso quantile models in predicting life expectancy values. The research begins with data exploration. Next, the regression models were trained. The model training was also done repeatedly by randomizing the data in the division of training and test data. Finally, the contribution of each independent variable was measured. The performance of the multiple linear regression model in the first iteration was quite good with r-square values greater than 85% in both training and test data. However, the performance of the lasso, ridge, elastic net, quantile, and quantile lasso models were not much different from the performance of the multiple linear regression model. When randomizing the training and test data, the quantile lasso regression model has a more consistent performance in predicting life expectancy values than other models. In each regression model, the birth rate and infant mortality rate are the variables that have the greatest contribution in predicting life expectancy, while the percentage of people attending formal school and the percentage of the population living in urban areas are not good enough independent variables to predict life expectancy.

Keywords**:** *life expectancy, regression model, training data, testing data.*

1. Pendahuluan

Angka harapan hidup (*life expectancy*) menjadi salah satu indikator penting dalam mengevaluasi kesejahteraan dan kualitas hidup suatu populasi atau negara. Angka harapan hidup menginterpretasikan rata-rata lamanya hidup seseorang serta menjadi dasar untuk perencanaan kebijakan kesehatan, investasi dalam sektor kesehatan, dan pendidikan serta evaluasi dampak berbagai faktor terhadap harapan hidup pada suatu populasi atau negara. Berdasarkan [1], rata-rata angka harapan hidup telah meningkat sebesar 80,4% sejak tahun 1950 dan sudah berada di atas 70 tahun. Peningkatan tersebut telah mencerminkan kemajuan yang signifikan dalam berbagai sektor, seperti kesehatan, teknologi, sanitasi, pendidikan, politik, dan kebijakan sosial.

Walaupun nilai angka harapan hidup terus meningkat setiap tahunnya, berdasarkan data yang diperoleh dari Kaggle, pada tahun 2023, nilai angka harapan hidup Indonesia berada di angka 71,5 tahun. Nilai tersebut masih berada di bawah rata-rata nilai angka harapan hidup di dunia, yaitu 72,3 tahun. Selain itu, data tersebut juga menunjukkan nilai angka harapan hidup di Indonesia berada pada peringkat 114 dari 195 negara. Posisi tersebut masih jauh di bawah Singapura yang menduduki posisi ke-5 di dunia dengan nilai angka harapan hidup sebesar 83,1 tahun. Hasil tersebut menunjukkan bahwa masih ada potensi untuk meningkatkan kondisi kesehatan serta kesejahteraan masyarakat. Oleh karena itu, diperlukan suatu metode yang dapat memprediksi angka harapan hidup berdasarkan berbagai macam faktor. Identifikasi faktor-faktor yang mempengaruhi tersebut juga penting agar dapat melakukan intervensi.

Model regresi linear berganda menjadi model yang acap digunakan karena modelnya yang sederhana. Model regresi linear berganda memodelkan hubungan kausalitas antar satu variabel dependen dengan beberapa variabel independen. Selain memodelkan hubungan kausalitas, model regresi juga dapat digunakan untuk memprediksi variabel dependen. Namun, agar model regresi linear berganda menghasilkan hasil prediksi dan interpretasi yang baik, terdapat asumsi-asumsi yang harus dipenuhi. Oleh karena itu, terdapat perkembangan dari model regresi linear berganda, seperti model regresi lasso, *ridge*, dan *elastic net* yang menggunakan metode regularisasi, model regresi kuantil yang dapat mengabaikan asumsi kenormalan pada sisaan, serta model kuantil lasso yang merupakan gabungan dari model regresi kuantil dan regresi lasso. Terdapat banyak peneliti yang sudah mengaplikasikan model-model tersebut, seperti [2] menggunakan regresi linear berganda, penelitian [3] meneliti regresi lasso dan *ridge*, penelitian [4] yang menganalisis model regresi *elastic net*, penelitian [5] menerapkan model regresi kuantil, serta penelitian [6] menerapkan model regresi kuantil lasso.

Menurut [7], faktor yang mempengaruhi nilai angka harapan hidup masih menjadi perdebatan di dunia kesehatan, ekonomi, serta politik. Namun, terdapat berbagai penelitian yang memprediksi nilai angka harapan hidup berdasarkan faktor-faktor tertentu, seperti penelitian [8] menyatakan bahwa tingginya angka harapan hidup berdampak terhadap kepedulian suatu individu kepada lingkungan, penelitian [9] menyatakan bahwa angka harapan hidup akan naik jika tingkat kematian menurun, serta penelitian [10] meneliti pengaruh variabel-variabel yang berkaitan dengan kesehatan dan ekonomi terhadap angka harapan hidup.

Metode *shapley values* yang pertama kali diperkenalkan oleh Lloyd Shapley pada tahun 1953 dapat digunakan untuk mengidentifikasi pengaruh dari setiap nilai amatan terhadap hasil prediksi pada model *machine learning*, seperti penelitian [11]. Metode *shapley values* memberikan solusi untuk menginterpretasikan model *machine learning* yang kompleks. Pengaruh setiap amatan dapat diagregasi untuk mengukur kontribusi dari setiap variabel independen. Metode tersebut dikenal sebagai *Mean Absolute Shapley Values* (MASV).

Oleh karena itu, tujuan dari penelitian ini adalah membandingkan model regresi linear berganda, regresi lasso, regresi ridge, regresi elastic net, regresi kuantil, serta regresi kuantil lasso dalam memprediksi angka harapan hidup. Selanjutnya, dilakukan analisis kontribusi variabel independen dalam memprediksi angka harapan hidup dengan metode MASV.

2. Metode Penelitian

2.1 Tahapan Penelitian

Penelitian ini dilakukan dengan bahasa pemrograman Python dengan menggunakan *package Scikit Learn* sebagai *package* utama untuk membangun model. Tahapan penelitian yang dilakukan adalah sebagai berikut:

1. Eksplorasi data

   Eksplorasi data dilakukan untuk melihat sebaran dari variabel dependen, korelasi antara variabel independen dengan variabel dependen, serta VIF dari masing-masing variabel independen.

2. Pembagian Data

   Pembagian data dilakukan dengan membagi data menjadi *data training* dan *data testing* dengan rasio 75% data pada *data training* dan 25% data pada *data testing*.

3. Normalisasi Data

   Normalisasi data mengubah skala data pada variabel independen menjadi seragam. Teknik normalisasi dapat dilihat pada Persamaan (1).

$$x_i^* = \frac{x_i}{x_{max} - x_{min}} (x_{new\ max} - x_{new\ min}) + x_{new\ min} \quad (1)$$

4. Membangun model

   Model regresi dibangun menggunakan *package Scikit Learn*. *Hyperparameter* pada model-model regresi di-*tunning* menggunakan metode *Grid Search Cross Validation*.

5. Evaluasi model

   Performa setiap model regresi dievaluasi dengan visualisasi antara data aktual dengan hasil prediksi. Selain itu, evaulasi model juga dilakukan menggunakan *Root Mean Square Error* (RMSE) dan koefisien determinasi yang diberikan pada Persamaan (2) dan (3).

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3)$$

6. Analisis konsistensi model

   Analisis konsistensi dilakukan untuk menilai seberapa konsisten performa model ketika dilatih menggunakan kombinasi data latih yang berbeda.

7. Analisis kontribusi variabel independen

   Kontribusi setiap variabel independen dilakukan untuk mengetahui variabel independen yang paling mempengaruhi hasil prediksi angka harapan hidup.

2.2 Regresi Linear Berganda

Regresi Linear berganda merupakan salah satu teknik pemodelan statistika yang digunakan untuk menganalisis hubungan antara satu variabel dependen dengan beberapa variabel independen. Selain itu, berdasarkan penelitian [12], model regresi linear berganda digunakan untuk memprediksi variabel berdasarkan kombinasi linear dari variabel numerik, variabel dikotomus, atau variabel *dummy* sebagai variabel independen. Model persamaan regresi linear berganda dinyatakan pada Persamaan (4).

$$y = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \epsilon \quad (4)$$

dengan $\hat{\beta}_i$, $i \geq 1$ merupakan koefisien regresi linear berganda yang merepresentasikan perubahan nilai variabel dependen y ketika variabel independen x_i berubah sebesar satu unit, sedangkan $\hat{\beta}_0$ merupakan titik ketika garis regresi berpotongan dengan dimensi variabel dependen y, yaitu ketika semua variabel independen x_k bernilai nol [13].

Untuk menduga nilai parameter terbaik pada model regresi linear berganda, digunakan metode *Ordinary Least Square* (OLS), yaitu menduga nilai parameter dengan meminimumkan jumlah kuadrat sisaan yang dinyatakan pada Persamaan (5).

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (5)$$

dengan

$$\hat{y}_i = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i \quad (6)$$

Model regresi linear berganda akan menggunakan semua set data yang dilatih. Namun, berdasarkan penelitian [14], model regresi tidak cukup baik dalam memprediksi nilai data baru. Hal tersebut mengakibatkan model regresi linear berganda mengalami *overfitting* ketika dilakukan ekstrapolasi. Untuk mendapatkan model regresi linear yang baik, asumsi-asumsi regresi linear berganda, seperti kenormalan sisaan serta tidak terdapat multikolinearitas harus terpenuhi.

2.3 Regresi Lasso, Ridge, dan Elastic Net

Regularisasi merupakan metode untuk memperbaiki hasil prediksi dengan cara membuat nilai koefisien regresi menuju nol [4]. Metode regularisasi pada model regresi linear berganda dilakukan dengan cara memberi penalti pada metode OLS agar tidak terjadi *overfitting* ketika dilakukan ekstrapolasi. Regresi lasso, *ridge*, dan *elastic net* merupakan model regresi linear berganda yang menggunakan tenik regularisasi. Model regresi lasso memiliki penalti *L1 norm*, sedangkan regresi *ridge* memiliki penalti *L2 norm*. Model regresi *elastic net* merupakan generalisasi dari regresi lasso dan regresi *ridge*. Model regresi lasso, ridge, dan elastic net dinyatakan pada Persamaan , tetapi untuk penduga model regresi lasso, *ridge*, dan *elastic net* dinyatakan berturut-turut pada Persamaan (7)-(9).

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\hat{\beta}_j| \quad (7)$$

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \left(\hat{\beta}_j\right)^2 \quad (8)$$

$$\hat{\beta}^{elastic\ net} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum_{j=1}^{k} |\hat{\beta}_j| + (1-\alpha) \sum_{j=1}^{k} \left(\hat{\beta}_j\right)^2 \right) \quad (9)$$

dengan $\lambda > 0$ dan $0 < \alpha < 1$ merupakan *hyperparameter* model regresi serta $\sum_{j=1}^{k} |\hat{\beta}_j|$ merupakan penalti *L1 norm* dan $\sum_{j=1}^{k} \left(\hat{\beta}_j\right)^2$ merupakan penalti *L2 norm*.

Parameter λ menentukan seberapa besar penalti yang diberikan pada model sehingga berdasarkan penelitian [3], ketika λ sangat besar, beberapa koefisien regresi lasso dan elastic net dapat bernilai nol, sedangkan regresi *ridge* hanya mendekati nol. Akibatnya, menurut penelitian [15], regresi lasso, regresi *ridge*, dan regresi *elastic net* dapat mengatasi multikolinearitas, tetapi hanya regresi lasso dan regresi *elastic net* yang dapat melakukan pemilihan variabel independen terbaik. Parameter α menentukan dominasi penalti *L1 norm* dan *L2 norm* pada model *elastic net*. Parameter α menentukan seberapa dominasi penalti *L1 norm* dan *L2 norm*. Jika $\alpha = 1$, model regresi *elastic net* sama dengan

model regresi lasso, sedangkan jika $\alpha = 0$, model regresi *elastic net* akan sama dengan model regresi *ridge*.

2.4 Regresi Kuantil dan Kuantil Lasso

Regresi kuantil merupakan model untuk mendapatkan suatu objektif tertentu pada variabel dependen karena model regresi kuantil dapat memodelkan hubungan antara variabel independen dengan kuantil yang spesifik pada variabel dependen [5]. Kemampuan tersebut disebabkan oleh penduga parameter regresi yang berbeda dengan OLS. Berdasarkan penelitian [6], penduga parameter regresi kuantil dinyatakan pada Persamaan (10).

$$\hat{\beta}^{kuantil} = \arg\min_{\beta} \sum_{i=1}^{n} \rho_{\tau} (y_i - \hat{y}_i)^2 \tag{10}$$

dengan $\rho_{\tau}(u)$ merupakan fungsi kerugian asimetrik yang dinyatakan pada Persamaan (11).

$$\rho_{\tau}(u) = u(\tau - I(u < 0)) \tag{11}$$

dengan $0 < \tau < 1$ dan I adalah fungsi indikator.

Menurut penelitian [16], implikasi dari penduga parameter tersebut adalah model regresi kuantil dapat mengabaikan asumsi kenormalan pada sisaan karena sisaan pada model regresi linear berganda menyebar bebas stokastik identik mengikuti suatu sebaran tertentu.

Model regresi kuantil lasso merupakan gabungan dari model regresi kuantil dan regresi lasso dengan memberikan penalti *L1 norm* pada model regresi kuantil sehingga diharapkan model regresi kuantil lasso dapat mengabaikan asumsi normalitas sisaan, mengatasi multikolinearitas, serta memilih fitur terbaik. Penduga parameter regresi kuantil lasso diberikan pada Persamaan (12).

$$\hat{\beta}^{kuantil} = \arg\min_{\beta} \sum_{i=1}^{n} \rho_{\tau} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\hat{\beta}_j| \tag{12}$$

dengan $\lambda > 0$.

2.5 *Shapley Values* dan *Mean Absolute Shapley Values* (MASV)

Metode *shapley values* digunakan pada model *machine learning*, termasuk model berbasis regresi linear untuk mengukur besarnya kontribusi setiap titik data dalam memprediksi variabel dependen. Menurut penelitian [17], metode *shapley values* merupakan teknik yang akurat dengan kontribusi amatan pada variabel independen yang konsisten serta menghasilkan kualitas interpretasi yang sangat baik. Persamaan *shapley values* dinyatakan pada Persamaan (13).

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} (f(S \cup i) - f(S)) \tag{13}$$

dengan ϕ_i merupakan *shapley values* pada amatan ke-i, N merupakan himpunan variabel independen, S merupakan subset dari N, serta f merupakan model yang digunakan.

Untuk menghitung kontribusi setiap variabel independen, dilakukan teknik agregasi untuk setiap amatan dengan menghitung rata-rata dari *shapley values* yang

dimutlakkan. Semakin besar MASV pada suatu variabel independen, semakin besar juga kontribusi variabel independen tersebut dalam memprediksi variabel dependen. Teknik tersebut dikenal dengan *Mean Absolute Shapley Values* (MASV) yang dinyatakan pada Persamaan (14).

$$MASV_j = \sum_{i=1}^{n} \frac{|\phi_i(f)|}{n}, \forall_{j \in N} \quad (14)$$

dengan $MASV_j$ merupakan nilai MASV pada variabel independen ke-j.

2.6 Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang diperoleh dari Kaggle. Dataset ini mencakup indikator kesehatan, ekonomi, edukasi, lingkungan, serta informasi mengenai demografi dari 195 negara di dunia pada tahun 2023. Dataset ini dapat diunduh di https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023. Variabel yang digunakan pada penelitian ini diberikan pada Tabel 1.

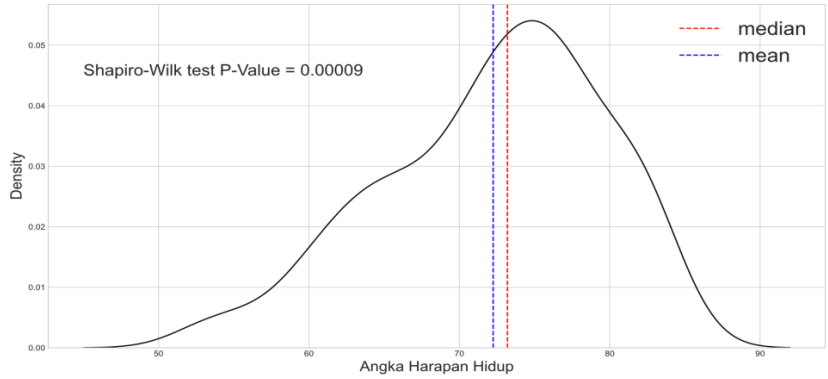
Tabel 1. Daftar Variabel Penelitian

No	Variabel	Keterangan
y	Angka Harapan Hidup	Harapan lamanya hidup individu dari lahir
X_1	Kepadatan penduduk	Banyaknya penduduk per km^2
X_2	Persentase lahan pertanian/perkebunan	Diukur dalam satuan persen
X_3	Tingkat kelahiran	Banyaknya kelahiran per 1000 populasi
X_4	Emisi CO_2	Emisi CO_2 dalam satuan ton
X_5	Tingkat kesuburan seorang Wanita	Banyaknya bayi yang lahir per satu wanita
X_6	Persentase area yang dipenuhi oleh hutan	Diukur dalam satuan persen
X_7	Produk Domestik Bruto (PDB)	Total barang dan jasa yang diproduksi
X_8	Persentase individu yang terdaftar pada sekolah formal	Diukur dalam satuan persen
X_9	Tingkat kematian bayi	Bayi yang meninggal per 1000 populasi sebelum 1 tahun
X_{10}	Persentase biaya tidak terencana untuk kesehatan	Diukur dalam satuan persen
X_{11}	Persentase populasi yang hidup di Perkotaan	Diukur dalam satuan persen

3. Hasil dan Pembahasan

3.1 Eksplorasi Data dan Normalisasi Data

Eksplorasi data yang pertama dilakukan adalah melihat sebaran dari variabel dependen. Hal itu dikarenakan model regresi linear berganda memiliki asumsi kenormalan sisaan. Sebaran pada variabel dependen menjadi diagnosis awal asumsi kenormalan sisaan terpenuhi. Sebaran variabel dependen dapat dilihat pada Gambar 1.



Gambar 1. Sebaran Angka Harapan Hidup (y)

Berdasarkan Gambar 1, sebaran variabel dependen cenderung tidak menyebar normal. Pernyataan tersebut didukung oleh uji Shapiro-Wilk yang menghasilkan *p-value* sebesar 0,00009, lebih kecil dari taraf nyata $\alpha = 0{,}05$ sehingga dapat disimpulkan bahwa variabel dependen tidak menyebar normal. Hal tersebut berimplikasi bahwa asumsi kenormalaan sisaan diduga tidak terpenuhi.

Selanjutnya, dilihat hubungan linear antara variabel dependen dengan variabel independen menggunakan korelasi pearson. Selain itu, dilihat juga adanya multikolinearitas pada variabel indepdenden menggunakan *Variance Inflation Factor* (VIF). Korelasi pearson serta VIF dapat dilihat pada Tabel 2.

Tabel 2. Korelasi Pearson dan VIF

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
$\rho(y, X_i)$	-0,15	-0,23	-0,87	-0,12	-0,84	0	0,18	0	-0,92	-0,31	0,01
VIF	1,06	1,30	31,91	21,79	26,85	1,37	8,46	18,08	4,13	1,20	9,36

Berdasarkan Tabel 2, X_3, X_4, dan X_8 memiliki korelasi yang kuat dengan variabel dependen y dengan nilai korelasi pearson yang mendekati -1. Selain itu, X_3, X_4, X_5, X_7, X_8, dan X_{11} memiliki nilai VIF≥ 5 sehingga terdapat multikolinearitas pada variabel-variabel tersebut.

Selanjutnya, data dibagi menjadi *data training* dan *data testing*. Namun, satuan pada variabel independen berbeda-beda sehingga variabel-variabel tersebut memiliki skala yang tidak seragam. Menurut [18], Skala yang tidak seragam menghasilkan hasil prediksi yang tidak baik. Oleh karena itu, dilakukan teknik normalisasi data untuk menyeragamkan skala pada variabel independen. Normalisasi data di-*fit* pada *data training* kemudian hasil *fitting* tersebut digunakan pada *data testing*.

3.2 Model Regresi

3.2.1 Menentukan Nilai *Hyperparameter*

Teknik pembentukan model diawali dengan mencari *hyperparameter* terbaik untuk setiap model regresi. Diharapkan *hyperparameter* yang dipilih dapat meningkatkan performa model serta mengatasi *overfitting*. Untuk mengevaluasi metode *grid search cross validation*, digunakan *loss function* RMSE. Nilai *hyperparameter* yang di-*tunning* dan nilai *hyperparameter* terbaik diberikan pada Tabel 3.

Tabel 3. Nilai *Hyperparameter* Model Regresi

No	Model Regresi	*Hyperparameter* yang di-*tunning*	Nilai *hyperparameter* terbaik
1	Regresi Linear Berganda	Tidak ada	Tidak ada
2	Regresi Lasso	$\lambda \in \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$	$\lambda = 0{,}01$
3	Regresi Ridge	$\lambda \in \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$	$\lambda = 0{,}1$
4	Regresi *Elastic Net*	$\lambda \in \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$, $\alpha \in \{0{,}05; 0{,}1; \dots; 0{,}95\}$	$\lambda = 0{,}01; \alpha = 0{,}8$
5	Regresi Kuantil	$q \in \{0{,}05; 0{,}1; \dots; 0{,}95\}$	$q = 0.5$
6	Regresi Kuantil-Lasso	$q \in \{0{,}05; 0{,}1; \dots; 0{,}95\}$, $\lambda \epsilon \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$	$\lambda = 0{,}01; q = 0{,}55$

λ pada penalti *L1 norm* pada model regresi lasso lebih kecil dibandingkan dengan λ pada penalti *L2 norm* pada model regresi ridge. Namun, pada model regresi elastic net, nilai λ sama dengan model regresi lasso. Nilai $\alpha = 0{,}8$ juga menginterpretasikan bahwa penalti *L1 norm* lebih dominan daripada penalti *L2 norm* sehingga model elastic net cenderung bergerak ke arah model regresi lasso.

Nilai q pada model regresi kuantil dan kuantil lasso menghasilkan nilai yang berbeda. Kuantil yang dipilih pada model regresi kuantil adalah 0,5, sedangkan kuantil yang dipilih pada model regresi kuantil lasso adalah 0,55. Hal ini menginterpretasikan bahwa metode regularisasi dapat mengubah kuantil yang spesifik pada variabel dependen. Selain itu, nilai λ pada model kuantil lasso sama dengan λ pada model regresi lasso.

3.2.2 Parameter Model Regresi

Setelah mendapatkan *hyperparameter* terbaik. Model regresi yang dibangun menggunakan metode *Grid Search Cross Validation* dievaluasi parameter modelnya. Parameter model regresi dapat dilihat pada Tabel 4.

Tabel 4. Nilai Parameter Model Regresi

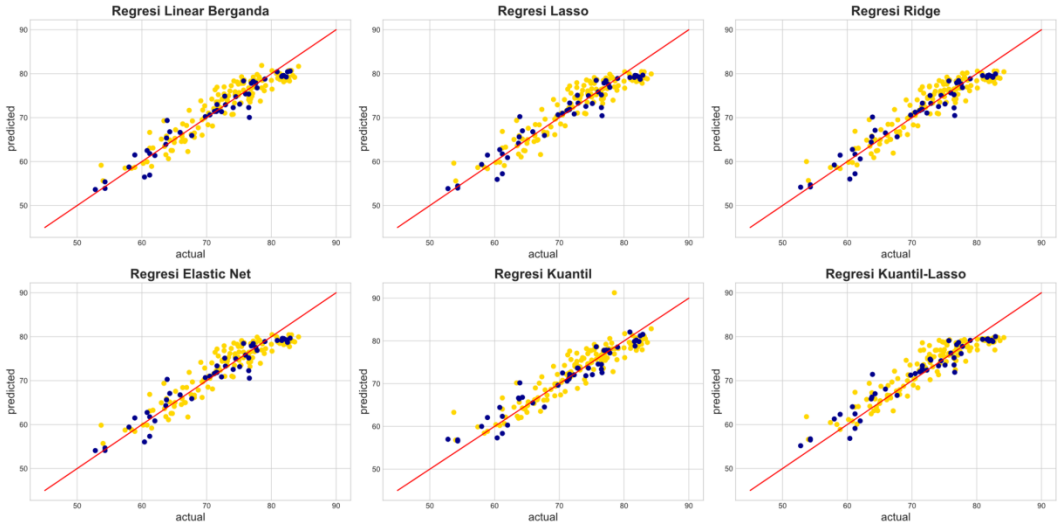
No	Parameter	Koefisien Regresi					
		OLS	Lasso	Ridge	*Elastic Net*	Kuantil	Kuantil-Lasso
1	β_1	-0,149	-0,179	-0,267	-0,270	0,439	0
2	β_2	-2,616	-2,191	-2,456	-2,251	-2,416	-0,221
3	β_3	-8,094	-7,234	-6,780	-6,260	-14,445	-8,851
4	β_4	-31,051	0	-8,823	-0,031	-30,163	0
5	β_5	1,359	0	-0,980	-2,197	3,834	0
6	β_6	-2,964	-2,380	-2,692	-2,342	-4,429	-1,289
7	β_7	14,915	0,662	5,422	0,798	27,144	0
8	β_8	-0,118	0	0,164	0	3,239	0
9	β_9	-18,002	-17,548	-16,976	-16,006	-12,804	-15,238
10	β_{10}	-1,4578	-1,276	-1,525	-1,606	-1,378	-0,418
11	β_{11}	14,844	0	-2,745	0	4,417	0
12	β_0	82,788	82,448	82,711	82,484	81,528	81,166

Pada beberapa model-model regresi, variabel X_8 dan X_{11} memiliki hubungan kausalitas yang lemah dengan variabel dependen. Hal itu didukung pada model regresi lasso, ridge, elastic net, dan kuantil lasso bahwa parameter tersebut mendekati nol, bahkan bernilai nol ketika diberikan penalti. Selain itu, Terdapat koefisien regresi yang tidak konsisten pada setiap model, yaitu bernilai positif pada beberapa model dan

bernilai negatif pada model lainnya. Parameter tersebut adalah β_1, β_5, β_8, dan β_{11}. Hal tersebut merupakan akibat dari multikolinearitas. Keberadaan multikolinearitas dapat memberikan informasi yang salah mengenai hubungan antara variabel independen dengan variabel dependen [19].

3.3 Evaluasi Model Regresi

Evaluasi model diawali dengan memvisualisasikan nilai aktual dengan nilai prediksi. Visualisasi diberikan dalam bentuk *scatter plot*. Semakin banyak titik yang dekat dengan garis $y = x$, semakin baik juga performa dari model regresi tersebut. Visualisasi antara nilai aktual dan prediksi disajikan pada Gambar 2.



Gambar 2. *Scatter Plot* antara Nilai Aktual dan Nilai Prediksi

Berdasarkan Gambar 2, performa dari setiap model regresi linear cukup baik karena titik-titik tersebut mendekati garis $y = x$. Namun, sulit untuk membandingkan performa setiap model regresi dari diagram tersebut. Oleh karena itu, digunakan RMSE dan koefisien determinasi untuk membandingkan performa model-model tersebut. RMSE mengukur rata-rata kuadrat dari selisih nilai aktual dengan nilai prediksi, sedangkan koefisien determinasi mengukur besarnya proporsi variasi yang dilprediksi oleh variabel independen terhadap variabel dependen. Nilai RMSE dan koefisien determinasi dilihat pada Tabel 5.

Tabel 5. RMSE dan Koefisien Determinasi Model Regresi

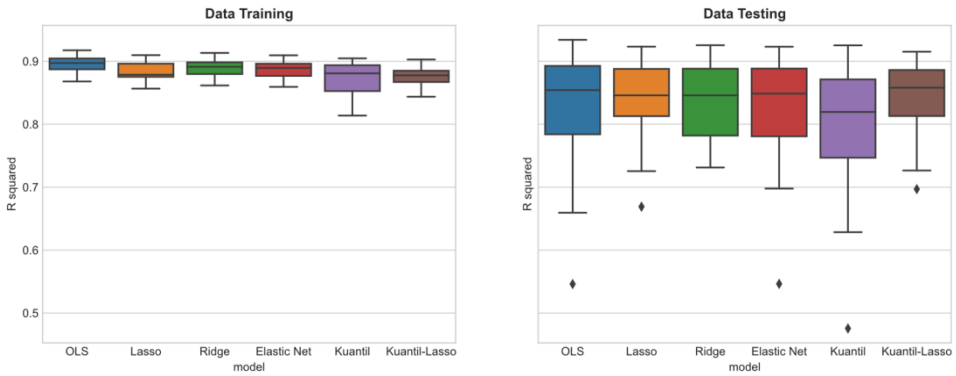
No	Model	RMSE		R^2	
		Data Training	*Data Testing*	*Data Training*	*Data Testing*
1	Regresi Linear Berganda	2,481	2,160	0,868	0,934
2	Regresi Lasso	2,560	2,332	0,859	0,923
3	Regresi Ridge	2,525	2,297	0,863	0,926
4	Regresi *Elastic Net*	2,571	2,384	0,858	0,920
5	Regresi Kuantil	2,748	2,301	0,838	0,925
6	Regresi Kuantil-Lasso	2,670	2,451	0,844	0,915

Nilai koefisien determinasi model-model regresi yang berada di atas 80% pada *data training* dan di atas 90% pada *data testing* juga menginterpretasikan bahwa semua model

regresi menduga nilai angka harapan hidup dengan baik. Regresi linear berganda memiliki performa yang paling baik dalam memprediksi angka harapan hidup dengan nilai koefisien determinasi 0,868 pada *data training* dan 0,934 pada *data testing* yang lebih tinggi dari model lainnya. Nilai RMSE model regresi linear pada *data training* dan *data testing* memiliki nilai yang paling kecil sehingga regresi linear berganda merupakan model terbaik yang dapat memprediksi nilai angka harapan hidup. Namun, ke-5 model lainnya memiliki nilai RMSE dan koefisien determinasi yang tidak jauh dari model regresi linear berganda. Selisih nilai RMSE model-model regresi dengan model regresi linear berganda tidak melebihi 0,3. Nilai koefisien determinasi model-model regresi dengan model regresi linear juga tidak melebihi 3%.

3.4 Analisis Konsistensi Model Regresi

Analisis konsistensi model dilakukan dengan cara mengacak data yang digunakan pada *data training* dan *data testing*. Pengacakan tersebut dilakukan sebanyak 20 kali. Data yang diacak tersebut digunakan untuk mem-*fit* ulang model-model regresi. Pengacakan dan Pembangunan model kembali dilakukan dalam satu iterasi. Setelah itu, performa dari setiap iterasi diamati sebarannya. Model yang memiliki variansi paling kecil serta tendensi sentral yang lebih besar merupakan model yang baik dan konsisten. Sebaran dari koefisien determinasi pada setiap model regresi dapat dilihat pada Gambar 3.



Gambar 3. Sebaran Koefisien Determinasi Model Regresi

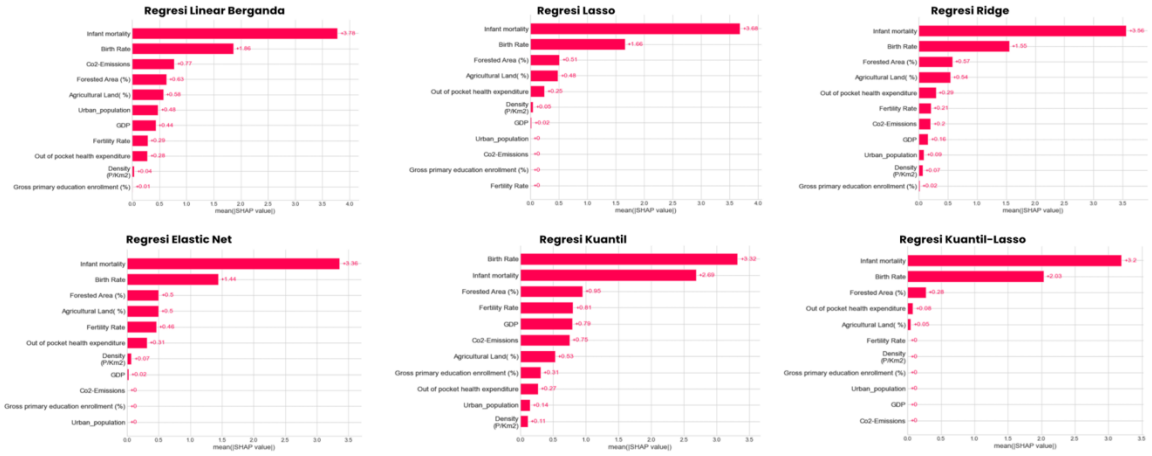
Pada *data training*, model regresi linear berganda bekerja dengan sangat baik. Jarak interkuartilnya menginterpretasikan bahwa model regresi linear berganda konsisten pada *data training*. Selain itu, nilai median koefisien determinasi pada model regresi linear lebih tinggi dari ke-5 model lainnya. Nilai *upperbound*-nya pun lebih tinggi dari model regresi lainnya. Namun, pada *data testing*, model regresi linear berganda tidak begitu konsisten karena jarak interkuartilnya cukup besar. Hal ini menginterpretasikan bahwa terdapat *overfitting* pada model regresi linear berganda.

Model regresi ridge dan *elastic net* cukup konsisten pada *data training* karena jarak interkuartil mereka tidak jauh berbeda. Namun, pada *data testing*, jarak interkuartil model regresi ridge dan *elastic net* juga lebih besar. Model yang sangat tidak konsisten baik di *data training* maupun *data testing* adalah model regresi kuantil. Perubahan *data training* yang digunakan menyebabkan pemilihan nilai kuantil ikut berubah. Perubahan nilai tersebut menyebabkan performa yang tidak konsisten, bahkan *overfitting*.

Berbeda dengan regresi lasso dan kuantil-lasso, kedua model regresi tersebut dapat dikatakan cukup konsisten pada *data training* dan *data testing*. Model regresi lasso dan kuantil-lasso mampu mempertahankan performa model pada *data training* dan *data testing*. Hal itu ditunjukkan dari jarak interkuartil pada *data training* dan *data testing* yang lebih kecil model lainnya. Dapat disimpulkan bahwa model yang hanya diberi penalti *L1 norm* lebih konsisten dari model yang tidak diberi penalti tersebut dalam memprediksi nilai angka harapan hidup. Model regresi kuantil-lasso menjadi model yang paling konsisten karena jarak interkuartil pada *data testing* relatif lebih kecil dan tidak berbeda jauh pada *data training* dibandingkan dengan regresi lasso.

3.5 Analisis Kontribusi Variabel Independen

Metode MASV dihitung menggunakan *package* SHAP yang tersedia pada bahasa pemrograman Python. Nilai MASV pada setiap model divisualisasikan menggunakan *bar chart* yang dapat dilihat pada Gambar 4.



Gambar 4. *Bar Chart* Nilai MASV pada Model Regresi

Pada model regresi linear berganda, ridge, dan kuantil, setiap variabel independen memiliki kontribusi terhadap hasil prediksi nilai angka harapan hidup. Namun, berbeda dengan model regresi lasso, *elastic net*, dan kuantil-lasso yang tidak semua variabel independen memiliki kontribusi pada hasil prediksi nilai angka harapan hidup. Hal ini dikarenakan penalti *L1 norm* yang dapat melakukan pemilihan variabel independen terbaik. Variabel independen yang mengandung multikolinearitas cenderung memiliki kontribusi yang sedikit. Hal itu terjadi karena variabel independen yang mengandung multikolinearitas tersebut sudah diwakili oleh variabel independen yang lebih dominan.

MASV juga menjelaskan bahwa variabel independen dengan nilai koefisien mutlak terbesar akan memiliki kontribusi yang terbesar, seperti emisi karbon dioksida pada model regresi linear yang memiliki koefisien sebesar -31,051 memiliki nilai MASV sebesar 0,77 yang masih berada di bawah tingkat kematian bayi. Tingkat kematian bayi dan tingkat kelahiran menjadi dua variabel yang memiliki kontribusi terbesar dalam memprediksi nilai angka harapan hidup pada setiap model regresi. Tingkat kematian bayi berada di urutan pertama pada setiap model regresi, keuali model regresi kuantil,

yaitu tingkat kelahiran. Selain itu, MASV menjelaskan bahwa persentase orang yang mengikuti sekolah formal dan persentase populasi yang tinggal di perkotaan tidak memiliki kontribusi yang berarti pada setiap model regresi. Informasi ini mendukung bahwa kedua variabel tersebut tidak baik untuk memprediksi nilai angka harapan hidup. Penelitian sebelumya juga menemukan bahwa tingkat kematian bayi dan tingkat kelahiran memiliki kontribusi besar dalam memprediksi angka harapan hidup di berbagai model regresi. Penelitian [20] menunjukkan tingginya tingkat kematian bayi seringkali mencerminkan buruknya kondisi kesehatan masyarakat dan layanan kesehatan yang secara signifikan mengurangi angka harapan hidup. Selain itu, persentase orang yang mengikuti sekolah formal dan persentase populasi yang tinggal di perkotaan tidak memiliki kontribusi juga didukung oleh penelitian [21] yang menunjukkan bahwa faktor-faktor ini memiliki hubungan yang lebih kompleks dengan angka harapan hidup yang mungkin dipengaruhi oleh variabel lain, seperti kualitas pendidikan dan infrastruktur perkotaan.

4. Kesimpulan

Model regresi linear berganda merupakan model yang sangat baik untuk memprediksi nilai angka harapan hidup. Namun, model regresi linear tidak memenuhi asumsi kenormalan serta multikolinearitas sehingga memungkinkan terjadinya informasi hubungan yang salah antara variabel dependen dengan variabel independen. Model regresi linear pun tidak konsisten pada *data testing* sehingga terdapat *overfitting* pada model regresi linear berganda. Model regresi lasso, ridge, *elastic net*, kuantil, serta kuantil-lasso memiliki performa model yang baik dan tidak jauh berbeda dengan model regresi linear berganda. Selain itu, model regresi kuantil-lasso merupakan model yang memenuhi asumsi-asumsi tersebut. Selain mengabaikan kenormalan dan menangani multikolinearitas, model regresi lasso-kuantil merupakan model yang paling konsisten. Pada setiap model regresi, tingkat kelahiran dan tingkat kematian bayi merupakan variabel yang memiliki kontribusi terbesar dalam memprediksi nilai angka harapan hidup, sedangkan persentase orang yang mengikuti sekolah formal dan persentase populasi yang tinggal di perkotaan bukan variabel independen yang cukup baik untuk memprediksi angka harapan hidup.

Daftar Pustaka

[1] U. Nation, "By Location | Pivot Table | Data Portal," *Population Division Data Portal*, 2022. https://population.un.org/dataportal/data/indicators/61/locations/360/start/1950/end/2023/table/pivotbylocation?df=d4c5d59b-bb22-4c62-b010-cfe63f0c5c37 (accessed Jul. 02, 2024).

[2] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, 2020, doi: 10.1016/j.dsx.2020.07.045.

[3] L. E. Melkumova and S. Y. Shatskikh, "Comparing Ridge and LASSO estimators for data analysis," in *Procedia Engineering*, 2017, vol. 201, pp. 746–755. doi:

10.1016/j.proeng.2017.09.615.
[4] C. Hans, "Elastic Net Regression Modeling With the Orthant Normal Prior," *J. Am. Stat. Assoc.*, vol. 106, p. 1383, 2011.
[5] Y. Sun, K. L. Teow, B. H. Heng, C. K. Ooi, and S. Y. Tay, "Real-time prediction of waiting time in the emergency department, using quantile regression," *Ann. Emerg. Med.*, vol. 60, no. 3, pp. 299–308, 2012, doi: 10.1016/j.annemergmed.2012.03.011.
[6] D. Santri and Y. Hanike, "Pemodelan Statistical Downscaling Regresi Kuantil Lasso dan Analisis Komponen Utama untuk Pendugaan Curah Hujan Ekstrim," *Math. Appl. J.*, pp. 47–57, 2020.
[7] S. Setyadi, A. Kustanto, and A. Widiastuti, "Life Expectancy in Indonesia: The Role of Health Infrastructure, Political, and Socioeconomic Status," *Iran. Econ. Rev.*, vol. 27, no. 3, pp. 965–1005, 2023, doi: 10.22059/ier.2023.329904.1007259.
[8] F. Mariani, A. Pérez-Barahona, and N. Raffin, "Life expectancy and the environment," *J. Econ. Dyn. Control*, vol. 34, no. 4, pp. 798–815, 2010, doi: 10.1016/j.jedc.2009.11.007.
[9] S. H. Woolf and H. Schoomaker, "Life Expectancy and Mortality Rates in the United States, 1959-2017," *JAMA - J. Am. Med. Assoc.*, vol. 322, no. 20, pp. 1996–2016, 2019, doi: 10.1001/jama.2019.16932.
[10] N. Shahira Pisal, S. Abdul-Rahman, M. Hanafiah, and S. I. Kamarudin, "Prediction of Life Expectancy for Asian Population Using Machine Learning Algorithms," *Malaysian J. Comput.*, vol. 7, no. 2, pp. 1150–1161, 2022.
[11] M. Smith and F. Alvarez, "Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19," *Expert Syst. Appl.*, vol. 176, 2021, doi: 10.1016/j.eswa.2021.114832.
[12] M. Esmaeili, M. Osanloo, F. Rashidinejad, A. Aghajani Bazzazi, and M. Taji, "Multiple regression, ANN and ANFIS models for prediction of backbreak in the open pit blasting," *Eng. Comput.*, vol. 30, no. 4, pp. 549–558, 2014, doi: 10.1007/s00366-012-0298-2.
[13] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*, 5th ed., vol. 95, no. 452. New Jersey: John Wiley & Sons, Inc., 2012. doi: 10.2307/2669806.
[14] D. M. McNeish, "Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences," *Multivariate Behav. Res.*, vol. 50, no. 5, pp. 471–484, 2015, doi: 10.1080/00273171.2015.1036965.
[15] S. Altelbany, "Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study," *J. Appl. Econ. Bus. Stud.*, vol. 5, no. 1, pp. 131–142, 2021, doi: 10.34260/jaebs.517.
[16] M. Maciak, "Quantile LASSO in arbitrage-free option markets," *Econom. Stat.*, vol. 18, pp. 106–116, 2021, doi: 10.1016/j.ecosta.2020.05.006.
[17] A. Messalas, Y. Kanellopoulos, and C. Makris, "Model-Agnostic Interpretability with Shapley Values," in *10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019*, 2019. doi: 10.1109/IISA.2019.8900669.
[18] S. Garcia, J. Luengo, and F. Herrera, *Data Preprocessing and Data Mining*. Warsaw: Polish Academy of Scieces, 2015.
[19] D. C. Montgomery, E. A. Peck, and G. G. Vinning, *Introduction to Linear Regression Analysis*, 5th ed. New Jersey: John Wiley & Sons, Inc., 2012.

[20] P. Roffia, A. Bucciol, and S. Hashlamoun, "Determinants of life expectancy at birth: a longitudinal study on OECD countries," *Int. J. Heal. Econ. Manag.*, vol. 23, no. 2, pp. 189–212, 2023, doi: 10.1007/s10754-022-09338-5.
[21] M. M. Rahman, R. Rana, and R. Khanam, "Determinants of life expectancy in most polluted countries: Exploring the effect of environmental degradation," *PLoS One*, vol. 17, no. 1 January, 2022, doi: 10.1371/journal.pone.0262802.