
**DATA MINING & ANALYTIC FORECASTING INDEKS STANDAR
PENCEMAR UDARA JAKARTA MENGGUNAKAN METODE LINEAR
REGRESSION (STUDI KASUS: DATASET INDEKS STANDAR
PENCEMAR UDARA JAKARTA 2021)**

Muhammad Abdul Latief, Yuli Karyanti
Magister Manajemen Sistem Informasi
abdullatief007@gmail.com, yuli@staff.gunadarma.ac.id

Received: 01-09-2022
Revised : 10-09-2022
Accepted: 20-09-2022

Abstrak (Indonesia)

Pencemaran udara merupakan hal yang harus diperhatikan saat ini. Besarnya partikel dan gas yang beracun di udara sudah melewati batasan yang cukup besar diberbagai negara salah satunya Indonesia. Ibukota Indonesia yaitu Jakarta merupakan pusat perekonomian yang dimana berdasarkan data yang didapatkan polusi tingkat pencemaran udara tergolong melewati batas standar AQG (Air Quality Guide) yang ditetapkan WHO (World Health Organization). Berdasarkan data Jakarta memiliki rata-rata polusi yang dapat dikategorikan sedang, yang dimana rentang indeks berada pada nilai 51-100 ($\mu\text{g}/\text{m}^3$). Hal ini diukur dari berbagai pengukuran gas dan partikel yakni PM_{10} , $\text{PM}_{2.5}$, NO_2 , SO_2 , CO , O_3 , dan HC . Pada penelitian ini akan dilakukan uji coba untuk memprediksi hasil pengukuran tersebut dimana dataset yang diperoleh merupakan hasil akuisisi dari website Opendata Jakarta. Penelitian ini bertujuan untuk menghitung, menghasilkan, serta memvisualisasikan hasil dari prediksi menggunakan pendekatan metode Linear Regression. Dengan Parameter yaitu faktor (Y_1) Critical, (Y_2) Max, dan (Y_3) Category dan Parameter factor yaitu dari (X_1) PM_{10} , (X_2) $\text{PM}_{2.5}$, (X_3) SO_2 , (X_4) CO , (X_5) O_3 , (X_6) NO_2 , (X_7) Tanggal, dan (X_8) Stasiun.

Kata kunci: Data Mining; Forecasting Analysis Linear Regression; Predictive Modelling; Indeks Standar Pencemar Udara;

Abstract (English)

Air pollution is something that must be considered at this time. The amount of toxic particles and gases in the air has crossed a fairly large limit in various countries, one of which is Indonesia. The capital city of Indonesia, namely Jakarta, is the center of the economy, which based on the data obtained, the level of air pollution is classified as exceeding the AQG (Air Quality Guide) standard set by WHO (World Health Organization). Based on the data, Jakarta has an average pollution that can be categorized as moderate, where the index range is at a value of 51-100 ($\mu\text{g}/\text{m}^3$). This is measured

from various gas and particle measurements, namely PM10, PM2.5, NO₂, SO₂, CO, O₃, and HC. . This study aims to calculate, generate, and visualize the results of predictions using the Linear Regression method approach. With parameters, namely (Y1) Critical, (Y2) Max, and (Y3) Category and Parameter factors, namely (X1) PM10, (X2) PM2.5, (X3) SO₂, (X4) CO, (X5) O₃, (X6) NO₂, (X7) Date, and (X8) Station.

Keywords: data mining; linear regression forecasting analysis; predictive models; air pollutant standard index

*Correspondent Author : Muhammad Abdul Latief

Email : abdullatief007@gmail.com



PENDAHULUAN

Perkembangan gadget dan teknologi mengakibatkan Data dan informasi berkembang secara luas dan eksponensial. Lebih dari 20 tahun investasi bisnis dalam infrastruktur teknologi, yang berdampak pesatnya kemampuan untuk menerima, menganalisa, mengolah, dan mengimplementasikan Data yang dapat digunakan sebagai informasi. Oleh karena itu sudah berkembang banyak perusahaan yang menggunakan Data Mining dan analisis. Menganalisa studi kasus membantu dalam mendapatkan pengetahuan untuk menjangkau tujuan dari suatu organisasi/perusahaan ataupun secara personal. Pemanfaatan teknik Data Mining dan analisis digunakan mulai dari perusahaan segi manufaktur, Stock (Saham), E-commerce, Industri, Kedokteran, Keuangan, Ekonomi, Cuaca dan lainnya yang hampir seluruh aspek dalam kehidupan dapat dilakukan analisa salah satunya adalah dalam penelitian ini adalah Lingkungan. Polusi udara merupakan salah satu fenomena lingkungan yang cukup besar saat ini. Dimana industri banyak pabrik-pabrik yang berkembang, transportasi yang semakin bertambah volume nya, bencana kebakaran hutan, penggerusan hutan, alat pembakaran rumah tangga, dan sebagainya merupakan sumber umum pencemaran udara yang membuat kontaminasi kualitas udara hampir diseluruh dunia. Kontaminasi lingkungan dalam atau luar ruangan oleh zat kimia, fisik, atau biologis apapun yang mengubah karakteristik alami atmosfer. Polutan yang menjadi perhatian utama kesehatan masyarakat termasuk Partikulat, Karbon Monoksida, Ozon, Nitrogen Dioksida dan Sulfur Dioksida.

Penelitian sebelumnya terkait algoritma *Linear Regression* adalah penelitian Oleh (Nurjannah, dkk, 2019). Dimana penelitian ini meneliti keterkaitan perilaku yang menimbulkan terjadi hipertensi dan menentukan factor determinasinya. Sample yang diolah untuk dianalisis berjumlah 200 sample. Metode yang digunakan adalah *Linear Regression*. Perilaku yang diujikan dalam kasus ini adalah efikasi diri/kepercayaan diri, pengetahuan, dukungan keluarga, dukungan tenaga Kesehatan, jenis kelamin, dan umur. Hasil yang didapatkan dari penelitian ini adalah efikasi diri, pengetahuan, dukungan keluarga, dan dukungan tenaga Kesehatan memiliki nilai yang signifikan terhadap pencegahan penyakit hipertensi, di mana confidence Interval yang digunakan 95% dengan hasil menunjukkan variabel bebas memiliki hasil hubungan yang *positive*, sedangkan untuk variabel umur dan jenis kelamin memiliki korelasi dengan pencegahan namun tidak terdapat signifikansi yang baik. Oleh karena itu dalam penelitian ini yang akan menjadi bahan pembahasan adalah Data tentang pencemaran udara mengenai Indeks Standar Pencemar Udara (ISPU) yang diukur dari 5 stasiun pemantau kualitas udara (SPKU) yang ada di Provinsi DKI Jakarta Tahun 2021. Pendekatan metode yang akan dilakukan adalah *Data Mining & Analytic Forecasting Indeks Standar Pencemar Udara Jakarta Menggunakan Metode Linear Regression (Studi Kasus: Dataset Indeks Standar Pencemar Udara Jakarta 2021)*

forecasting Data Analysis menggunakan Linear Regression. Dengan *Parameter* yaitu faktor (Y_1) *Critical*, (Y_2) *Max*, dan (Y_3) *Category* dan *Parameter factor* yaitu dari (X_1) PM_{10} , (X_2) $PM_{2.5}$, (X_3) SO_2 , (X_4) CO , (X_5) O_3 , (X_6) NO_2 , (X_7) Tanggal, dan (X_8) Stasiun.

METODE PENELITIAN

Metode Linear Regression merupakan salah satu dari metode statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antara variabel faktor Independent (bebas) terhadap variabel Dependent (terikat). Faktor Independent pada umumnya dilambangkan dengan X atau disebut juga dengan Predictor sedangkan variabel Dependent dilambangkan dengan Y atau disebut juga dengan Response. SLR (Simple Linear Regression) juga merupakan salah satu Metode Statistik yang dipergunakan dalam produksi untuk melakukan peramalan ataupun prediksi tentang karakteristik kualitas maupun Kuantitas, Berikut persamaanya

$$Y=a+bx$$

$$Y=a+bx$$

Dari rumusan tersebut dimana :

Y = Variabel Response atau Variabel Akibat (Dependent).

X = Variabel Predictor atau Variabel Faktor Penyebab (Independent).

a = konstanta.

b= koefisien regresi (kemiringan) atau besaran Response yang ditimbulkan oleh Predictor. Dalam penelitian ini akan membandingkan berbagai metode Linear Regression yang merupakan turunan rumus dari Simple Linear Regression yaitu metode :

1. Multi Linear Regression

Multiple Linear Regression adalah model regresi berganda jika variabel terikatnya berskala *Data Interval* atau rasio (kuantitatif atau numerik). Sedangkan variabel bebas pada umumnya juga berskala *Data Interval* atau rasio. *Multiple Linear Regression* memiliki variabel bebas yang lebih dari satu. Berikut merupakan rumusannya :

$$Y = a + \beta_1 X_1 + \beta_2 X_2 \dots \beta_n X_n$$

2. Ridge Regression

Ridge Regression merupakan salah satu dari beberapa metode yang telah diusulkan untuk mengatasi Multikolinearitas dengan memodifikasi metode kuadrat terkecil untuk memperbolehkan estimator koefisien regresi yang bias. *Ridge Linear Regression* dapat digunakan untuk mengatasi korelasi yang tinggi antara beberapa variabel bebas. Berikut merupakan rumusannya:

$$L_{ridge}(\beta^{\wedge}) = \sum_{i=1}^n (y_i - x_i' \beta^{\wedge})^2 + \lambda \sum_{j=1}^m \beta_j^2$$

3. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) atau biasa disebut juga Operator Penyusutan dan Seleksi Terkecil Absolut, secara konseptual hampir mirip dengan regresi *Ridge*. Dimana menambahkan penalti dengan membatasi jumlah nilai absolut dari perkiraan untuk koefisien bukan nol. Lasso membatasi ukuran dengan nilai Absolute dari koefisien regresi, dengan menetapkan nilai yang tidak relevan menjadi 0. Berikut merupakan rumusannya:

$$L_{lasso}(\beta^{\wedge}) = \sum_{i=1}^n (y_i - x_i' \beta^{\wedge})^2 + \lambda \sum_{j=1}^m |\beta_j|$$

4. Elastic Net Regression

Elastic Net adalah menggabungkan Batasan dari regresi *Ridge* dan laso untuk mendapatkan yang terbaik dari kedua dunia. Yaitu mengambil nilai jumlah kuadrat eror dan jumlah nilai *Absolute*. Dimana terdapat *Parameter* pencampuran antara *punggungan* ($\alpha = 0$) dan laso ($\alpha = 1$). Sekarang, terdapat dua *Parameter* untuk diberikan. *Elastic Net* bertujuan untuk meminimalkan fungsi kerugian, berikut merupakan rumusan sederhana:

$$L_{enet}(\beta^{\wedge}) = \sum_{i=1}^n \frac{(y_i - x_i' \beta^{\wedge})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^m \beta_j^2 + \alpha \sum_{j=1}^m |\beta_j| \right)$$

5. Polynomial Regression

Regresi linier sederhana dapat diperluas dengan membangun fitur polinomial dari koefisien. Dalam kasus regresi linier standar, Regresi polinomial merupakan regresi di mana fungsinya adalah kuadratik. Perbedaan persamaannya bisa dilihat sebagai berikut:

$$Y = \alpha_0 + (\alpha_1 * x_1) + (a_2 * x_2)^2 + .. + a_n * x_n^n$$

6. Uji Hipotesis ANOVA

Anova (Analysis of variance) adalah sebuah analisis statistik yang menguji perbedaan rerata antar grup. Anova adalah dapat menguji perbedaan lebih dari dua kelompok. Hipotesis nol dari uji Anova adalah bahwa data adalah simple random dari populasi yang sama sehingga memiliki ekspektasi mean dan varians yang sama. Jika nilai pada *f* hitung lebih besar dari *f* tabel, maka dapat disimpulkan bahwa analisis anova adalah *H1* atau analisis diterima. Sebaliknya, apabila nilai *f* tabel lebih kecil daripada *f* hitung, maka hasil analisis uji anova ditolak atau *H0*. Prinsip dari anova adalah dengan melakukan analisis data dalam dua sumber variasi, yaitu variasi dalam kelompok dan variasi luar kelompok.

HASIL DAN PEMBAHASAN

A. Hasil Penelitian

	Train CV	Test CV
Multiple Linear Regression (MLR)	0.935	0.936
Ridge Linear Regression (RLR)	0.935	0.936
Lasso Linear Regression (LLR)	0.934	0.936
Elastic-Net Regression (ENR)	0.930	0.932
Polynomial Regression (PNR)	-6.020	0.234

	Train-R2	Test-R2	Train-MAPE	Test-MAPE	Train-MSE	Test-MSE	Train-RMSE	Test-RMSE	Train-RSS	Test-RSS
Multiple Linear Regression (MLR)	0.938	-1.872	0.043	0.393	23.961	1182.350	4.895	34.385	15191.401	7.507932e+05
Ridge Linear Regression (RLR)	0.938	-1.872	0.043	0.393	23.961	1182.152	4.895	34.382	15191.413	7.506664e+05
Lasso Linear Regression (LLR)	0.936	-1.832	0.043	0.391	24.468	1165.665	4.946	34.142	15512.784	7.401917e+05
Elastic-Net Regression (ENR)	0.932	-1.718	0.045	0.383	26.042	1118.768	5.103	33.448	16510.834	7.104117e+05
Polynomial Regression (PNR)	1.000	-60.203	0.000	1.310	0.000	25194.887	0.000	158.729	0.000	1.599875e+07

Gambar 5.1 Metrics Variabel Max (Y_1)

Tabel 5.1 Hasil prediksi Nilai Max

No.	Tanggal Prediksi	Actual Y	PrediksiY Ridge Max Kedepan	PrediksiY Multi Linear Max Kedepan
1	01/11/2021	53.00	52.73	52.73
2	02/11/2021	47.00	46.06	46.06
3	03/11/2021	58.00	60.40	60.40
4	04/11/2021	48.00	46.15	46.15
5	05/11/2021	53.00	53.18	53.18
6	06/11/2021	58.00	61.58	61.58
...	...	...	...	...
1504	25/08/2022	76.00	80.77	80.77
1505	26/08/2022	77.00	81.47	81.47
1506	27/08/2022	70.00	76.34	76.34
1507	28/08/2022	74.00	78.26	78.26
1508	29/08/2022	74.00	77.32	77.32

	Train_CV	Test_CV
Multiple Linear Regression (MLR)	0.543	0.548
Ridge Linear Regression (RLR)	0.543	0.548
Lasso Linear Regression(LLR)	0.246	0.278
Elastic-Net Regression (ENR)	0.375	0.399
Polynomial Regression (PNR)	-23.824	-6.270

	Train-R2	Test-R2	Train-MAPE	Test-MAPE	Train-MSE	Test-MSE	Train-RMSE	Test-RMSE	Train-RSS	Test-RSS
Multiple Linear Regression (MLR)	0.576	-1.075	0.125	0.278	0.087	0.468	0.294	0.684	54.929	296.990
Ridge Linear Regression (RLR)	0.576	-1.074	0.125	0.278	0.087	0.468	0.294	0.684	54.929	296.948
Lasso Linear Regression(LLR)	0.266	-0.447	0.137	0.206	0.150	0.326	0.387	0.571	95.063	207.136
Elastic-Net Regression (ENR)	0.397	-0.590	0.135	0.229	0.123	0.358	0.351	0.599	78.086	227.550
Polynomial Regression (PNR)	1.000	-217.745	0.000	2.264	0.000	49.309	0.000	7.022	0.000	31311.203

Gambar 5.2 Metrics Variabel Category (Y₂)

Tabel 5.2 Hasil prediksi Nilai Category

No.	Tanggal Prediksi	Actual Y	PrediksiY Ridge Category Kedepan	PrediksiY Multi Linear Category Kedepan
1	01/11/2021	2.00	1.58	1.58
2	02/11/2021	1.00	1.47	1.47
3	03/11/2021	2.00	1.72	1.72
4	04/11/2021	1.00	1.47	1.47
5	05/11/2021	2.00	1.60	1.60
6	06/11/2021	2.00	1.76	1.76
...	...	...	...	...
1504	25/08/2022	2.00	2.10	2.10
1505	26/08/2022	2.00	2.11	2.11
1506	27/08/2022	2.00	2.01	2.01
1507	28/08/2022	2.00	2.04	2.04
1508	29/08/2022	2.00	2.04	2.04

	Train CV	Test CV
Multiple Linear Regression (MLR)	-0.198	-0.186
Ridge Linear Regression (RLR)	-0.198	-0.186
Lasso Linear Regression (LLR)	-0.047	-0.082
Elastic-Net Regression (ENR)	-0.036	-0.086
Polynomial Regression (PNR)	-69.775	-30.669

	Train-R2	Test-R2	Train-MAPE	Test-MAPE	Train-MSE	Test-MSE	Train-RMSE	Test-RMSE	Train-RSS	Test-RSS
Multiple Linear Regression (MLR)	0.142	-0.067	0.105	0.103	0.234	0.251	0.483	0.501	148.059	159.334
Ridge Linear Regression (RLR)	0.142	-0.067	0.104	0.103	0.234	0.251	0.483	0.501	148.059	159.326
Lasso Linear Regression (LLR)	0.014	0.007	0.069	0.061	0.268	0.234	0.518	0.483	170.121	148.254
Elastic-Net Regression (ENR)	0.070	0.016	0.074	0.067	0.253	0.231	0.503	0.481	160.529	146.848
Polynomial Regression (PNR)	1.000	-234.160	0.000	1.960	0.000	55.285	0.000	7.435	0.000	35105.931

Gambar 5.3 Metrics Variabel Critical (Y₃)

Tabel 5.3 Hasil prediksi Nilai Critical

No.	Tanggal Prediksi	Actual Y	Prediksi Y Ridge Critical Kedepan	Prediksi Y Multi Linear Critical Kedepan
1	01/11/2021	2.00	2.54	2.54
2	02/11/2021	5.00	3.03	3.03
3	03/11/2021	2.00	2.63	2.63
4	04/11/2021	2.00	2.72	2.72
5	05/11/2021	2.00	2.54	2.54
6	06/11/2021	2.00	2.76	2.76
...	...	...	...	...
1504	25/08/2022	2.00	1.93	1.93
1505	26/08/2022	2.00	1.91	1.91
1506	27/08/2022	2.00	1.80	1.80
1507	28/08/2022	2.00	1.97	1.97
1508	29/08/2022	2.00	1.97	1.97

Berdasarkan Metrics Training yang didapatkan berdasarkan gambar 5.1, 5.2, dan 5.3 diperoleh hasil yang paling baik saat Training adalah Polynomial, namun saat Testing Metricsnya berbeda jauh yang sangat tinggi. Kemudian skor untuk R2 nilai yang cukup baik saat Training adalah Multi Linear dan Ridge. Yang kemudian metode tersebut juga mendapatkan hasil MAP, MSE, RMSE, dan RSS yang lebih rendah dibandingkan 2 metode lainnya yaitu Lasso dan Elastic Net. Oleh karena itu Setelah melakukan berbagai proses yang telah dilakukan sebelumnya, telah diperoleh hasil dari menggunakan metode apa yang cukup tepat dan cocok untuk kasus Dataset ini, dimana akan digunakan 2 metode untuk melakukan prediksi pada Data dummy untuk prediksi dalam 10 bulan kedepan. Metode yang akan digunakan adalah Multi Linear Regression dan Elastic Net Regression. Dikarenakan hasilnya yang sedikit mirip maka untuk melihat hasil Data akhirnya akan menggunakan metode tersebut. Dapat dilihat hasil prediksi Multi Linear dan Ridge tersebut pada tabel 5.1, 5.2, dan 5.3 dari hasil diperoleh hasil yang mirip antara kedua metode tersebut. Dimana prediksi untuk 10 bulan kedepan terlihat sedikit beberapa perbedaan dengan Data aktual yang dimana merupakan Data 10 bulan saat ini. Tetapi hasil terlihat untuk (Y1) Max akan didominasi nilai “79.99” yang didapatkan dari hasil averagenya, (Y2) Category akan didominasi nilai “2 = Sedang” yang didapatkan dari hasil averagenya yaitu 2.09. (Y3) Critical akan didominasi nilai “2 = PM2.5” yang didapatkan dari hasil averagenya yaitu 2,14.

	df	sum_sq	mean_sq	F	PR(>F)
PM10	1.0	535761.29386	535761.29386	14184.05950	0.00000
PM25	1.0	165565.69824	165565.69824	4383.28364	0.00000
SO2	1.0	1946.37935	1946.37935	51.52959	0.00000
CO	1.0	1890.37472	1890.37472	50.04689	0.00000
O3	1.0	4047.29562	4047.29562	107.15048	0.00000
NO2	1.0	1041.71658	1041.71658	27.57902	0.00000
Tanggal	1.0	46.03509	46.03509	1.21876	0.26978
Stasiun	1.0	8205.68137	8205.68137	217.24203	0.00000
Residual	1499.0	56620.33351	37.77207	NaN	NaN

Gambar 5.4 ANOVA Max (Y₁)

	df	sum_sq	mean_sq	F	PR(>F)
PM10	1.0	177.91717	177.91717	1891.96313	0.00000
PM25	1.0	58.97486	58.97486	627.13600	0.00000
SO2	1.0	0.89784	0.89784	9.54763	0.00204
CO	1.0	1.68533	1.68533	17.92171	0.00002
O3	1.0	0.81477	0.81477	8.66427	0.00330
NO2	1.0	1.28495	1.28495	13.66415	0.00023
Tanggal	1.0	0.02196	0.02196	0.23347	0.62903
Stasiun	1.0	2.23267	2.23267	23.74213	0.00000
Residual	1499.0	140.96355	0.09404	NaN	NaN

Gambar 5.5 ANOVA Category (Y₂)

	df	sum_sq	mean_sq	F	PR(>F)
PM10	1.0	70.74475	70.74475	192.28817	0.00000
PM25	1.0	0.24508	0.24508	0.66614	0.41453
SO2	1.0	0.83722	0.83722	2.27562	0.13163
CO	1.0	0.15606	0.15606	0.42417	0.51496
O3	1.0	124.53304	124.53304	338.48772	0.00000
NO2	1.0	1.10805	1.10805	3.01174	0.08287
Tanggal	1.0	3.90271	3.90271	10.60778	0.00115
Stasiun	1.0	0.10786	0.10786	0.29317	0.58827
Residual	1499.0	551.49719	0.36791	NaN	NaN

Gambar 5.6 ANOVA Critical (Y₃)

Uji coba hipotesis untuk melihat signifikansi antara variabel dependent terhadap variabel independent akan diperoleh hasil uji coba ANOVA yang didapatkan dari hasil metode *ordinary least squares (OLS)*. Dimana fungsi OLS ini memberikan beragam stats tentang statistic dari variabel yang diuji, yang kemudian diambil nilai F dan P-value. Dari hasil uji cob aini jika nilai tersebut diatas *Confidence Interval* yaitu 95% dengan toleransi kesalahan yaitu 5% (0.05). Jika $0.05 > H_0$, maka H_0 “Tidak berpengaruh terhadap variabel dependent”, sebaliknya maka H_1 yang akan diterima yaitu “Terdapat pengaruh terhadap variabel dependent”. Berdasarkan gambar 5.4, diperoleh hasil yang signifikan terhadap Max hanyalah Tanggal dengan nilai kurang lebih 0,27. Hal ini memberikan hasil H_1 diterima untuk Tanggal yaitu $0.27 > 0.05$. Dimana hal ini berarti terdapat tanggal-tanggal dimana nilai max akan menjadi sangat tinggi melebihi dari nilai rata-rata max. Berdasarkan gambar 5.5, diperoleh hasil yang hamper mirip dengan signifikan terhadap *Category* hanyalah Tanggal dengan nilai kurang lebih 0,63 lebih besar pengaruhnya terhadap Max. Hal ini memberikan hasil H_0 diterima untuk Tanggal yaitu $0.63 > 0.05$. Dimana hal ini berarti terdapat tanggal-tanggal dimana nilai *Category* akan menjadi menjadi sangat dominan terhadap suatu kondisi tanggal yang dimana berdasarkan hasil tabel 5.1, 5.2, dan 5.3 adalah category “Sedang”.

Berdasarkan gambar 5.6, diperoleh hasil yang signifikan terhadap Critical adalah PM_{2.5}, SO₂, CO, NO₂ dan Stasiun. Hal ini memberikan hasil H_0 diterima untuk variabel tersebut dengan nilai yaitu diatas 0,05. Dimana hal ini berarti gas dan partikel PM_{2.5}, SO₂, CO, dan NO₂ memiliki pengaruh yang terhadap perubahan *Critical* indeks

udara. Kemudian juga didapatkan pengaruh dari berbagai Stasiun, yang menunjukkan terdapat lokasi Stasiun yang memiliki dampak critical terhadap partikel gas tersebut.

Berdasarkan dataset Indeks Standar Pencemar Udara yang diperoleh, terdapat beberapa factor sebagai Atribut yang akan digunakan dalam penelitian ini yaitu sebagai berikut:

Tabel 5.4
Factor Atribut

Atribut	Keterangan
Tanggal	Tanggal saat pemeriksaan
PM ₁₀	Partikulat Udara 10 Mikrometer
PM _{2.5}	Partikulat Udara 2.5 Mikrometer
SO ₂	Gas Sulfur Dioksida
CO	Gas Karbon Monoksida
O ₃	Gas Ozon
NO ₂	Gas Nitrogen Dioksida
Max	Nilai Faktor tertinggi saat pemeriksaan
Critical	Gas atau Partikulat yang kritis
Category	Kategori udara berdasarkan pemeriksaan
Stasiun	Tempat pengambilan sample

Pembersihan data dilakukan terlebih dahulu, dengan menghilangkan data yang tidak memiliki nilai, hal tersebut dapat dilihat pada gambar 4.1 dan 4.2.

1	tanggal	stasiun	pm10	pm25	so2	co	o3	no2	max	critical	category
118	24/01/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
119	25/01/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
236	24/02/2021	DK03 (Jagakarsa)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
241	01/02/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
424	04/03/2021	DK05 (Kebon Jeruk) Jakarta Barat	---	---	---	---	---	---	0	---	TIDAK ADA DATA
498	17/04/2021	DK02 (Kelapa Gading)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
499	18/04/2021	DK02 (Kelapa Gading)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
1011	12/07/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
1325	19/09/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
1326	20/09/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
1436	08/10/2021	DK03 (Jagakarsa)	---	---	---	---	---	---	0	---	TIDAK ADA DATA
1464	05/10/2021	DK04 (Lubang Buaya)	---	---	---	---	---	---	0	---	TIDAK ADA DATA

Gambar 4.1 Data masih terdapat nilai kosong

68	05/01/2021	DK03 (Jagakarsa)	59	89	15	19	42	7	89	PM25	SEDANG
69	06/01/2021	DK03 (Jagakarsa)	52	81	18	28	52	11	81	PM25	SEDANG
70	07/01/2021	DK03 (Jagakarsa)	22	52	16	12	44	5	52	PM25	SEDANG
71	08/01/2021	DK03 (Jagakarsa)	26	68	21	22	40	9	68	PM25	SEDANG
72	09/01/2021	DK03 (Jagakarsa)	25	77	17	32	31	8	77	PM25	SEDANG
73	10/01/2021	DK03 (Jagakarsa)	15	33	16	7	23	4	33	PM25	BAIK
74	11/01/2021	DK03 (Jagakarsa)	15	72	17	15	56	4	72	PM25	SEDANG
75	12/01/2021	DK03 (Jagakarsa)	15	58	18	9	37	4	58	PM25	SEDANG
76	13/01/2021	DK03 (Jagakarsa)	26	47	16	6	39	3	47	PM25	BAIK
77	14/01/2021	DK03 (Jagakarsa)	29	78	20	14	40	8	78	PM25	SEDANG
78	15/01/2021	DK03 (Jagakarsa)	24	82	19	16	64	6	82	PM25	SEDANG
79	16/01/2021	DK03 (Jagakarsa)	70	92	19	16	60	7	92	PM25	SEDANG
80	17/01/2021	DK03 (Jagakarsa)	58	86	22	11	43	9	86	PM25	SEDANG
81	18/01/2021	DK03 (Jagakarsa)	51	64	21	7	38	5	64	PM25	SEDANG
82	19/01/2021	DK03 (Jagakarsa)	42	56	19	7	37	6	56	PM25	SEDANG

Gambar 4.2 Setelah dibersihkan

Hal ini bertujuan agar saat penelitian dapat dengan mudah untuk diolah. Kemudian setelah pembersihan data dilakukan perubahan format typedata, karena saat melakukan implementasi metode Linear Regression data hanya dapat dihitung jika memiliki typedata berformat integer atau numerik. Hal ini dapat dilihat pada gambar 4.3.

1	pm10	pm25	so2	co	o3	no2	max	Stasiun	Bulan	Critical	Category
2	38	53	29	6	31	13	53	1	1	2	2
3	27	46	27	7	47	7	47	1	1	5	1
4	44	58	25	7	40	13	58	1	1	2	2
5	30	48	24	4	32	7	48	1	1	2	1
6	38	53	24	6	31	9	53	1	1	2	2
7	41	58	23	13	46	13	58	1	1	2	2
8	35	47	22	6	39	10	47	1	1	2	1
9	37	54	26	16	17	10	54	1	1	2	2
10	47	61	16	27	22	12	61	1	1	2	2
11	23	25	16	11	33	8	33	1	1	5	1
12	38	54	17	14	27	10	54	1	1	2	2
13	29	50	20	12	26	15	50	1	1	2	1
14	34	44	17	13	20	9	44	1	1	2	1
15	36	57	15	15	26	13	57	1	1	2	2
16	42	62	14	15	32	12	62	1	1	2	2
17	52	73	15	17	25	14	73	1	1	2	2
18	52	71	17	15	26	14	71	1	1	2	2
19	40	52	16	11	26	12	52	1	1	2	2
20	31	41	13	12	25	12	41	1	1	2	1

Gambar 4.3 Konversi Typedata Numerik

Setelah keseluruhan data siap digunakan untuk diolah untuk dilakukan analisa maka, kemudian metode Linear Regrssion akan diimplementasikan menggunakan Bahasa *Data Mining & Analytic Forecasting Indeks Standar Pencemar Udara Jakarta Menggunakan Metode Linear Regression (Studi Kasus: Dataset Indeks Standar Pencemar Udara Jakarta 2021)*

Pemrograman Python. Dimana parameter yang akan ditetapkan adalah *Parameter* faktor (Y_1) *Critical*, (Y_2) *Max*, dan (Y_3) *Category* dan *Parameter* factor yaitu dari (X_1) PM_{10} , (X_2) $PM_{2.5}$, (X_3) SO_2 , (X_4) CO , (X_5) O_3 , (X_6) NO_2 , (X_7) Tanggal, dan (X_8) Stasiun. Akan terlihat metrics yang akan dilakukan pengukuran seperti R^2 , MSE, RMSE, MAPE, RSS, dan Cross Validation dari berbagai metode Regresi yang ditetapkan.

KESIMPULAN

Hasil dari seluruh perbandingan yang dilakukan ternyata didapatkan tidak terlalu jauh dari seperti nilai 10 bulan lalu, karena memang jika untuk memprediksi masa depan yang tepat adalah memiliki juga gambaran untuk variabel X. Pada penelitian ini hanya digunakan untuk uji coba penerapan Dataset terhadap fungsi prediksi dari Linear Regression Model yang berkaitan hubungan variabel Independent dan Dependen cenderung memberikan hasil yang positif. Oleh karena itu maka berarti faktor dari atribut Max, Category, dan Critical cukup dipengaruhi oleh variabel Independent. Dari penelitian ini didapatkan kesimpulan bahwa metode Regression yang cocok untuk melakukan prediksi pada studi kasus ini adalah metode Multi Linear dan Ridge dikarenakan hasil Metrics dari metode tersebut tidak terlalu jauh perbedaannya antara nilai validasi dan Test. Untuk metode Polynomial memiliki hasil yang sangat bagus tetapi memiliki perbedaan yang jauh dengan validasi dan Testing, hal ini kemungkinan dikarenakan faktor Multikolinearitas yang tinggi, karena Polynomial melakukan pemangkatan yang mana hal itu saling menambahkan ketergantungan antara prediktor dalam masalah regresi berdimensi ganda. Hal ini membatasi model menjadi tidak pas dengan benar pada kumpulan Data.

Berdasarkan hasil ANOVA, gas dan partikel $PM_{2.5}$, SO_2 , CO , dan NO_2 memiliki pengaruh yang terhadap perubahan Critical indeks udara. Kemudian juga didapatkan pengaruh dari berbagai Stasiun, yang menunjukkan terdapat lokasi Stasiun yang memiliki dampak Kritis terhadap partikel gas tersebut. Kemudian faktor yang berpengaruh terhadap Max dan Category adalah Tanggal yang berarti nilai Category akan menjadi sangat dominan terhadap suatu kondisi dan nilai max akan menjadi sangat tinggi melebihi dari nilai rata-rata max. Berdasarkan hasil uji coba tersebut didapatkan pengaruh yang kritis terhadap Partikulat yaitu $PM_{2.5}$. Hasil yang diperoleh nilai partikulat ini berada di dalam kategori sedang. Dalam suatu penelitian ([Elsa Try Julita Sembiring, 2020](#)) yang masih dalam lingkup suatu daerah di Jakarta mengatakan konsentrasi polutan $PM_{2.5}$ di lokasi penelitian masih memenuhi baku mutu ($<65 \mu g/m^3$). menurut Peraturan Pemerintah Republik Indonesia No. 41 tahun 1999 pada kondisi rata-rata $56,078 \mu g/m^3$ dan kondisi minimum $35,1 \mu g/m^3$ dan sudah melebihi baku mutu pada kondisi maksimum $116 \mu g/m^3$. Namun bila dibandingkan dengan National Ambient Air Quality Standards (NAAQS) US-EPA, kualitas udara sudah melewati baku mutu pada kondisi rata-ratanya. Berbagai material yang terkandung dalam $PM_{2.5}$ ini dapat menyebabkan berbagai gangguan saluran pernafasan seperti infeksi saluran pernafasan akut (ISPA), kanker paru-paru, kardiovaskular, kematian dini, dan penyakit paru-paru obstruktif kronis ([Novirsa et al., 2012; WHO, 2006](#)). Oleh karena itu pada hasil penelitian ini ternyata mendapatkan hasil yang tidak jauh berbeda dibandingkan penelitian sebelumnya dan kemungkinan dari prediksi diperoleh hasil yang memungkinkan selama 10 bulan ke depan juga diperkirakan partikulat ini masih dalam kondisi dengan potensi polusi sedang. Berdasarkan hal itu pemerintah ataupun pembuat kebijakan agar lebih memperhatikan hal tersebut untuk mengurangi berbagai partikel dan gas berbahaya, terutama partikulat $PM_{2.5}$.

BIBLIOGRAFI

- Nurjannah, Setyo Sri Rahardjo, dan Rossi Sanusi. (2019). "Linear Regression Analysis on the Determinants of Hypertension Prevention Behavior". <https://doi.org/10.26911/thejhp.2019.04.01.03>, e-ISSN: 2549-1172. [Google Scholar](#)
- Tri Indarwati, Tri Irawati dan Elistya Rimawati. (2018). "Penggunaan Metode Linear Regression untuk Prediksi Penjualan Smartphone", Jurnal TIKomSiN, Vol. 6, No. 2. ISSN Cetak : 2338-4018, DOI: <https://doi.org/10.30646/tikomsin.v6i2.369> ISSN Online : 2620-7532. [Google Scholar](#)
- T.L Wasilaine, M.W Talakua, dan Y.A Lenussa. (2014). "Model Regresi Ridge Untuk Mengatasi Model Regresi Linier Berganda Yang Mengandung Multikolinieritas". Barekeng Vol. 8 No. 1 Hal. 31 – 37. [Google Scholar](#)
- Ghebyla Najla Ayuni dan Devi Fitriana. (2019). "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ". Jurnal Telematika, Vol.14, No.2. Institut Teknologi Harapan Bangsa, Bandung p-ISSN: 1858-2516 e-ISSN: 2579-377. [Google Scholar](#)
- Stephen Chen, Khairil, Anwar Notodiputro, dan Septian Rahardiantoro. (2020). "Penerapan Analisis Lasso dan Group Lasso dalam Mengidentifikasi Faktor-faktor yang Berhubungan dengan Tuberkulosis di Jawa Barat". Indonesian Journal of Statistics and Its Applications, (eISSN:2599-0802) Vol.4, No.1 (2020), 39 – 54. [Google Scholar](#)
- Hizkia Edwar Sinaga dan Dewi Retno Sari Saputro. (2021). "Performa Metode Elastic-Net dalam Kasus Multikolinearitas pada Analisis Linear Berganda". e-ISSN No. 2721-6802. [Google Scholar](#)
- Lia Nurlia. (2020). "Pengaruh Lokasi Terhadap Keputusan Pembelian Pengunjung di Minimarket Menggunakan Metode Regresi Linier". Jursistekni Vol 2, ISSN. P: 2715-1875, E: 2715-1883. [Google Scholar](#)
- Jammaludin dan La Tommeng. (2018). "Pengaruh Teknologi Informasi terhadap Kualitas Layanan di UPT Perpustakaan Universitas Hasanuddin". Media Pustakawan, Vol. 25 No. 4. [Google Scholar](#)
- Nur Nafi'iyah dan Eka Rakhmawati. (2021). "Analisis Regresi Linear Dan Moving Average Dalam Memprediksi Data Penjualan Supermarket". ISSN: 2087-0868, Vol.12, No. 1. [Google Scholar](#)
- Andika Eka Putra dan Asep Juarna. (2021). "Prediksi Produksi Daging Sapi Nasional dengan Metode Regresi Linier dan Regresi Polinomial". Jurnal Ilmiah Komputasi, Vol.20, No.2, p-ISSN 1412-9434/e-ISSN 2549-7227. [Google Scholar](#)
- Ilyas Aji Prakoso, Kusnadi, dan Billy Nugraha. (2021). "Peramalan Penjualan Produk Dengan Metode Regresi Linear Dan Aplikasi Pom-Qm Di Pt Xyz". Scientific Journal Widya Teknik ,Vol.20, No.1, ISSN 1412-7350 eISSN 2621. [Google Scholar](#)

- Agus M Soleh dan Aunuddin. (2013). “Lasso: Solusi Alternatif Seleksi Peubah Dan Penyusutan Koefisien Model Regresi Linier”. Indonesian Journal of Statistics, Vol.18, No.1, April 2013, p: 21-27, ISSN 0853-8115. [Google Scholar](#)
- Elsa Try Julita Sembiring. (2020). “Risiko Kesehatan Paparan Pm_{2,5} Di Udara Ambien Pada Pedagang Kaki Lima Di Bawah Flyover Pasar Pagi Asemka Jakarta”, Jurnal Teknik Lingkungan Vol.26, No.1. 7.-Jurnal-TL-ITB-Elsa_SD_REV1-101-120. [Google Scholar](#)
- Denny Kurniawan. (2008), “Regresi Linier (Linear Regression), statistical computing. R Foundation for Statistical Computing”, Vienna, Austria. ISBN 3-900051-07-0. [Google Scholar](#)
- G. Shaddick, M. L. Thomas, P. Mudu, G. Ruggeri, dan S. Gumy. (2020). “Half The World’s Population Are Exposed To Increasing Air Pollution, npj Climate and Atmospheric Science”. 3:23, Half the world’s population are exposed to increasing air pollution | Science (nature.com). [Google Scholar](#)
- Foster Provost dan Tom Fawcett. (2013). “Data Science for Business”. First Edition, Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. [Google Scholar](#)
- Peter Bruce, Andrew Bruce, dan Peter Gedeck. (2020). “Practical Statistics for Data Scientist”. Second Edition, Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. [Google Scholar](#)
- Jagostat. “Analisis Regresi”. Analisis Regresi (jagostat.com) diakses pada tanggal 20 Januari 2022. [Google Scholar](#)
- Sklearn. “Linear Models User Guide”. 1.1. Linear Models — scikit-learn 1.0.2 documentation diakses pada tanggal 22 Januari 2022. [Google Scholar](#)
- Dinas Lingkungan Hidup Provinsi DKI Jakarta. (2021). “Laporan Pemantauan Kualitas Udara DKI Jakarta”. URL: Laporan Pemantauan Kualitas Udara | Dinas Lingkungan Hidup DKI Jakarta. [Google Scholar](#)
- Dasrul Chaniago, Annisa Zahara, Indah Suci Ramadhani. (2020). “Indeks Standar Pencemar Udara (Ispu) Sebagai Informasi Mutu Udara Ambien Di Indonesia”. URL: Portal Direktorat Pengendalian Pencemaran Udara Ditjen Ppkl Kllhk (Menlhk.Go.Id) diakses pada tanggal 6 Maret 2022. [Google Scholar](#)



© 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).