Perbandingan Kinerja Model Regresi Linier, LASSO, Ridge, dan Elastic Net pada Seoul *Bike Sharing Demand Dataset*

Muhamad Haris Hartanto[1], Alya Mirza Safira[2], Fiqih Pavita Andharluana[3], Hajjar Ayu Cahyani K.[4]

[1, 2, 3, 4]*Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur*

{[1]21083010045, [2]21083010039, [3]21083010042, [4]21083010044}@student.upnjatim.ac.id

Corresponding author email: 21083010045@student.upnjatim.ac.id

Abstract: *Bike sharing is a system where bicycles provided by the operator can be borrowed by users to be used within a specified time at an affordable cost. Over time, this system is growing rapidly and has become one of the main choices of transportation in many cities around the world, such as in the City of Seoul, South Korea. These developments have increased the demand for bicycle loans, so a prediction model is needed that can accurately predict the number of requests. Therefore, this study aims to create and compare the performance of ordinary linear regression models with regularized linear regression models, such as the Ridge, LASSO, and Elastic Net models by using a data set that contains historical data on the number of bicycles borrowed in the city of Seoul during the year 2017-2018 which is publicly available on the UCI Machine Learning Repository website. The results showed that the model with the best performance was the LASSO regression model with R-Square values for training and testing data of 0.700421 and 0.702623.*

Keywords: *Bike sharing, regression, linear regression, Ridge, LASSO, Elastic Net*

Abstrak: *Bike sharing* adalah sistem di mana sepeda yang disediakan oleh operator dapat dipinjam oleh pengguna untuk digunakan dalam waktu yang ditentukan dengan biaya yang terjangkau. Seiring dengan berjalannya waktu, sistem tersebut semakin berkembang pesat dan menjadi salah satu pilihan utama transportasi di banyak kota di seluruh dunia, seperti di Kota Seoul, Korea Selatan. Perkembangan tersebut membuat permintaan peminjaman sepeda menjadi meningkat sehingga dibutuhkan suatu model prediksi yang dapat memprediksi jumlah permintaan tersebut secara akurat. Oleh karena itu, pada penelitian ini bertujuan untuk membuat dan membandingkan kinerja model regresi linier biasa dengan model regresi linier yang diregularisasi, seperti model Ridge, LASSO, dan Elastic Net dengan menggunakan kumpulan data yang berisi data historis mengenai jumlah peminjaman sepeda di kota Seoul selama tahun 2017-2018 yang tersedia secara publik pada situs UCI Machine Learning Repository. Hasil penelitian menunjukkan bahwa model dengan kinerja yang terbaik adalah model regresi LASSO dengan nilai R-Square pada data *training* dan *testing* sebesar 0.700421 dan 0.702623.

Kata kunci: Peminjaman sepeda, regresi, regresi linier, Ridge, LASSO, Elastic Net

I. PENDAHULUAN

Bike sharing adalah sistem di mana sepeda yang disediakan oleh operator dapat dipinjam oleh pengguna untuk digunakan dalam waktu yang ditentukan dengan biaya yang terjangkau. Seiring dengan berjalannya waktu, sistem tersebut menjadi salah satu pilihan utama transportasi di banyak kota di seluruh dunia, seperti di Kota Seoul, Korea Selatan, di mana sistem tersebut telah berkembang pesat dalam beberapa tahun terakhir [1].

Sistem *bike sharing* yang semakin berkembang membuat permintaan peminjaman sepeda yang terus meningkat, membuat diperlukannya suatu model prediksi yang dapat memprediksi jumlah permintaan tersebut secara akurat. Berkaitan dengan sistem *bike sharing* di Kota Seoul, tersedia kumpulan data atau *dataset* yang berisi data historis mengenai jumlah peminjaman sepeda di kota tersebut selama tahun 2017-2018 dengan data waktu dan cuaca sebagai variabel prediktor [2]. Kumpulan data tersebut dapat diperoleh melalui situs UCI Machine Learning Repository. Berbagai penelitian telah dilakukan untuk mengembangkan model prediksi pada sistem *bike sharing* dengan menggunakan kumpulan data tersebut.

Sathishkumar V. E., dkk. melakukan penelitian dengan membuat dan membandingkan model regresi linier, Gradien Boosting Machine, Boosted Trees, Support Vector Machine, dan Extreme Gradient Boosting Tree dengan menerapkan teknik *data mining*. Pembuatan model regresi pada penelitian ini berfokus pada eksplorasi hubungan antara waktu dan jumlah sepeda yang dipinjam dengan menggunakan variabel jam sebagai sebuah variabel numerik. Rangkaian proses dari penelitian ini diawali dengan persiapan data, praproses data, eksplorasi data, pembuatan model regresi, dan diakhiri dengan pengevaluasian kinerja dari model regresi yang telah dibuat. Tahapan praproses data yang dilakukan pada penelitian ini meliputi pembersihan, pengintegrasian, transformasi data. Kinerja

dari model regersi yang telah dibuat dievaluasi dengan menggunakan metrik R-Square, root mean squared error (RMSE), mean absolute error (MAE), dan coefficient of variation (CV). Hasil dari penelitian ini diperoleh informasi model regresi dengan kinerja terbaik adalah model *Gradien Boosting Tree* dengan nilai R-Square pada data *training* dan *testing* sebesar 0.96 dan 0.92. Selain itu, juga ditunjukkan bahwa model regresi linier tidak memberikan hasil yang baik dengan nilai R-Square pada data training dan data testing yang hanya sebesar 0.55 [2].

Pada penelitian yang lain, kumpulan data Seoul *Bike Sharing Demand* juga diteliti oleh Syarif Hidayatulloh, dkk. Pada penelitian tersebut dibuat model prediksi dengan menggunakan metode Neural Nets, Generelized Linear Models, Support Vector Machines, Random Forests, dan Deep Learning, serta dilakukan percobaan penggunaan otimasi atribut untuk meningkatkan akurasi dari setiap model. Dilakukan 10 percobahan dengan validasi silang terhadap kelima model tersebut. Hasil dari percobaan yang telah dilakukan, diperoleh metode terbaik, yaitu Deep Learning dengan akurasi sebesar 90% yang berhasil ditingkatkan hingga 95.63% [3].

Seperti yang telah diuraikan di atas, beberapa peneliti telah melakukan pembuatan model regresi untuk memprediksi jumlah permintaan sepeda dengan menggunakan kumpulan data sistem *bike sharing* di Kota Seoul pada tahun 2017-2018. Oleh karena itu, pada penelitian ini dilakukan pengujian untuk membuat dan membandingkan model prediksi yang lebih sederhana, yaitu model regresi linier dan model regresi linier yang diregularisasi, seperti model Ridge, LASSO, dan Elastic Net pada kumpulan data yang sama dengan menggunakan pendekatan yang berbeda. Pendekatan yang dilakukan adalah dengan membuat model regresi yang berfokus pada pola musiman sehingga variabel jam (waktu) diubah menjadi bentuk kategorik. Selain itu, pada penelitian ini dilakukan serangkaian proses data mining yang dimulai dari praproses data (pembersihan dan transformasi), eksplorasi data, feature engineering, pembuatan model, dan pengevaluasian kinerja model yang telah dibuat. Hasil yang diharapkan dari penelitian ini adalah model regresi yang dibuat dapat memiliki kinerja yang baik untuk memprediksi jumlah permintaan peminjaman sepeda.

II. TINJAUAN PUSTAKA

2.1 Regresi Linier

Regresi linier merupakan suatu metode analisis data yang digunakan untuk memprediksi nilai variabel yang tidak diketahui menggunakan nilai variabel lain yang terkait dan diketahui. Regresi linier terbagi menjadi dua jenis, yaitu regresi linier sederhana dan regresi linier berganda. Regresi linier sederhana, hanya memiliki satu variabel bebas dan satu variabel terikat. Sementara itu, regresi linier berganda memiliki lebih dari satu variabel bebas dan satu variabel terikat. Persamaan regresi linier berganda dapat dituliskan sebagai berikut [4]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \quad (1)$$

Keterangan:

Y : nilai variabel terikat (dependen)

β_0 : intercept (perbedaan besarnya rata-rata variabel Y ketika nilai variabel $X = 0$)

$\beta_1, \beta_2, \ldots, \beta_n$: koefisien regresi (perkiraan besarnya perubahan nilai variabel Y jika nilai variabel X berubah satu unit pengukuran)

$X_1, X_2, \ldots, X_n$: nilai variabel bebas (independen)

ε : komponen error random yang saling bebas (diasumsikan berdistribusi normal dengan $E(\varepsilon_i) = 0$ dan varians konstan $Var(\varepsilon_i) = \sigma^2$)

2.2 Regresi Ridge

Regresi Ridge adalah teknik yang digunakan untuk menstabilkan nilai koefisien regresi dalam kondisi multikolinieritas. Teknik ini merupakan modifikasi dari metode ordinary least square (OLS). Regresi Ridge menghasilkan estimasi koefisien yang lebih bias tetapi memiliki varians yang lebih kecil daripada OLS. Perbandingan mean square error (MSE) antara penduga Ridge dengan penduga OLS menunjukkan bahwa penduga Ridge mendekati nilai parameter yang sebenarnya. Regresi Ridge mirip dengan OLS dalam hal meminimumkan sum of squares error (SSE) pada pendugaan koefisien regresi, tetapi terdapat tambahan penalti penyusutan dalam persamaan untuk mengurangi dampak multikolinieritas yang dituliskan sebagai berikut [5]:

$$SS_E(\beta) + \lambda \sum_{j=1}^{p} \beta_j{}^2 \quad (2)$$

Keterangan:
λ : parameter penyusutan
$SS_E(\beta)$: $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$
$\sum_{j=1}^{p} \beta_j^2$: penalti penyusutan

2.3 *Least Absolute Shrinkage and Selection Operator* (LASSO)

LASSO merupakan metode penyusutan yang serupa dengan regresi ridge yang digunakan untuk menstabilkan nilai koefisien regresi dalam kondisi multikolinieritas. Regresi ridge menyusutkan penduga koefisien regresi tetapi tidak dapat melakukan seleksi variabel secara otomatis, sementara LASSO dapat menyusutkan beberapa koefisien dan mengatur yang lain menjadi 0, sehingga dapat menghasilkan model dengan variabel penjelas yang lebih sedikit dan sederhana. LASSO dan Regresi Ridge memiliki persamaan yang serupa, dengan penalti penyusutan yang berbeda. Penalti penyusutan pada regresi ridge diganti dengan $\sum_{j=1}^{p}|\beta_j|$ pada LASSO seperti yang ditunjukkan oleh persamaan berikut:

$$SS_E(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

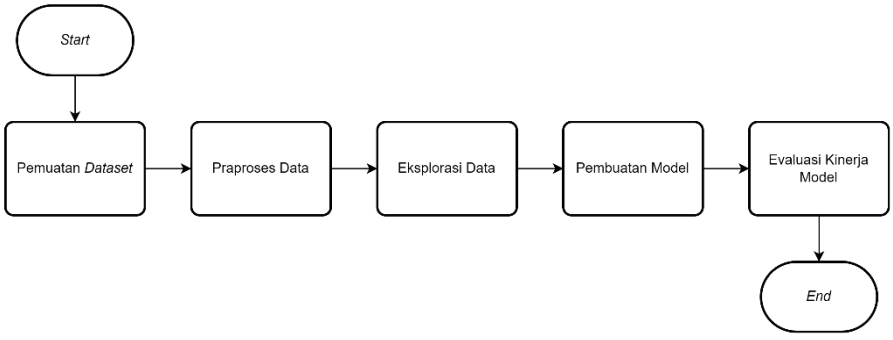
Hal tersebut menyebabkan variabel prediktor yang kurang penting disusutkan sampai nol dan terseleksi dari model sehingga model menjadi lebih efisien [5].

2.4 Elastic Net

Elastic Net merupakan metode penyeleksi variabel inovatif untuk estimasi regresi linear dengan menggabungkan dua metode yaitu regularisasi model Ridge dan regularisasi model LASSO dengan menggunakan nilai rasio campuran *r* yang dapat dikontrol. Saat *r* bernilai 0, regresi ini akan sama dengan regresi Ridge sedangkan saat *r* bernilai 1, maka akan sama dengan Regresi LASSO seperti yang ditunjukan oleh rumus berikut [6]:

$$SS_E(\beta) + SS_E(\beta) + r\,\lambda \sum_{j=1}^{p}|\beta_j| + (1-r)\lambda \sum_{j=1}^{p} \beta_j{}^2 \quad (4)$$

III. METODE PENELITIAN



Gambar 1. Diagram Alir *Proses* Penelitian

Tahapan proses yang dilakukan pada penelitian ini ditunjukkan oleh diagram alir yang ditunjukkan oleh Gambar 1. Proses diawali dengan pembacaan atau pemuatan dataset. Selanjutnya, dilakukan praproses data untuk mendapatkan data yang bersih. Setelah didapatkan data yang bersih, dilanjutkan dengan pembuatan model regresi linier. Proses diakhiri dengan evaluasi kinerja dari model regresi yang telah dibuat. Seluruh proses tersebut dilakukan dengan menggunakan bahasa pemrograman Python.

3.1. *Dataset*

Kumpulan data (*dataset*) yang digunakan pada penelitian ini adalah Seoul *Bike Sharing Demand Dataset*. Kumpulan data tersebut tersedia secara publik pada situs UCI Machine Learning Repository dan dapat diakses melalui tautan berikut: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand. *Dataset* tersebut terdiri dari 8760 *record* dengan 13 atribut dan 1 variabel target yang mencakup informasi tentang tanggal dan waktu, kondisi cuaca, serta informasi tentang jumlah sepeda yang disewa dalam satu jam. Rincian mengenai atribut dari *dataset* tersebut tercantum dalam tabel berikut [2]:

Tabel 1. Rincian Atribut pada *Dataset*

Atribut	Tipe Data	Deskripsi
Date	Teks	Tanggal (hari/bulan/tahun)
Rented Bike Count	Numerik	Jumlah sepeda yang disewa
Hour	Numerik	Waktu (jam) dalam satu hari
Temperature (°C)	Numerik	Suhu udara dalam skala Celcius (°C)
Humidity (%)	Numerik	Kelembaban relatif udara dalam satuan persen (%)
Windspeed (m/s)	Numerik	Kecepatan angin dalam satuan meter per detik (m/s)
Visibility (10m)	Numerik	Jarak pandang (10 m)
Dew Point Temperature (°C)	Numerik	Suhu titik embun dalam skala Celcius (°C)
Solar Radiation (MJ/m2)	Numerik	Radiasi matahari dalam satuan Megajoule per meter persegi (MJ/m^2)
Rainfall (mm)	Numerik	Curah hujan dalam satuan milimeter (mm)
Snowfall (cm)	Numerik	Salju yang turun dalam satuan centimeter (cm)
Seasons	Kategorik	Musim dalam setahun (Winter, Spring, Summer, Autumn)
Holiday	Kategorik	Status hari libur (No Holiday/Holiday)
Functional Day	Kategorik	Status hari fungsional (Yes/No)

3.2. Praproses Data

Praproses data atau *preprocessing data* adalah serangkaian proses yang diterapkan terhadap kumpulan data (*dataset*) dengan tujuan mendapatkan data yang bersih untuk digunakan pada tahapan proses selanjutnya [7]. Pada tahap ini, dilakukan pembersihan dan transformasi data. Pembersihan data yang dilakukan meliputi perbaikan format penulisan, perubahan tipe data, serta pemeriksaan, pengisian atau penghapusan nilai data yang tidak valid. Sementara itu, proses transformasi yang dilakukan adalah penambahan atribut baru yang berasal dari nilai atribut yang telah ada dan normalisasi nilai data.

3.2.1. Pembersihan Data

a. Perbaikan Format Penulisan

Pada *dataset* yang digunakan, terdapat sejumlah nama atribut yang dituliskan beserta dengan satuan pengukurannya, seperti “Temperature (°C)”. Penulisan nama atribut dengan format tersebut akan menyulitkan proses lainnya yang memerlukan penyertaan nama atribut, seperti proses manipulasi data dengan teknik *slicing* DataFrame. Oleh karena itu, dilakukan perubahan format penulisan nama atribut menjadi karakter dengan huruf kecil (*lowercase*) dengan spasi yang diganti dengan karakter *underscore*, serta menghapus karakter “(...)” beserta dengan karakter di dalamnya. Perbedaan sebelum dan sesudah perubahan ditunjukkan oleh Tabel 2.

Tabel 2. Perubahan Nama Atribut

Nama Atribut

Sebelum	Sesudah
Date	date
Rented Bike Count	rented_bike_count
Hour	hour
Temperature (°C)	temperature
…	…
Snowfall (cm)	snowfall
Seasons	seasons
Holiday	holiday
Functional Day	functional_day

b. Perubahan tipe data

Dilakukan perubahan tipe data untuk atribut "hour" dari yang sebelumnya berupa numerik menjadi kategorik. Perubahan ini dilakukan untuk menyesuaikan dengan tujuan model regresi linier yang akan dibuat, di mana berfokus pada pola musiman (hari, minggu, bulan) sehingga atribut "Hour" lebih cocok dengan tipe kategorik.

c. Pemeriksaan Nilai Data

Pemeriksaan nilai data yang dilakukan meliputi identifikasi nilai kosong (*missing value*) dan *outlier*. Nilai kosong (*missing value*) yang terdapat dalam *dataset*, apabila dibiarkan dan tidak ditangani dengan tepat dapat memengaruhi kinerja model regresi dengan menghasilkan pendugaan parameter yang tidak efisien karena berkurangnya ukuran data [8]. Nilai kosong (*missing value*) dapat diidentifikasi dengan menggunakan *method* isnull() terhadap DataFrame data dan didapatkan jumlah nilainya dengan menggunakan *method* sum(). Setelah dilakukan pendeteksian dengan menggunakan kedua *method* tersebut, tidak ditemukan adanya nilai kosong (*missing value*) pada *dataset* yang digunakan.

Sementara itu, *outlier* merupakan observasi atau nilai yang secara signifikan berbeda dari nilai-nilai lain pada variabel atau atribut yang sama dalam suatu *dataset*. Perbedaan tersebut bisa ditimbulkan oleh kesalahan pengukuran, kegagalan saat proses pengumpulan data, atau karakteristik unik dari sebuah sampel [9]. *Outlier* dapat memengaruhi hasil estimasi parameter regresi dan dapat menyebabkan sisaan yang besar dari model yang dibuat [10]. Namun, *outlier* tidak selamanya buruk. Dalam studi kasus nyata, terkadang ditemukan nilai data yang jauh berbeda dengan nilai lainnya, tetapi nilai tersebut wajar dan masuk akal. Dilakukan identifikasi *outlier* terhadap seluruh nilai pada atribut dengan tipe numerik menggunakan pendekatan statistik menggunakan metode *Interquartile Range* (IQR) dengan tahapan proses sebagai berikut:

1. Menghitung nilai kuartil ke-1 (Q_1) dan kuartil ke-3 (Q_3)
2. Menghitung nilai IQR dengan mengurangi nilai Q_3dengan nilai Q_1
3. Menghitung batas bawah dan batas atas *outlier* yang terbagi menjadi dua kategori, yaitu:
   - *Minor Outlier*

$$Q_1 - 1.5\ IQR\ s/d\ Q_3 + 1.5\ IQR \tag{5}$$

   - *Major Outlier*

$$Q_1 - 3\ IQR\ s/d\ Q_3 + 3\ IQR \tag{6}$$

   Dalam penerapannya kategori nilai batas *outlier* dipilih menyesuaikan dengan kebutuhan penelitian.
4. Mengidentifikasi nilai data yang terletak di luar batas atas dan batas bawah *outlier*

Seluruh atribut dengan tipe numerik yang diidentifikasi berisi informasi mengenai kondisi cuaca. Berdasarkan identifikasi yang telah dilakukan, didapatkan sejumlah 1773 nilai data yang termasuk dalam kategori *minor outlier*. Setelah itu, dilakukan validasi nilai *outlier* yang diperoleh dengan membandingkan nilai tersebut dengan nilai pengamatan kondisi cuaca di lokasi dan pada waktu yang sesuai. Setelah dilakukan validasi, diketahui bahwa seluruh nilai *outlier* yang teridentifikasi merupakan nilai yang wajar untuk kondisi cuaca di lokasi dan waktu yang sesuai dari setiap nilai. Oleh karena itu, nilai outlier yang teridentifikasi tersebut tidak akan dihapus dan tetap digunakan.

3.2.2. Transformasi Data

Transformasi data dapat diartikan sebagai sebuah proses kreatif pengolahan data dan bergantung pada jenis atau pola informasi dari *dataset*. Pada tahap ini dilakukan penambahan atribut baru dengan menggunakan nilai dari atribut yang sudah ada. Penambahan atribut tersebut dilakukan dengan ekstraksi nama hari dan bulan dari atribut "date" yang kemudian disimpan ke dalam atribut baru, yaitu "day" dan "month". Setelah itu, dilakukan kembali penambahan atribut baru, yaitu "week" yang merupakan kategori dari atribut "date" dengan nilai yang berupa hari kerja (*weekdays*) atau akhir pekan (*weekend*). Atribut tersebut diperoleh dengan menerapkan percabangan IF-ELSE terhadap nilai atribut "day" dengan kondisi IF berupa "jika hari adalah sabtu (*saturday*) atau minggu (*sunday*) , maka diperoleh nilai *weekend*" dan kondisi ELSE diperoleh nilai *weekdays*. Setelah didapatkan ketiga atribut tersebut, dilakukan penghapusan terhadap atribut "date" karena tidak dapat digunakan dalam pembuatan model regresi.

Selain itu, pada tahap ini juga dilakukan normalisasi data untuk menskalakan dan menyeimbangkan nilai data pada setiap atribut dengan tipe numerik. Hal tersebut dilakukan karena pada pembuatan model regresi dengan model yang diregularisasi (Ridge, Lasso, Elastic Net) sensitif terhadap skala dari nilai atribut independen [6]. Normalisasi data juga dapat memberikan kenaikan kinerja dan stabilitas dari suatu model *machine learning*. Pada penelitian ini, dilakukan normalisasi data dengan menggunakan metode Min-Max yang dituliskan dalam rumus perhitungan sebagai berikut:

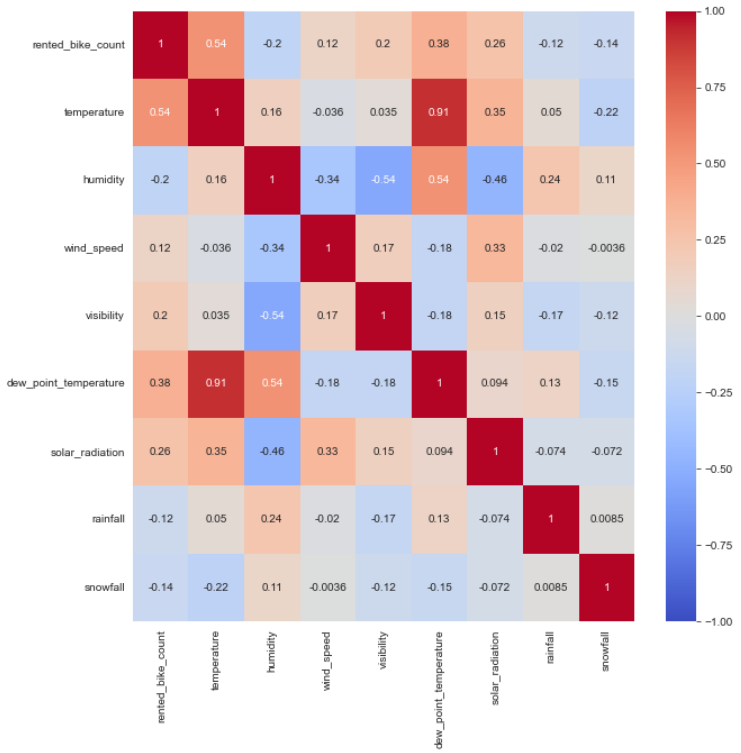
$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} (new_{max} - new_{min}) + new_{min} \quad (7)$$

Keterangan:
X_{norm} : nilai hasil normalisasi
x : nilai data
x_{min} : nilai data minimum di dalam atribut
x_{max} : nilai data maksimum di dalam atribut
new_{max} : nilai maksimal dari rentang nilai normalisasi
new_{min} : nilai minimum dari rentang nilai normalisasi

Normalisasi data dengan metode tersebut diterapkan menggunakan bantuan modul MinMaxScaler yang terdapat dalam pustaka scikit-learn dengan parameter nilai new_{max} dan new_{min} bawaan, yaitu sebesar 1 dan 0.

3.3. Eksplorasi Data

Eksplorasi data merupakan suatu proses yang berkaitan dengan pemahaman terhadap suatu data yang diteliti [7]. Pemahaman tersebut dapat diperoleh melalui statistika deskriptif dan teknik visualisasi sederhana menggunakan parameter yang berbeda untuk meringkas data. Gambar 2 menampilkan grafik *heatmap* nilai korelasi antar atribut tipe numerik dengan variabel target yang dibentuk dengan menggunakan fungsi heatmap() yang tersedia dalam modul *seaborn*. Sementara itu, nilai korelasi diperoleh melalui penggunaan fungsi corr() yang terdapat dalam modul pandas. Secara bawaan perhitungan korelasi dari fungsi tersebut menggunakan metode Pearson.



Gambar 2. Korelasi antar Atribut Numerik dan Variabel Target

Berdasarkan grafik *heatmap* korelasi yang ditampilkan Gambar 2, diketahui bahwa terdapat korelasi positif antara "rented_bike_count" dan "temperature" sebesar 0.54. Hal tersebut mengindikasikan permintaan sewa sepeda meningkat ketika suhu meningkat. Korelasi positif lainnya juga terlihat antara "rented_bike_count" dengan "wind_speed", "visibility", "dew_point_temperature", dan solar_radiation". Terdapat juga korelasi negatif antara "rented_bike_count" dengan "humidity", "rainfall", dan "snowfall" yang menandakan bahwa kelembaban, curah hujan dan salju adalah faktor yang memengaruhi permintaan sewa sepeda berkurang ketika nilai persentase kelembaban, curah hujan, atau salju meningkat.

Sementara itu, nilai korelasi lainnya merupakan korelasi antara atribut kondisi cuaca. Terdapat korelasi positif yang kuat antara atribut "temperature" dengan "dew_point_temperature" yang mengakibatkan multikolinieritas. Hal yang umum dilakukan apabila terdapat multikolinearitas adalah menghapus salah satu atribut yang berkaitan. Namun, karena pada penelitian ini akan dibuat model regresi linier dengan regularisasi yang tidak terpengaruh oleh multikolinieritas, maka atribut yang memiliki hubungan multikolinieritas tetap dipertahankan untuk dibandingkan kinerjanya dengan model regresi linier biasa.

3.4. *Feature Engineering*

Feature engineering merupakan proses mendapatkan serangkaian *feature* atau atribut bagus (sesuai) untuk dilatih dalam model *machine learning* [6]. Pada tahap ini dilakukan transformasi atribut dengan tipe kategorik menjadi variabel atau atribut *dummy*. Variabel atau atribut *dummy* merupakan suatu variabel yang diterapkan untuk mengubah data yang bersifat kualitatif menjadi kuantitatif dengan menggunakan nilai 1 dan 0. Masing-masing dari nilai tersebut menunjukkan keberadaan (*presence*) atau ketidakberadaan (*absence*) dari kualitas atau suatu variabel atau atribut [11].

Terdapat aturan umum dari pembentukan variabel *dummy*, yaitu jumlah variabel yang dibentuk adalah sebanyak nilai kategori dikurangi satu. Hal tersebut dilakukan karena satu variabel yang tidak dibentuk tersebut akan ditangkap oleh intersep dan ditentukan ketika semua variabel *dummy* dengan jenis yang sama bernilai 0. Ketika dibentuk atribut *dummy* dengan jumlah yang sama dengan nilai kategorinya, maka akan menyebabkan kegagalan regresi karena ada terlalu banyak parameter untuk diestimasi ketika intersep juga disertakan [12].

Transformasi variabel *dummy* dilakukan dengan menggunakan fungsi get_dummies() yang terdapat pada modul pandas dengan menggunakan parameter "drop_first=True" untuk mendapatkan sejumlah $k - 1$ dari *k* nilai kategori dengan menghapus variabel *dummy* kategori pertama.

3.5. Model Regresi

Pada penelitian ini, dibuat model regresi linier, regresi Ridge, regresi LASSO, dan regresi Elastic Net. Pembuatan dan pelatihan model regresi linier, Ridge, dan LASSO dapat dilakukan dengan menggunakan bantuan modul yang disediakan oleh pustaka scikit-learn dengan rasio pembagian data *training* dan *testing* sebesar 75:25. Selain itu, juga diterapkan*hyperparameter tuning* untuk mendapatkan parameter terbaik yang memberikan tingkat eror yang lebih kecil atau akurasi yang lebih baik untuk model regresi Ridge, LASSO, dan Elastic Net. Teknik *hyperparameter tuning* yang digunakan pada penelitian ini adalah *grid search*. Penggunaan eknik tersebut memungkinakan untuk menguji beberapa nilai parameter yang diinisialisasikan secara sekaligus pada suatu model.

3.6. Evaluasi Kinerja Model

Evaluasi kinerja dari seluruh model regresi linier yang telah dibuat dilakukan dengan menggunakan 2 metrik, yaitu nilai *root mean square error* (RMSE) dan R-square. RMSE merupakan metrik yang sering digunakan untuk mengukur perbedaan antara nilai prediksi dengan nilai yang yang sebenarnya dalam pengukuran kesalahan prediksi model. Secara matematis, RMSE didefinisikan sebagai akar dari rata-rata kuadrat selisih antara nilai prediksi dan nilai sebenarnya yang dituliskan dalam rumus sebagai berikut [13]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}} \tag{8}$$

Keterangan:
$\widehat{Y_i}$: nilai sebenarnya
Y_i : nilai prediksi
n : jumlah data

Sementara itu, R-square atau koefisien determinasi merupakan ukuran statistik yang digunakan untuk mengevaluasi seberapa baik model regresi dengan data yang Nilai R-square berkisar antara 0 dan 1, dengan nilai 1 menunjukkan bahwa model regresi memberikan penjelasan yang sangat baik terhadap variasi variabel terikat, sedangkan nilai 0 menunjukkan bahwa model tidak memberikan penjelasan sama sekali terhadap variasi variabel terikat [14].Metrik ini dihitung dengan menggunakan rumus sebagai berikut:

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y}_i)^2} \tag{9}$$

Keterangan:
SS_R: nilai sebenarnya (*sum of squares regression*)
SS_T: jumlah kuadrat total (*sum of squares total*)

Dilakukan penghitungan untuk mendapatkan nilai kedua metrik tersebut terhadap data latih dan uji. Hal tersebut dilakukan agar nilai metrik dari setiap bagian data dapat dibandingkan untuk melihat apakah model yang dibuat *underfir* atau *overfit*. Penghitungan nilai RMSE dan R-square dilakukan dengan menggunakan modul mean_squared_error dan r2_score yang tersedia dalam pustaka scikit-learn. Pada penggunaan modul mean_squared_error ditambahkan parameter berupa "square=False" untuk mendapatkan nilai RMSE.

IV. HASIL DAN PEMBAHASAN

Evaluasi model regresi dilakukan untuk mengukur seberapa akurat model dalam memprediksi variabel respon (*Y*) dengan menggunakan variabel prediktor (*X*), serta untuk mencari model regresi yang paling baik dalam melakukan prediksi jumlah permintaan peminjaman sepeda pada kumpulan data Seoul *Bike Sharing Demand*.

4.2 Hasil

Berdasarkan pengujian yang telah dilakukan, didapatkan hasil evaluasi kinerja dari model regresi linier, Ridge, LASSO dan Elastic Net yang dibuat dengan menggunakan kumpulan data Seoul *Bike Sharing Demand* dengan rasio pembagian data training dan testing sebesar 75:25 yang disajikan dalam Tabel 3.

Tabel 3. Hasil Evaluasi Model Regresi

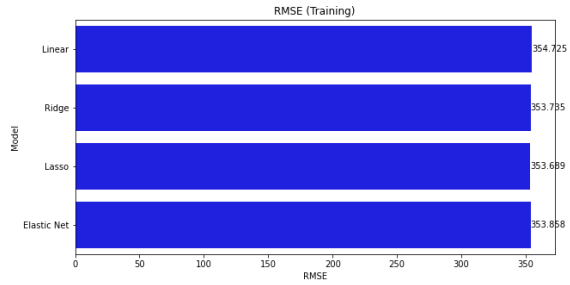
Model	Data *Training*		Data *Testing*	
	RMSE	***R-square***	**RMSE**	***R-square***
Linier	354.725006	0.698664	351.240580	0.699927
Ridge	353.734668	0.700344	350.093389	0.701884
LASSO	353.689207	0.700421	349.659031	0.702623
Elastic Net	353.858377	0.700134	350.490746	0.701207

Berdasarkan Tabel 3, didapatkan bahwa nilai RMSE data training dari model linier sebesar 354.725006 dengan nilai RMSE data testing sebesar 351.240580 dan nilai R-square data training sebesar 0.698664 dengan nilai R-square data testing sebesar 0.699927. Pada model regresi Ridge didapatkan bahwa nilai RMSE data training sebesar 353.734668 dengan nilai RMSE data testing sebesar 350.093389 dan nilai R-square data training sebesar 0.700344 dengan nilai R-square data testing sebesar 0.701884. Dari model LASSO didapatkan bahwa nilai RMSE data training sebesar 353.689207 dengan nilai RMSE data testing sebesar 349.659031 dan nilai R-square data training sebesar 0.700421 dengan nilai R-square data testing sebesar 0.702623. Sementara itu, pada model Elastic Net didapatkan bahwa nilai RMSE data training sebesar 353.858377 dengan nilai RMSE data testing sebesar 350.490746 dan nilai R-square data training sebesar 0.700134 dengan nilai R-square data testing sebesar 0.701207

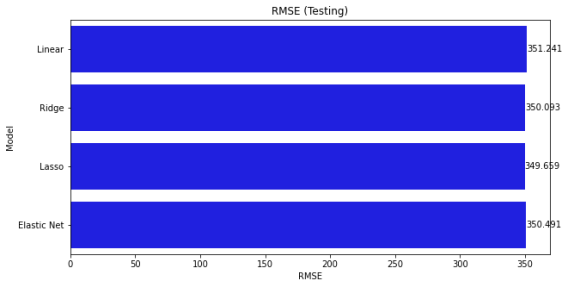
4.2 Pembahasan

Pada penelitian ini, penulis tertarik untuk membandingkan model regresi Elastic Net dengan model regresi Ridge dan LASSO. Model regresi Elastic Net dipilih karena merupakan gabungan dari kedua model tersebut dan diharapkan dapat memberikan solusi yang lebih stabil dan *robust*. Penulis juga ingin melakukan perbandingan kinerja dari regresi linier biasa dengan regularisasi model dari regresi linier.

Regresi Elastic Net dapat memperbaiki kelemahan masing-masing dari model regresi Ridge dan LASSO sehingga dapat memberikan solusi yang lebih stabil dan *robust*. Namun, pada studi kasus yang diujikan, didapatkan nilai RMSE dan R-square yang menunjukkan bahwa kinerja model regresi Elastic Net tidak lebih baik dibandingkan dengan regresi Ridge dan LASSO secara terpisah. Di sisi lain, model regresi linier biasa menunjukkan kinerja yang baik dengan nilai R-square dan RMSE sebesar 0.699927 dan 351.240580. Kinerja tersebut berbeda dengan hasil yang didapatkan pada penelitian terdahulu yang dilakukan oleh Sathishkumar V. E., dkk. dengan diperoleh nilai R-square yang hanya sebesar 0.55 [2]. Perbedaan hasil kinerja dari model regresi linier tersebut disebabkan oleh pendekatan pengolahan data yang berbeda. Pada penelitian ini, variabel "hour" diubah menjadi kategorik dari yang sebelumnya numerik. Grafik dari hasil evaluasi kinerja model regresi dengan menggunakan nilai R-square dan RMSE ditampilkan pada Gambar 3 dan Gambar 4.



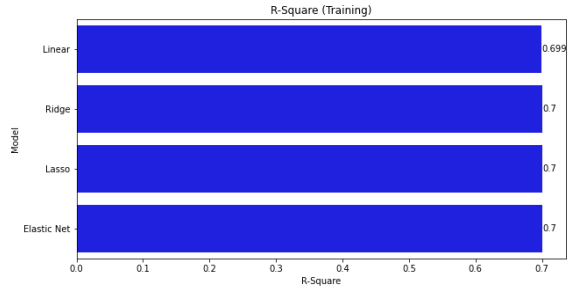
Gambar 3a. Grafik Nilai RMSE Data *Training*



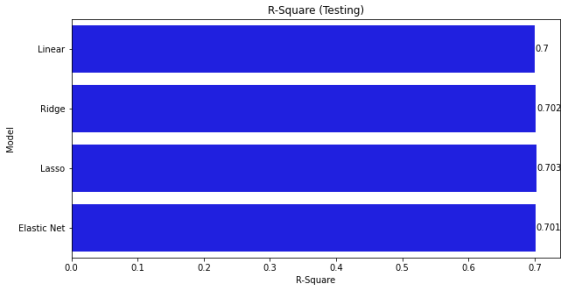
Gambar 3b. Grafik Nilai RMSE Data *Testing*

Berdasarkan Gambar 3a dan Gambar 3b, diperoleh selisih nilai RMSE antara data testing dan data training pada model regresi linier sebesar 3.4844. Pada model regresi Ridge, selisih antara nilai RMSE data testing dan data training adalah sebesar 3.641279, sedangkan pada model regresi LASSO selisih

antara nilai RMSE data training dan data testing adalah sebesar 4.030176. Pada model regresi Elastic Net didapatkan selisih nilai RMSE antara data testing dan data training sebesar 3.367631. Keempat model regresi tersebut menunjukkan selisih yang relatif kecil antara data training dan data testing sehingga didapatkan hasil bahwa model-model tersebut mempunyai kinerja yang baik dalam mempelajari pola-pola dari data dan mampu menggeneralisasi dengan baik pada data baru.



Gambar 4a. Grafik Nilai R-square Data *Training*



Gambar 4b. Grafik Nilai R-square Data *Testing*

Berdasarkan gambar 4a dan Gambar 4b, didapatkan selisih nilai R-square data training dan data testing dari model regresi linier sebesar 0.001283. Pada model regresi Ridge selisih antara nilai R-square data testing dan data training sebesar 0.00154. Model regresi LASSO menghasilkan selisih antara R-square pada data testing dan data training sebesar 0.002202. Sementara itu, pada model regresi Elastic Net didapatkan selisih nilai R-square data training dan data testing sebesar 0.001073. Keempat model regresi tersebut tidak memiliki selisih yang besar antara data training dan data testing sehingga didapatkan hasil bahwa tidak ada model yang underfit maupun overfit dan dapat menggeneralisasi dengan baik pada data baru.

Pada penelitian ini, parameter yang digunakan untuk menguji keakuratan suatu model adalah nilai R-square dan RMSE. Kedua nilai tersebut digunakan karena nilai R-square menggambarkan seberapa baik model regresi linier yang cocok dengan data yang diamati, sementara nilai RMSE menggambarkan seberapa jauh prediksi model dari nilai aktual pada data yang diuji. Oleh karena itu, model regresi yang memiliki nilai R-square tertinggi dan RMSE paling rendah adalah model yang paling cocok. Berdasarkan keempat model regresi yang diujikan, model LASSO merupakan model terbaik karena model ini memiliki nilai R-Square yang paling tinggi serta RMSE yang paling rendah diantara ketiga model lainnya. Oleh sebab itu, maka model regresi LASSO cocok untuk digunakan memprediksi jumlah permintaan sewa sepeda pada kumpulan data Seoul Bike Sharing Demand. Secara teori, model regresi LASSO dikatakan cocok digunakan karena model tersebut dapat menangani variabel independen yang berkorelasi atau terjadi multikolinieritas

Dari penelitian yang telah dilakukan, didapatkan bahwa keempat model regresi yang diuji telah optimal karena memiliki nilai R-square diatas 0.5 atau 50% dan memiliki nilai RMSE yang rendah. Namun, dari keempat model tersebut, model yang paling optimal adalah model regresi LASSO. Selain itu, juga terbukti dari hasil evaluasi model dengan menggunakan parameter R-square dan RMSE pada didapatkan bahwa model Elastic Net tidak lebih baik dari model Ridge dan LASSO. Namun, tidak dipungkiri bahwa model Elastic net akan lebih baik dari model Ridge dan LASSO jika evaluasi dengan parameter yang berbeda.

V. KESIMPULAN

Berdasarkan hasil evaluasi kinerja yang telah dilakukan, dapat disimpulkan bahwa model regresi linier terbaik pada studi kasus Seoul bike sharing demand dataset adalah dengan model regresi LASSO. Model tersebut terpilih karena diperoleh nilai R-square yang paling tinggi dan nilai RMSE terendah di antara keempat model regresi yang diuji. Nilai R-square yang semakin tinggi menunjukkan bahwa model tersebut semakin baik. Sementara itu, selisih nilai R-square data training dan testing yang tidak berbeda jauh mengindikasikan bahwa tidak ada model yang underfit maupun overfit, serta model dapat menggeneralisasi dengan baik pada data baru. Selain itu, ditemukan bahwa model regresi Elastic Net memiliki kinerja yang tidak lebih baik dari regresi Ridge dan LASSO.

Oleh karena itu, model linier terbaik untuk dilakukan regresi pada studi kasus tersebut adalah Model Regresi LASSO. Tingginya nilai R-square dan rendahnya nilai RMSE membuktikan bahwa data

model yang dibangun memiliki tingkat akurasi yang tinggi dan cukup baik dalam menjelaskan variabel dari *dataset* tersebut.

REFERENSI

[1] H. Kim, "Seasonal Impacts of Particulate Matter Levels on Bike Sharing in Seoul, South Korea," *International Journal of Environmental Research and Public Health,* vol. 17, pp. 1-17, 2020.

[2] S. V. E., J. Park and Y. Cho, "Using data mining techniques for bike sharing demand prediction in metropolitan city," *Computer Communication,* vol. 153, pp. 353-366, 2020.

[3] S. Hidayatulloh, M. A. Mustajab and Y. Ramdhani, "PENGGUNAAN OTIMASI ATRIBUT DALAM PENINGKATAN AKURASI PREDIKSI DEEP LEARNING PADA BIKE SHARING DEMAND," *INFOTECH Journal,* vol. 9, pp. 54-61, 2023.

[4] A. Widarjono, Analisis Multivariat Terapan, Edisi Kedua, Yogyakarta: UPP STIM YKPN, 2015.

[5] G. W. Kusuma and I. Y. Wulansari, "ANALISIS KEMISKINAN DAN KERENTANAN KEMISKINAN DENGAN REGRESI RIDGE, LASSO, DAN ELASTIC-NET DI PROVINSI JAWA TENGAH TAHUN 2017," in *Seminar Nasional Official Statistics*, Jakarta, 2019.

[6] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Second Edition, Canada: O'Reilly Media, Inc., 2019.

[7] D. Cielen, A. D. B. Meysman and M. Ali, Introducing Data Science, New York: Manning Publications, 2016.

[8] T. Hendrawati, "Kajian Metode Imputasi dalam Menangani Missing Data," in *Seminar Nasional Matematika dan Pendidikan Matematika*, Surakarta, 2015.

[9] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis Fifth Edition, New Jersey: Wiley, 2012.

[10] I. G. A. M. Srinadi, "Pengaruh Outlier Terhadap Estimator Parameter Regresi dan Metode Regresi Robust," in *Konferensi Nasional Matematika*, Surabaya, 2014.

[11] I. Ghozali, Ekonometrika. Teori, Konsep dan Aplikasi dengan SPSS 17, Semarang: Badan Penerbit Universitas Diponegoro, 2009.

[12] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2nd Edition, Melbourne: OTexts, 2018.

[13] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono and C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar dengan Metode Regresi Linier," *Jurnal Nasional Teknologi dan Sistem Informasi,* pp. 8-17, 2021.

[14] N. D. Nachrowi and H. Usman, Pendekatan Populer dan Praktis Ekonometrika Untuk Analisis Ekonomi dan Keuangan, Jakarta: Lembaga Penerbit Fakultas Ekonomi Universitas Indonesia, 2006.