SIGMA
JURNAL TEKNOLOGI PELITA BANGSA
Vol.12 No.1 Maret 2020 p-ISSN: 2407-3903 e-ISSN: 2407-3903
Diterima 2 Februari 2020 | Direvisi 20 Februari 2020 | Dipublikasikan 10 Maret 2020

Penerapan Metode Regresi Linear untuk Prediksi Rata-rata Lama Sekolah di Kabupaten dan Kota di Pulau Jawa

Ahmad Nirwana[1], Ananto Tri Sasongko[2]
[1,2] Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]ahmadnirwana07@gmail.com, [2]ananto@pelitabangsa.ac.id

Abstract

Education is a process that involves imparting knowledge, skills, values and experience to individuals. A high level of education contributes to economic growth and social stability. This study focuses on factors that affect the average length of schooling in cities and districts in Java, which is a region with a huge education gap. The average length of schooling is an important indicator of the Human Development Index, which reflects the number of years of formal education the population receives. Statistics show that there are significant differences between regions, with some regions having lower average length of schooling. Factors such as education gaps, quality of education, limited resources, socioeconomic disparities, population mobility and low motivation may be the causes of this low attainment. Through a linear regression approach, this study aims to understand the relationship between certain variables and the average length of schooling in Java. Predictions using this method are expected to provide deep insight into the factors affecting education level. The data used in this study is data on the average length of schooling by city and district in Java Island from 2013 to 2023 obtained from the Central Bureau of Statistics in Indonesia. The results of the conclusions and analysis of the prediction of the average length of schooling by City and Regency in Java Island from 2013-2023 using the multiple linear regression method, it was concluded that the results of the multiple linear regression analysis obtained an RMSE value of 3.87836, an MAE value of 3.02546 and an R2 score value of 0.1 or 100%. The results of this research are expected to be the basis for the development of more appropriate policies and strategies in improving education in cities and districts in Java.

Keywords: *Education, Average length of schooling, Java Island, Multiple Linear Regression.*

Abstrak

Pendidikan adalah suatu proses yang melibatkan pemberian pengetahuan, keterampilan, nilai-nilai dan pengalaman kepada individu. Tingkat pendidikan yang tinggi berkontribusi terhadap pertumbuhan ekonomi dan stabilitas sosial. Penelitian ini berfokus pada faktor-faktor yang mempengaruhi rata-rata lama sekolah di kota dan kabupaten di Pulau Jawa, yang merupakan wilayah dengan kesenjangan pendidikan yang sangat besar. Rata-rata lama sekolah merupakan indikator penting Indeks Pembangunan Manusia, yang mencerminkan jumlah tahun pendidikan formal yang diterima penduduk. Statistik menunjukkan bahwa terdapat perbedaan yang signifikan antar daerah, dimana beberapa daerah mempunyai rata-rata lama sekolah yang lebih rendah. Faktor-faktor seperti kesenjangan pendidikan, kualitas pendidikan, keterbatasan sumber daya, kesenjangan sosial ekonomi, mobilitas penduduk dan rendahnya motivasi mungkin menjadi penyebab rendahnya pencapaian ini. Melalui pendekatan regresi linier, penelitian ini bertujuan untuk memahami hubungan antara variabel-variabel tertentu dengan rata-rata lama sekolah di Pulau Jawa. Prediksi menggunakan metode ini diharapkan dapat memberikan wawasan yang mendalam tentang faktor-faktor yang mempengaruhi tingkat pendidikan. Data yang digunakan pada penelitian ini yaitu data rata-rata lama sekolah menurut kota dan kabupaten di Pulau

Jawa dari tahun 2013 sampai tahun 2023 yang di peroleh dari Badan Pusat Statistik di Indonesia. Hasil kesimpulan dan analisis prediksi rata-rata lama sekolah menurut Kota dan Kabupaten di Pulau Jawa dari tahun 2013-2023 menggunakan metode regresi linear berganda, maka didapatkan kesimpulan bahwa hasil dari analisa regresi linear berganda di dapatkan nilai RMSE sebesar 3.87836, nilai MAE sebesar 3.02546 dan nilai R2 score didapatkan sebesar 0.1 atau 100%. Hasil penelitian ini diharapkan dapat menjadi dasar untuk pengembangan kebijakan dan strategi yang lebih tepat dalam meningkatkan pendidikan di kota-kota dan kabupaten-kabupaten di Pulau Jawa.

Kata kunci: Pendidikan, Rata-rata lama sekolah, Pulau Jawa, Regresi Linear Berganda.

1. Pendahuluan

Pendidikan adalah proses yang melibatkan pemberian pengetahuan, keterampilan, nilai, dan pengalaman kepada individu. Ini bukan hanya tentang apa yang dipelajari di sekolah, tetapi juga melibatkan pembelajaran sepanjang hayat di berbagai konteks, mulai dari rumah, lingkungan, hingga institusi formal. Perkembangan teknologi dan pendidikan telah mengubah cara kita belajar. Pendidikan memainkan peran kunci dalam perkembangan sosial dan ekonomi suatu masyarakat. Tingkat pendidikan yang tinggi sering kali terkait dengan pertumbuhan ekonomi yang lebih baik dan stabilitas sosial. Ini adalah kunci untuk memahami dunia, mengembangkan potensi, dan menciptakan perubahan positif. Manusia merupakan salah satu faktor pembentuk kekayaan bangsa yang nyata. Tujuan utama pembangunan adalah sebagai salah satu upaya yang dilakukan oleh pemerintah untuk mewujudkan suatu lingkungan yang mampu menguatkan setiap penduduknya dalam menikmati umur panjang, sehat, memiliki wawasan luas, dan menjalankan kehidupan yang *profitable* agar terbentuk kesejahteraan dan kemakmuran hidup yang efektif dan efisien.

Pulau Jawa merupakan salah satu wilayah terpadat di Indonesia, memiliki beragam karakteristik dan disparitas dalam sistem pendidikannya antara kabupaten dan kota. Oleh karena itu, penelitian mengenai faktor-faktor yang memengaruhi rata-rata lama sekolah di kabupaten dan kota di Pulau Jawa menjadi relevan untuk dipelajari. Rata-rata Lama Sekolah merupakan salah satu indikator penting dari komponen indeks pendidikan dalam pencapaian indeks pembangunan manusia yang optimal. Rata-rata lama sekolah berguna untuk menggambarkan jumlah tahun yang digunakan oleh penduduk dalam menjalani pendidikan formal sesuai dengan kesepakatan United Nations Development Program (UNDP) perhitungan rata-rata lama sekolah memiliki batas maksimum 15 tahun dan batas minimum 0 tahun. Berdasarkan data Badan Pusat Statistik, pada tahun 2013-2023 dari 119 kota dan kabupaten di pulau jawa, diketahui rata-rata lama sekolah tertinggi berada di Kota Tangerang Selatan mencapai 11,7 tahun. Kota Yogyakarta berada di peringkat kedua yang mencapai 11,55 tahun, sedangkan rata-rata lama sekolah terendah berada di Kabupaten Sampang yang berada di Pulau Madura Provinsi Jawa Timur yang hanya mencapai 4,28 tahun. Kota-kota dan Kabupaten-kabupaten yang masih minim rata-rata lama sekolah di Pulau Jawa membutuhkan perhatian khusus dan terampil terhadap setiap penduduknya agar mampu meningkatkan indeks pembangunan manusia secara merata.

Rendahnya pencapaian rata-rata lama sekolah disuatu wilayah dikarenakan beberapa faktor tertentu seperti kesenjangan pendidikan, kualitas pendidikan yang beragam, keterbatasan sumber daya, perbedaan sosial-ekonomi, mobilitas penduduk, dan rendahnya motivasi penduduk terhadap pendidikan. Seiring berkembangnya teknologi dan informasi yang mendunia dan berdasarkan pemaparan masalah diatas, sangat dibutuhkan penanganan secara objektif dan serius oleh pemerintah dalam meningkatkan rata-rata lama sekolah khususnya di Pulau Jawa sehingga pemerintah dapat meningkatkan kesejahteraan penduduk melalui pendidikan dalam pencapaikan indeks Pembangunan manusia yang optimal dan berkualitas.

Melakukan prediksi dengan pendekatan metode statistik dan heuristik menjadi salah satu poin utama untuk mengetahui meningkatnya rata-rata lama sekolah diperiode tertentu. Metode regresi linier merupakan salah satu alat statistik yang kuat untuk menganalisis hubungan antara variabel independen dan dependen [1]. Dalam konteks ini, penggunaan metode regresi linier untuk memprediksi rata-rata lama sekolah di berbagai wilayah di Pulau Jawa dapat memberikan pemahaman yang lebih mendalam tentang faktor-faktor apa saja yang berkontribusi terhadap tingkat pendidikan di sana.

Penelitian ini bertujuan untuk menjelaskan dan menganalisis hubungan antara variabel-variabel dengan rata-rata lama sekolah di kabupaten dan kota di Pulau Jawa. Berdasarkan pemaparan latar belakang

masalah, diharapkan penelitian dapat memprediksi rata-rata lama sekolah sebagai indikator dalam pencapaian indeks pembangunan manusia di Kota-kota dan Kabupaten-kabupaten khususnya di Pulau Jawa dan menjadi landasan yang kuat untuk pengembangan strategi dan kebijakan yang lebih tepat guna dalam meningkatkan tingkat pendidikan di Pulau Jawa. Dari hasil penelitian ini, diharapkan dapat menjadi masukan, saran, dan upaya khususnya kepada pemerintah daerah di Pulau Jawa dalam meningkatkan rata-rata lama sekolah penduduk di setiap Kota dan Kabupaten sesuai dengan pendekatan maksimal United National Development Program (UNDP) yaitu menjalani pendidikan maksimal 15 tahun sehingga indeks Pembangunan manusia berkembang secara optimal serta dapat meningkatkan kesejahteraan penduduk.

2. Metode Penelitian

Metode penelitian adalah cara untuk mengumpulkan data dan informasi melalui beberapan tahapan yaitu studi literatur dan mengumpulkan data yang akan digunakan untuk menyelesaikan sebuah kasus. Jenis penelitian yang diterapkan merupakan jenis penelitian kuantitatif yang menekankan pada data numerik. Variabel yang digunakan yaitu variabel independen atau variabel yang mempengaruhi dan variabel dependen atau variabel yang dipengaruhi [2]. Data yang digunakan menggunakan data *time series*. Data *time series* adalah kumpulan data yang didapatkan dari hasil perhitungan waktu ke waktu seperti jumlah penduduk di pulau jawa perbulan, jumlah angka kelahiran perbulan, dan pertumbuhan ekonomi suatu provinsi per tahun. Jenis data yang digunakan adalah data sekunder. Data sekunder merupakan data yang diperoleh dari sumber yang sudah tersedia sebelumnya. Metode yang digunakan pada penelitian ini yaitu menggunakan metode regresi linear berganda.

2.1. Data Penelitian

Data yang digunakan pada penelitian ini yaitu data rata-rata lama sekolah menurut Kota dan Kabupaten di Pulau Jawa selama 11 Tahun mulai Tahun 2013 sampai Tahun 2023 yang di peroleh dari Badan Pusat Statistik di Indonesia. Berikut ini merupakan contoh datasetnya.

Provinsi/Kabupaten/Kota	[Metode Baru] Rata-rata Lama Sekolah (Tahun)	
	2022	2023
ACEH	9,44	9,55
Simeulue	9,73	9,81
Aceh Singkil	8,69	8,7
Aceh Selatan	8,89	8,91
Aceh Tenggara	9,92	10,09
Aceh Timur	8,32	8,47
Aceh Tengah	9,87	9,89
Aceh Barat	9,87	9,98
Aceh Besar	10,35	10,36
Pidie	9,02	9,03
Bireuen	9,31	9,32
Aceh Utara	8,73	8,85
Aceh Barat Daya	8,68	8,77
Gayo Lues	8,41	8,42
Aceh Tamiang	9,04	9,24
Nagan Raya	8,95	8,96
Aceh Jaya	8,72	8,74
Bener Meriah	10,01	10,12

Gambar 1. Dataset

2.2. Regresi Linear Berganda

Regresi Linier Berganda adalah algoritma yang digunakan untuk mengukur hubungan antara korelasi dua variabel atau lebih yang digunakan untuk prediksi melalui garis lurus [3]. Variabel sendiri merupakan ukuran yang memiliki nilai yang berubah ubah. Model persamaan regresi linear berganda adalah sebagai berikut [4]:

$$Y = a + b1x1 + b2x2$$

Keterangan:

Y = Variabel terikat

a = Konstanta (*intercept*)

b = koefiensi regresi

X = Variabel bebas

Berikut ini rumus yang digunakan untuk menentukan nilai a dan b [5]:

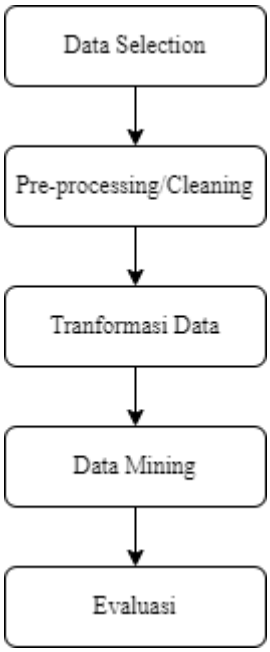
$$a = \frac{\sum y(\sum x^2) - \sum x.\sum xy}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - (\sum x)^2}$$

2.3. Data Mining

Pada tahapan penelitian dilakukan tahapan-tahapan dalam data mining yang disebut dengan *Knowledge Data Discovery* ataupun *pattern recognition* [6]. *Knowledge Data Discover* adalah proses pengumpulan informasi yang dapat mengekstrak informasi data yang berarti pada data [7]. Data *mining* (DM) menjadi inti dari proses KDD, yaitu dengan menggunakan algoritma tertentu untuk mengeksplorasi data, membangun model dan menemukan pola yang belum diketahui [8]. Model digunakan untuk memahami fenomena data, analisa

maupun prediksi [9]. Tahap dari proses *Knowledge Data Discovery* dapat dilihat berikut ini [10]:



Gambar 2. *Knowledge Data Discovery*

3. Hasil dan Pembahasan

3.1. Data *Selection*

Data yang digunakan pada penelitian ini yaitu data rata-rata lama sekolah menurut Kota dan Kabupaten di Pulau Jawa selama 11 Tahun mulai Tahun 2013 sampai Tahun 2023 yang di peroleh dari Badan Pusat Statistik di Indonesia. Kemudian data tersebut diseleksi dan akan digunakan untuk memprediksi prediksi rata-rata lama sekolah menurut Kota dan Kabupaten di Pulau Jawa.

```
data.show()
```

```
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|  Kota dan Kabupaten| 2013| 2014| 2015| 2016| 2017| 2018| 2019| 2020| 2021| 2022| 2023|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|         Kep. Seribu| 7.99| 8.03| 8.04| 8.24| 8.25| 8.46| 8.47| 8.68| 8.81| 9.02| 9.03|
|Kota Jakarta Selatan|10.95|10.97|11.23|11.42|11.47|11.57|11.62|11.63|11.64|11.66|11.75|
|  Kota Jakarta Timur|11.18|11.21|11.32|11.52| 11.6|11.64|11.65|11.66|11.67|11.73|11.88|
|  Kota Jakarta Pusat|10.85|10.87|10.88|11.01|11.02|11.24|11.25|11.38|11.39|11.53|11.54|
|  Kota Jakarta Barat|10.04|10.13|10.15|10.36|10.37|10.38| 10.4|10.63|10.78|11.13|11.23|
|  Kota Jakarta Utara| 9.85| 9.85|10.05|10.23| 10.6|10.69| 10.7| 10.8|10.81|10.82|10.84|
|               Bogor|  7.4|  7.7|  7.7|  7.8|  7.8|  7.8|  7.8|  7.8|  7.8|  7.8|  7.8|
|            Sukabumi| 6.32| 6.36| 6.51| 6.74| 6.79|  6.8| 7.02| 7.07|  7.1| 7.11| 7.33|
|             Cianjur|  6.5| 6.52| 6.54| 6.61| 6.92| 6.93| 6.97| 7.18| 7.19|  7.2| 7.22|
|             Bandung| 8.18| 8.34| 8.41|  8.5| 8.51| 8.58| 8.79| 8.96| 9.07| 9.08|  9.1|
|              Garut|  6.8| 6.83| 6.84| 6.88| 7.28|  7.5| 7.51| 7.52| 7.53| 7.83| 7.84|
|          Tasikmalaya| 6.69| 6.87| 6.88| 6.94| 7.12| 7.13| 7.17| 7.35| 7.48| 7.73| 7.96|
|             Ciamis|  7.2| 7.44| 7.45| 7.55| 7.59|  7.6| 7.69|  7.7|  7.9|  8.0| 8.09|
|            Kuningan| 6.98| 7.04|  7.2| 7.34| 7.35| 7.36| 7.38| 7.57|  7.8| 7.88| 7.89|
|             Cirebon| 6.08| 6.31| 6.32| 6.41| 6.61| 6.62| 6.71| 6.92|  7.1|  7.4| 7.64|
|           Majalengka| 6.72| 6.75|  6.8| 6.89|  6.9| 6.91| 7.09| 7.27| 7.31| 7.49| 7.52|
|            Sumedang| 7.51| 7.66| 7.66| 7.72| 7.98| 8.17| 8.27| 8.51| 8.52| 8.72| 8.73|
|            Indramayu| 5.29| 5.45| 5.46| 5.56| 5.97| 5.98| 5.99|  6.3| 6.52| 6.83| 6.94|
|              Subang| 6.29| 6.44| 6.45| 6.58| 6.83| 6.84| 6.85|  7.1| 7.11|  7.2| 7.45|
|           Purwakarta| 7.11| 7.17| 7.35| 7.42| 7.74| 7.75| 7.92| 8.09|  8.1| 8.11| 8.13|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

Gambar 3. Data *Selection*

3.2. *Pre-processing/Cleaning*

Tahap *preprocessing* merupakan tahap awal dari proses KDD. Pada tahapan ini data yang *missing value* harus dibersihkan. Hal ini dikarenakan data yang *missing value* merupakan syarat awal dalam dalam melakukan data mining. Suatu data dikatakan *missing value* jika terdapat atribut dalam dataset yang tidak berisi nilai atau kosong.

```
from pyspark.sql.functions import col

# Looping untuk menghitung jumlah nilai null dalam setiap kolom
null_counts = [data.where(col(c).isNull()).count() for c in data.columns]

# Menggabungkan nama kolom dengan jumlah null-nya
nulls_in_columns = zip(data.columns, null_counts)

# Menampilkan jumlah null dalam setiap kolom
for column, null_count in nulls_in_columns:
    print(f"Kolom '{column}' memiliki {null_count} nilai null")

```

```
Kolom 'Kota dan Kabupaten' memiliki 0 nilai null
Kolom '2013' memiliki 0 nilai null
Kolom '2014' memiliki 0 nilai null
Kolom '2015' memiliki 0 nilai null
Kolom '2016' memiliki 0 nilai null
Kolom '2017' memiliki 0 nilai null
Kolom '2018' memiliki 0 nilai null
Kolom '2019' memiliki 0 nilai null
Kolom '2020' memiliki 0 nilai null
Kolom '2021' memiliki 0 nilai null
Kolom '2022' memiliki 0 nilai null
Kolom '2023' memiliki 0 nilai null
```

Gambar 4. Data *Missing Value*

3.3. *Transformation*

Pada tahap ini tidak diperlukan transformasi karena *value* pada atribut yang digunakan semuanya telah berupa data *numeric*.

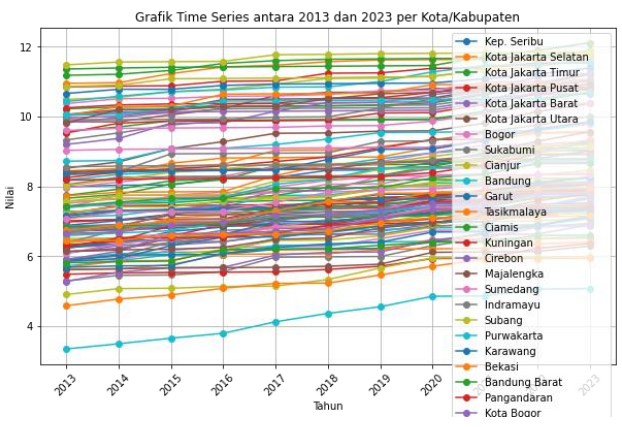
```
data.printSchema()
```

```
root
 |-- Kota dan Kabupaten: string (nullable = true)
 |-- 2013: double (nullable = true)
 |-- 2014: double (nullable = true)
 |-- 2015: double (nullable = true)
 |-- 2016: double (nullable = true)
 |-- 2017: double (nullable = true)
 |-- 2018: double (nullable = true)
 |-- 2019: double (nullable = true)
 |-- 2020: double (nullable = true)
 |-- 2021: double (nullable = true)
 |-- 2022: double (nullable = true)
 |-- 2023: double (nullable = true)
```

Gambar 5. Data *Transformation*

3.4. Data *Mining*

Berikut ini adalah data *mining* menampilkan data per tahun dari kota dan kabupaten di pulau jawa.



Gambar 6. Visualisasi Data

3.4.1. Menghitung rata-rata

Pada tahap ini diperlukan rata-rata setiap baris dari kolom-kolom setiap tahun untuk menentukan variabel Y dengan menambahkan kolom rata-rata dari tahun 2013-2023. Berikut ini adalah tampilan hasil perhitungan.

```
# Menghitung rata-rata per baris dari kolom-kolom tahunan dan menambahkan kolom avg
data = data.withColumn("Rata-rata 2013-2023", sum(col(c) for c in data.columns[1:]) / len(data.columns[1:]))

# Menampilkan hasil
data.show()
```

```
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
|  Kota dan Kabupaten| 2013| 2014| 2015| 2016| 2017| 2018| 2019| 2020| 2021| 2022| 2023|Rata-rata 2013-2023|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
|         Kep. Seribu| 7.99| 8.03| 8.04| 8.24| 8.25| 8.46| 8.47| 8.68| 8.81| 9.02| 9.03|  8.456363636363635|
|Kota Jakarta Selatan|10.95|10.97|11.23|11.42|11.47|11.57|11.62|11.63|11.64|11.66|11.75| 11.446363636363637|
|  Kota Jakarta Timur|11.18|11.21|11.32|11.52| 11.6|11.64|11.65|11.66|11.67|11.73|11.88|  11.55090909090909|
|  Kota Jakarta Pusat|10.85|10.87|10.88|11.01|11.02|11.24|11.25|11.38|11.39|11.53|11.54| 11.178181818181816|
|  Kota Jakarta Barat|10.04|10.13|10.15|10.36|10.37|10.38| 10.4|10.63|10.78|11.13|11.23| 10.509090909090908|
|  Kota Jakarta Utara| 9.85| 9.85|10.05|10.23| 10.6|10.69| 10.7| 10.8|10.81|10.82|10.84| 10.476363636363637|
|              Bogor|  7.4|  7.74|  7.75|  7.83|  7.84|  7.88|  8.29|  8.3|  8.31|  8.34|  8.37|  8.004545454545456|
|            Sukabumi| 6.32| 6.36| 6.51| 6.74| 6.79|  6.8| 7.02| 7.07|  7.1| 7.11| 7.33|  6.831818181818181|
|             Cianjur|  6.5| 6.52| 6.54| 6.61| 6.92| 6.93| 6.97| 7.18| 7.19|  7.2| 7.22|  6.889090909090908|
|             Bandung| 8.18| 8.34| 8.41|  8.5| 8.51| 8.58| 8.79| 8.96| 9.07| 9.08|  9.1|  8.683636363636364|
|              Garut|  6.8| 6.83| 6.84| 6.88| 7.28|  7.5| 7.51| 7.52| 7.53| 7.83| 7.84|  7.305454545454546|
|        Tasikmalaya| 6.69| 6.87| 6.88| 6.94| 7.12| 7.13| 7.17| 7.35| 7.48| 7.73| 7.96|  7.210909090909092|
|              Ciamis|  7.2| 7.44| 7.45| 7.55| 7.59|  7.6| 7.69|  7.7|  7.9|  8.0| 8.09|  7.655454545454546|
|            Kuningan| 6.98| 7.04|  7.2| 7.34| 7.35| 7.36| 7.38| 7.57|  7.8| 7.88| 7.89|  7.435454545454545|
|              Cirebon| 6.08| 6.31| 6.32| 6.41| 6.61| 6.62| 6.71| 6.92|  7.1|  7.4| 7.64|  6.738181818181818|
|           Majalengka| 6.72| 6.75|  6.8| 6.89|  6.9| 6.91| 7.09| 7.27| 7.31| 7.49| 7.52|  7.059090909090909|
|            Sumedang| 7.51| 7.66| 7.66| 7.72| 7.98| 8.17| 8.27| 8.51| 8.52| 8.72| 8.73|  8.131818181818183|
|            Indramayu| 5.29| 5.45| 5.46| 5.56| 5.97| 5.98| 5.99|  6.3| 6.52| 6.83| 6.94| 6.0263636363636355|
|              Subang| 6.29| 6.44| 6.45| 6.58| 6.83| 6.84| 6.85|  7.1| 7.11|  7.2| 7.45|  6.830909090909091|
|           Purwakarta| 7.11| 7.17| 7.35| 7.42| 7.74| 7.75| 7.92| 8.09|  8.1| 8.11| 8.13|  7.717272727272728|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
only showing top 20 rows
```

Gambar 7. Menambahkan Data Rata-rata

3.4.2. Pemilihan fitur yang di inginkan

Pada tahapan ini dilakukan pemilihan fitur (*features*) yang akan digunakan dalam proses pembuatan model regresi linear. Fitur-fitur tersebut melibatkan nama kota atau kabupaten, serta tahun-tahun dari 2013 hingga 2023, dan kolom Rata-rata 2013-2023 yang merupakan target atau variabel yang akan diprediksi.

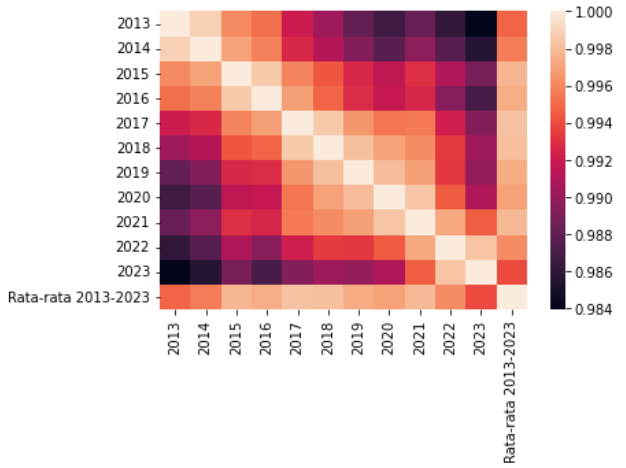
```
#PEMILIHAN FITUR YANG DIINGINKAN

selected_data = data.select('Kota dan Kabupaten','2013', '2014','2015','2016','2017',
                            '2018','2019','2020','2021','2022','2023','Rata-rata 2013-2023')
selected_data.show()
```

```
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
|  Kota dan Kabupaten| 2013| 2014| 2015| 2016| 2017| 2018| 2019| 2020| 2021| 2022| 2023|Rata-rata 2013-2023|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
|         Kep. Seribu| 7.99| 8.03| 8.04| 8.24| 8.25| 8.46| 8.47| 8.68| 8.81| 9.02| 9.03|  8.456363636363635|
|Kota Jakarta Selatan|10.95|10.97|11.23|11.42|11.47|11.57|11.62|11.63|11.64|11.66|11.75| 11.446363636363637|
|  Kota Jakarta Timur|11.18|11.21|11.32|11.52| 11.6|11.64|11.65|11.66|11.67|11.73|11.88|  11.55090909090909|
|  Kota Jakarta Pusat|10.85|10.87|10.88|11.01|11.02|11.24|11.25|11.38|11.39|11.53|11.54| 11.178181818181816|
|  Kota Jakarta Barat|10.04|10.13|10.15|10.36|10.37|10.38| 10.4|10.63|10.78|11.13|11.23| 10.509090909090908|
|  Kota Jakarta Utara| 9.85| 9.85|10.05|10.23| 10.6|10.69| 10.7| 10.8|10.81|10.82|10.84| 10.476363636363637|
|              Bogor|  7.4|  7.74|  7.75|  7.83|  7.84|  7.88|  8.29|  8.3|  8.31|  8.34|  8.37|  8.004545454545456|
|            Sukabumi| 6.32| 6.36| 6.51| 6.74| 6.79|  6.8| 7.02| 7.07|  7.1| 7.11| 7.33|  6.831818181818181|
|             Cianjur|  6.5| 6.52| 6.54| 6.61| 6.92| 6.93| 6.97| 7.18| 7.19|  7.2| 7.22|  6.889090909090908|
|             Bandung| 8.18| 8.34| 8.41|  8.5| 8.51| 8.58| 8.79| 8.96| 9.07| 9.08|  9.1|  8.683636363636364|
|              Garut|  6.8| 6.83| 6.84| 6.88| 7.28|  7.5| 7.51| 7.52| 7.53| 7.83| 7.84|  7.305454545454546|
|        Tasikmalaya| 6.69| 6.87| 6.88| 6.94| 7.12| 7.13| 7.17| 7.35| 7.48| 7.73| 7.96|  7.210909090909092|
|              Ciamis|  7.2| 7.44| 7.45| 7.55| 7.59|  7.6| 7.69|  7.7|  7.9|  8.0| 8.09|  7.655454545454546|
|            Kuningan| 6.98| 7.04|  7.2| 7.34| 7.35| 7.36| 7.38| 7.57|  7.8| 7.88| 7.89|  7.435454545454545|
|              Cirebon| 6.08| 6.31| 6.32| 6.41| 6.61| 6.62| 6.71| 6.92|  7.1|  7.4| 7.64|  6.738181818181818|
|           Majalengka| 6.72| 6.75|  6.8| 6.89|  6.9| 6.91| 7.09| 7.27| 7.31| 7.49| 7.52|  7.059090909090909|
|            Sumedang| 7.51| 7.66| 7.66| 7.72| 7.98| 8.17| 8.27| 8.51| 8.52| 8.72| 8.73|  8.131818181818183|
|            Indramayu| 5.29| 5.45| 5.46| 5.56| 5.97| 5.98| 5.99|  6.3| 6.52| 6.83| 6.94| 6.0263636363636355|
|              Subang| 6.29| 6.44| 6.45| 6.58| 6.83| 6.84| 6.85|  7.1| 7.11|  7.2| 7.45|  6.830909090909091|
|           Purwakarta| 7.11| 7.17| 7.35| 7.42| 7.74| 7.75| 7.92| 8.09|  8.1| 8.11| 8.13|  7.717272727272728|
+--------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-------------------+
only showing top 20 rows
```

Gambar 8. Pemilihan Fitur

3.4.3. Korelasi Antar Data

Analisa korelasi digunakan untuk melihat keterkaitan tentang derajat ikatan variabel sehingga bisa mengenali ikatan variabel yang terdapat.



Gambar 9. Korelasi Antar Data

3.4.4. Proses Data *Test* dan Data *Training*

Pada tahapan ini dilakukan proses input data *test* dan data *training* yang dibagi menjadi data test 20% dan data training 80%.

```
#PEMISAHAN DATA TRAINING DAN DATA TESTING
# Memisahkan data menjadi data latih dan data uji dengan perbandingan 80:20
train_data, test_data = selected_data.randomSplit([0.8, 0.2], seed=42)
train_data.show()
```

Gambar 10. Pembagian data *test* dan data *training*

3.4.5. Membuat Model Regresi Linear

Membuat model regresi linear merupakan suatu proses melatih suatu model statistik yang dapat digunakan untuk memahami dan memodelkan hubungan linier antara variabel dependen dan variabel independen.

```
# Pembuatan Model Regresi Linear
assembler = VectorAssembler(inputCols=['2013', '2014','2015','2016','2017','2018','2019',
                                       '2020','2021','2022','2023','Rata-rata 2013-2023'],
                            outputCol='rata-rata lama sekolah')
train_data = assembler.transform(train_data)
test_data = assembler.transform(test_data)
```

```
# Inisialisasi model regresi linear
lr = LinearRegression(featuresCol="rata-rata lama sekolah", labelCol="Rata-rata 2013-2023")
```

```
# Melatih model menggunakan data latih
model = lr.fit(train_data)
```

```
# Melakukan prediksi menggunakan data uji
predictions = model.transform(test_data)
```

Gambar 11. Membuat Model Regresi Linear

3.4.6. Menampilkan Hasil Prediksi

Pada tahapan ini bertujuan untuk menampilkan hasil prediksi dari model regresi linear pada kolom nama kota atau kabupaten, tahun-tahun dari 2013 hingga 2023, nilai rata-rata lama sekolah dari 2013 hingga 2023, dan kolom *prediction* yang berisi nilai prediksi yang dihasilkan oleh model. Berikut ini adalah tampilan hasil proses.

```
# Tampilkan hasil prediksi
predictions.select('Kota dan Kabupaten', '2013', '2014', '2015', '2016', '2017',
                   '2018', '2019', '2020', '2021', '2022', '2023','Rata-rata 2013-2023','prediction').show()
```

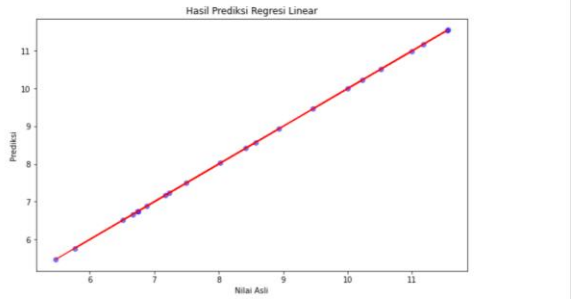
```
+-------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----------------+-----------------+
|Kota dan Kabupaten|2013|2014|2015|2016|2017|2018|2019|2020|2021|2022|2023|Rata-rata 2013-2023|       prediction|
+-------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----------------+-----------------+
|           Bangkalan|  4.9| 5.07| 5.08| 5.13| 5.14| 5.33| 5.66| 5.95| 5.96| 5.97| 5.99|5.4709090909090909|5.4709090908851...|
|         Banyuwangi| 6.84| 6.87| 6.88| 6.93| 7.11| 7.12| 7.13| 7.16| 7.42| 7.66| 7.76|7.1709090909090912|7.1709090909097295|
|             Bekasi| 8.34| 8.38| 8.66| 8.81| 8.82| 8.84| 8.84| 9.12|  9.3| 9.53| 9.57|  8.928181818181818|8.9281818181818645|
|           Bondowoso| 5.48| 5.52| 5.53| 5.54| 5.55| 5.62| 5.71| 5.93| 5.94| 6.22| 6.36|5.7636363636363636|5.7636363636362044|
|             Cirebon| 6.08| 6.31| 6.32| 6.41| 6.61| 6.62| 6.71| 6.92|  7.1|  7.4| 7.64|6.7381818181818181|6.7381818181819984|
|            Grobogan| 6.25| 6.32| 6.33| 6.62| 6.66| 6.67| 6.86| 6.91| 7.11| 7.26| 7.28|6.7518181818181813|6.7518181818183311|
|          Karanganyar| 8.38| 8.47| 8.48| 8.49|  8.5| 8.51| 8.52| 8.56| 8.57| 8.79| 9.02|  8.571818181818181|8.5718181818181448|
|             Klaten| 7.74| 7.92| 8.16| 8.22| 8.23| 8.24| 8.31| 8.58| 8.81| 9.09| 9.27|  8.415454545454546|8.4154545454545439|
|         Kota Depok|10.43|10.58|10.71|10.76|10.84|10.85| 11.0|11.28|11.46|11.47|11.58|10.996363636363636|10.996363636363597|
|  Kota Jakarta Barat|10.04|10.13|10.15|10.36|10.37|10.38| 10.4|10.63|10.78|11.13|11.23| 10.50909090909091|10.5090909090903...|
|  Kota Jakarta Pusat|10.85|10.87|10.88|11.01|11.02|11.24|11.25|11.38|11.39|11.53|11.54|11.178181818181816|11.178181818181592|
|  Kota Jakarta Timur|11.18|11.21|11.32|11.52|  11.6|11.64|11.65|11.66|11.67|11.73|11.88|11.558090909090909|11.558090909092297|
|         Kota Kediri| 9.57|  9.7| 9.88| 9.89|  9.9| 9.91| 9.92| 9.93|10.15|10.45|10.69| 9.999090909090908|9.9990909090909672|
|       Kota Mojokerto| 9.91| 9.91| 9.92| 9.93| 9.98| 9.99|10.24|10.25|10.47| 10.8|11.05|10.222727272727273|10.222727272727216|
|       Kota Sukabumi| 8.52|  8.7| 9.08| 9.28| 9.52| 9.53| 9.58| 9.59| 9.81|10.14|10.37| 9.465454545454547|9.4654545454546666|
|     Kota Yogyakarta|11.36|11.39|11.41|11.42|11.43|11.44|11.45|11.46|11.72|11.89|12.11|11.552727272727275|11.552727272727785|
|           Magelang| 6.88| 7.02| 7.19|  7.4| 7.41| 7.57| 7.77| 7.78| 7.79| 7.81| 7.82| 7.494545454545457|7.4945454545453231|
|             Serang| 6.65| 6.69|  6.9| 6.98| 7.17| 7.18| 7.33|  7.5| 7.51| 7.78| 7.79| 7.225454545454546|7.2254545454537515|
|             Tegal| 5.85| 5.93|  6.3| 6.54| 6.55|  6.7| 6.86| 6.98| 6.99| 7.25| 7.34| 6.662727272727272|6.6627272727269431|
|          Tulungagung| 7.44| 7.45| 7.72| 7.73| 7.82| 8.06| 8.07| 8.33| 8.34| 8.65| 8.66| 8.024545454545454|8.0245454545450271|
+-------------------+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----------------+-----------------+
only showing top 20 rows
```

Gambar 12. Hasil Prediksi

3.4.7. Visualisasi Hasil Prediksi

Pada tahapan ini menggunakan Matplotlib untuk membuat visualisasi hasil prediksi. Scatter plot digunakan dengan sumbu x (horizontal) adalah nilai aktual (Rata-rata 2013-2023) dan sumbu y (vertikal) adalah nilai prediksi (prediction). Selain itu, garis merah (Ideal) ditambahkan pada plot untuk menunjukkan kondisi ideal di mana nilai aktual sama dengan nilai prediksi. Label sumbu x dan y, serta judul plot, juga ditambahkan untuk memberikan konteks visual pada hasil prediksi. Dengan tahapan ini dapat dilihat seberapa baik model regresi linear memprediksi nilai target dengan membandingkan sebaran nilai aktual dan nilai prediksi. Berikut ini tampilan visualisasi.

```
# Mengonversi data prediksi menjadi Pandas DataFrame
pred_df = predictions.select('Rata-rata 2013-2023', "prediction").toPandas()

# Visualisasi hasil prediksi
plt.figure(figsize=(10, 6))
plt.scatter(pred_df['Rata-rata 2013-2023'], pred_df['prediction'], color='blue', alpha=0.5)
plt.plot(pred_df['Rata-rata 2013-2023'], pred_df['Rata-rata 2013-2023'],
         label="Ideal", color="red")  # Garis ideal (nilai aktual = prediksi)
plt.xlabel('Nilai Asli')
plt.ylabel('Prediksi')
plt.title('Hasil Prediksi Regresi Linear')
plt.show()
```



Gambar 13. Visualisasi Hasil Prediksi

3.4.8. Evaluasi Model Regresi

Pada tahapan ini dilakukan evaluasi model regresi linear dengan cara yang dilakukan yaitu RMSE, MAE, dan R^2. RMSE memberikan informasi tentang seberapa besar deviasi prediksi dari nilai yang sebenarnya. MAE memberikan ukuran kesalahan mutlak rata-rata antara prediksi dan nilai sebenarnya. R-squared memberikan gambaran seberapa baik model kita cocok dengan variasi dalam data. Semakin tinggi nilainya, semakin baik modelnya. Berikut ini adalah tampilan hasil dari prosesnya.

```
from pyspark.ml.evaluation import RegressionEvaluator

# Evaluasi model
evaluator_rmse = RegressionEvaluator(labelCol='Rata-rata 2013-2023', predictionCol='prediction', metricName='rmse')
evaluator_mae = RegressionEvaluator(labelCol='Rata-rata 2013-2023', predictionCol='prediction', metricName='mae')
evaluator_r2 = RegressionEvaluator(labelCol='Rata-rata 2013-2023', predictionCol='prediction', metricName='r2')

# Menghitung nilai untuk model
rmse = evaluator_rmse.evaluate(predictions)
mae = evaluator_mae.evaluate(predictions)
r2 = evaluator_r2.evaluate(predictions)

# Tampilkan hasil evaluasi
print(f'Root Mean Squared Error (RMSE): {rmse}')
print(f'Mean Absolute Error (MAE): {mae}')
print(f'R-squared (R2): {r2}')
```

```
Root Mean Squared Error (RMSE): 3.8783582457018895e-13
Mean Absolute Error (MAE): 3.0254586714694265e-13
R-squared (R2): 1.0
```

Gambar 14. Evaluasi Model Regresi Linear

4. Kesimpulan

Berdasarkan hasil dan analisis diatas terkait prediksi rata-rata lama sekolah menurut Kota dan Kabupaten di Pulau Jawa dari tahun 2013-2023 menggunakan metode regresi linear berganda, maka didapatkan kesimpulan bahwa hasil dari analisa regresi linear berganda di dapatkan nilai RMSE sebesar 3.87836, nilai MAE sebesar 3.02546 dan nilai R^2 score didapatkan sebesar 0.1 atau 100%.

Ucapan Terima Kasih

Sebutkan nama pemberi dana dan pemberi fasilitas yang membantu.

Referensi

[1] P. R. Linear, U. Prediksi, H. Beras, D. Indonesia, V. Arinal, and M. Azhari, "Penerapan Regresi Linear Untuk Prediksi Harga Beras Di Indonesia," *Jurnal Sains dan Teknologi*, vol. 5, no. 1, p. |pp, 2023, doi: 10.55338/saintek.v5i1.1417.

[2] W. Andriani, Gunawan, and A. E. Prayoga, "PREDIKSI NILAI EMAS MENGGUNAKAN ALGORITMA REGRESI LINEAR," *Jurnal Ilmiah Informatika Komputer*, vol. 28, no. 1, pp. 27–35, 2023, doi: 10.35760/ik.2023.v28i1.8096.

[3] H. Husdi and H. Dalai, "Penerapan Metode Regresi Linear Untuk Prediksi Jumlah Bahan Baku Produksi Selai Bilfagi," *Jurnal Informatika*, vol. 10, no. 2, pp. 129–135, Oct. 2023, doi: 10.31294/inf.v10i2.14129.

[4] M. Arif and M. Faisal, "Penerapan Model Regresi Linear Untuk Estimasi Mobil Bekas Menggunakan Bahasa Python," *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 11, no. 2, pp. 182–191, Nov. 2023, doi: 10.37905/euler.v11i2.20698.

[5] D. Wulandari, "Pemodelan dan Prediksi Produksi Padi Menggunakan Regresi Linear." [Online]. Available:

https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera

[6] A. Supriyadi Sunge and A. Turmudi Zy, "ANALISIS PREDIKSI PENJUALAN DENGAN METODE REGRESI LINEAR DI PT. EAGLE INDUSTRY INDONESIA," 2023.

[7] A. Junia Karlina, M. Irsyad, F. Insani, and E. Pandu Cynthia, "KLIK: Kajian Ilmiah Informatika dan Komputer Estimasi Hasil Panen Ayam Pedaging Menggunakan Algoritma Regresi Linear Berganda," *Media Online)*, vol. 3, no. 6, pp. 966–976, 2023, doi: 10.30865/klik.v3i6.920.

[8] F. Ekawati and H. Adi Chandra, "PREDIKSI PERTUMBUHAN PENDUDUK KAL-SEL MENGGUNAKAN METODE REGRESI LINEAR BERGANDA," *Technologia*, vol. 14, no. 4, 2023, doi: 10.31602/tji.v14i3.12574.

[9] M. Fadhilah and I. Ali, "Prediksi Jumlah Produksi Sablon Tahun Menggunakan Algoritma Regresi Linear di Nolbas SVNR," *INTERNAL (Information System Journal*, vol. 6, no. 1, pp. 22–32, doi: 10.32627.

[10] A. Maulana and I. Ali, "PREDIKSI HASIL PRODUKSI PANEN BAWANG MERAH MENGGUNAKAN METODE REGRESI LINIER SEDERHANA," 2023.