

**Laporan Studi Kasus Ridge & Lasso Regression  
Berdasarkan Indeks Pembangunan Manusia (IPM) Provinsi  
Maluku Menggunakan Data Tahun 2020 - 2023**



**Disusun Oleh:**

<b>Aurelia Hapsari Dyah Rinjani</b>	<b>3323600035</b>
<b>Evinda Eka Ayudia Lestari</b>	<b>3323600039</b>
<b>Robi'Atul Adawiyah</b>	<b>3323600041</b>
<b>Wahyu Ikbal Maulana</b>	<b>3323600056</b>

**Mata Kuliah Praktikum Pemodelan Statistik Terapan  
Program Studi D4 Sains Data Terapan  
Departemen Teknik Informatika dan Komputer  
Politeknik Elektronika Negeri Surabaya  
2024**

## LEMBAR PENGESAHAN

Laporan resmi ini untuk memenuhi tugas mata kuliah praktikum pemodelan statistik terapan yang disusun oleh.

1. Nama : Aurelia Hapsari Dyah Rinjani  
NRP : 3323600035
2. Nama : Evinda Eka Ayudia Lestari  
NRP : 3323600039
3. Nama : Robi' Atul Adawiyah  
NRP : 3323600041
4. Nama : Wahyu Iqbal Maulana  
NRP : 3323600056

Surabaya, Maret 2024  
Mengetahui,  
Dosen Pengampu

**Ronny Susetyoko, S.Si., M.Si**  
NIP. 197112111995011001

## **ABSTRAK**

Indonesia terdiri dari 38 provinsi yang tersebar mulai dari Sabang hingga Merauke, salah satunya adalah provinsi Maluku. Di Provinsi Maluku, seperti halnya di banyak daerah di Indonesia, peningkatan IPM dianggap sebagai bagian penting dari upaya pembangunan yang berkelanjutan. Pemanfaatan teknik analisis data seperti Ridge dan Lasso Regression dapat memberikan pemahaman yang berharga terhadap faktor-faktor yang mempengaruhi IPM di Provinsi Maluku. Data yang digunakan dalam analisis ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Maluku menggunakan data tahun 2020-2023. Analisis dilakukan dengan menerapkan tiga metode yaitu, metode regresi linear, Ridge dan Lasso. Model Ridge dan Lasso keduanya menunjukkan nilai yang bagus dalam menangani overfitting, dimana keduanya mampu menangani overfitting dengan baik. Model Lasso dengan koefisien determinasi data train dan data test sebesar 89% dan 70%, dan untuk model Ridge menangani sedikit lebih baik dengan koefisien determinasi data train dan data test sebesar 83% dan 78%.

Kata kunci: Linear regresi, Ridge, Lasso

## DAFTAR ISI

<b>LEMBAR PENGESAHAN.....</b>	<b>2</b>
<b>ABSTRAK.....</b>	<b>3</b>
<b>DAFTAR ISI.....</b>	<b>4</b>
<b>DAFTAR TABEL DAN GAMBAR.....</b>	<b>5</b>
<b>BAB 1 PENDAHULUAN.....</b>	<b>6</b>
1.1. Latar Belakang.....	6
1.2. Tujuan.....	6
1.3. Rumusan Masalah.....	7
<b>BAB 2 TINJAUAN PUSTAKA.....</b>	<b>8</b>
2.1. Indeks Pembangunan Manusia.....	8
2.2. Umur Harapan Hidup.....	8
2.3. Sarana Kesehatan.....	8
2.4. Jumlah Siswa (SMA).....	8
2.5. Harapan Lama Sekolah.....	9
2.6. Penduduk Miskin.....	9
2.7. Produk Domestik Regional Bruto (PDRB).....	9
2.8. Pengeluaran perkapita.....	9
<b>BAB 3 METODE PENELITIAN.....</b>	<b>11</b>
3.1. Dataset.....	11
3.2. Metodologi.....	11
3.3. Algoritma.....	11
3.3.1. Regresi Linear.....	11
3.3.2. Regresi Ridge.....	12
3.3.3. Regresi Lasso.....	13
<b>BAB 4 HASIL PENELITIAN DAN LUARAN YANG DICAPAI.....</b>	<b>15</b>
4.1. Membaca dataset.....	15
4.2. Nilai statistik deskriptif.....	15
4.3. Visualisasi.....	16
4.4. Linear Regresi.....	20
4.5. Regresi Lasso.....	21
4.6. Regresi Ridge.....	22
<b>BAB 5 KESIMPULAN DAN SARAN.....</b>	<b>25</b>
5.1. Kesimpulan.....	25
5.2. Saran.....	25
<b>DAFTAR PUSTAKA.....</b>	<b>27</b>

## DAFTAR TABEL DAN GAMBAR

Tabel 1. Dataset Provinsi Maluku.....	15
Tabel 2. Nilai Statistik Deskriptif.....	15
Gambar 1. Visualisasi Heatmap.....	16
Gambar 2. Visualisasi Scatter plot.....	17
Gambar 3. Visualisasi Boxplot.....	18
Gambar 4. Visualisasi Parallel Coordinate.....	18
Gambar 5. Visualisasi Correlation Heatmap.....	19

# **BAB 1**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Indonesia terdiri dari 38 provinsi yang tersebar mulai dari Sabang hingga Merauke, dengan masing-masing provinsi memiliki ibukota di wilayahnya sendiri. Salah satunya adalah provinsi Maluku, yang memiliki Kota Ambon sebagai ibu kotanya. Provinsi Maluku terbagi menjadi 9 kabupaten dan 2 kotamadya. Di setiap provinsi perlu adanya pembangunan berkelanjutan dalam upaya peningkatan mutu di provinsi tersebut seperti peningkatan mutu ekonomi, hidup manusia, dan lain-lain.

Peningkatan mutu hidup manusia telah menjadi perhatian utama dalam proses pembangunan suatu provinsi. Salah satu metode yang digunakan untuk mengukur mutu hidup adalah Indeks Pembangunan Manusia (IPM), yang meliputi aspek-aspek seperti kesehatan, pendidikan, dan standar hidup. IPM mencerminkan kemajuan suatu wilayah dalam memenuhi kebutuhan dasar penduduknya. Di Provinsi Maluku, seperti halnya di banyak daerah di Indonesia, peningkatan IPM dianggap sebagai bagian penting dari upaya pembangunan yang berkelanjutan.

Pemanfaatan teknik analisis data seperti Ridge dan Lasso Regression dapat memberikan pemahaman yang berharga terhadap faktor-faktor yang mempengaruhi IPM di Provinsi Maluku. Pendekatan ini memungkinkan identifikasi dan evaluasi dampak relatif dari berbagai variabel yang berkontribusi pada peningkatan IPM. Dengan demikian, penggunaan analisis data tersebut dapat mendukung perancangan kebijakan yang lebih efektif untuk meningkatkan mutu hidup penduduk Maluku.

### **1.2. Tujuan**

1. Mengetahui faktor-faktor yang berpengaruh signifikan terhadap Indeks Pembangunan Manusia (IPM) di Provinsi Maluku dari tahun 2020 hingga 2023.
2. Mengetahui cara kerja model regresi linear dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023.
3. Mengetahui cara kerja model regresi Ridge dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023.
4. Mengetahui cara kerja model regresi Lasso dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023.

### **1.3. Rumusan Masalah**

1. Apa saja faktor-faktor yang berpengaruh signifikan terhadap Indeks Pembangunan Manusia (IPM) di Provinsi Maluku dari tahun 2020 hingga 2023?
2. Bagaimana cara kerja model regresi linear dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023?
3. Bagaimana cara kerja model regresi Ridge dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023?
4. Bagaimana cara kerja model regresi Lasso dalam memprediksi IPM Provinsi Maluku berdasarkan data tahun 2020 hingga 2023?

## **BAB 2**

### **TINJAUAN PUSTAKA**

#### **2.1. Indeks Pembangunan Manusia**

Indeks Pembangunan Manusia (IPM) adalah sebuah alat ukur yang digunakan untuk mengevaluasi kualitas pembangunan manusia dari berbagai aspek, baik yang bersifat fisik maupun non-fisik. Dalam konteks fisik, hal ini mencakup kesehatan dan kesejahteraan, yang dapat tercermin dalam angka harapan hidup dan daya beli masyarakat. Sementara itu, aspek non-fisik melibatkan kualitas pendidikan masyarakat. IPM memberikan gambaran yang luas mengenai kinerja pembangunan suatu wilayah dengan memperhatikan aspek-aspek penting seperti harapan hidup, pendidikan, dan standar hidup layak.

#### **2.2. Umur Harapan Hidup**

Umur Harapan Hidup adalah rata-rata kesempatan atau waktu hidup yang tersisa. Usia harapan hidup bisa diartikan pula dengan banyaknya tahun yang ditempuh penduduk yang masih hidup sampai umur tertentu. Besar kecilnya umur harapan hidup suatu generasi sangat dipengaruhi oleh banyaknya penduduk yang mampu melewati umur tertentu.

Peningkatan umur harapan hidup bisa menjadi tolok ukur keberhasilan program kesehatan dan program pembangunan sosial ekonomi. Umur harapan hidup yang rendah di suatu daerah bisa menjadi perhatian pemerintah untuk lebih banyak program pembangunan, kesehatan, dan program sosial lainnya seperti kesehatan lingkungan, kecukupan gizi dan kalori, termasuk program pemberantasan kemiskinan.

#### **2.3. Sarana Kesehatan**

Sarana kesehatan atau bisa disebut dengan fasilitas pelayanan kesehatan adalah alat dan/atau tempat yang digunakan untuk menyelenggarakan upaya pelayanan kesehatan, baik promotif, preventif, kuratif, maupun rehabilitatif yang dilakukan oleh Pemerintah, Pemerintah Daerah, dan/atau masyarakat. (UU No. 38 Tahun 2014).

#### **2.4. Jumlah Siswa (SMA)**

Jumlah Siswa dalam dataset ini mengacu pada jumlah siswa SMA di Provinsi Maluku. Jumlah siswa mungkin bisa memberikan pengaruh terhadap Indeks Pembangunan Manusia pada suatu daerah dikarenakan pendidikan adalah salah satu dari



dimensi penentu IPM. Ketika jumlah siswa meningkat maka diharapkan IPM pada daerah tersebut juga meningkat.

## **2.5. Harapan Lama Sekolah**

Menurut Badan Pusat Statistik (2021), angka harapan lama sekolah (HLS) mencerminkan durasi sekolah yang diharapkan akan dialami oleh anak pada usia tertentu di masa mendatang. Keberadaan HLS sangat penting dalam menilai kemajuan pembangunan sistem pendidikan pada berbagai tingkat, yang menggambarkan estimasi lama pendidikan yang diharapkan dapat dicapai oleh setiap anak. Angka HLS dihitung untuk populasi yang berusia 7 tahun ke atas.

## **2.6. Penduduk Miskin**

Penduduk Miskin menurut BPS adalah penduduk yang memiliki rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan. Untuk mengukur kemiskinan, Biro Pusat Statistik (BPS) menggunakan konsep kemampuan memenuhi kebutuhan dasar (*basicneedsapproach*). Dengan pendekatan ini, kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran.

Menurut Piven dan Clowed dan Swanson dalam Suharto (2009:15) kemiskinan menggambarkan adanya kelangkaan materi atau barang-barang yang diperlukan dalam kehidupan sehari-hari seperti makanan, pakaian dan perumahan.

## **2.7. Produk Domestik Regional Bruto (PDRB)**

PDRB merupakan jumlah nilai tambah yang dihasilkan oleh seluruh unit usaha dalam suatu daerah tertentu, atau merupakan jumlah nilai barang dan jasa akhir (neto) yang dihasilkan oleh seluruh unit ekonomi. PDRB atas dasar harga berlaku menggambarkan nilai tambah barang dan jasa yang dihitung menggunakan harga yang berlaku pada setiap tahun, sedang PDRB atas dasar harga konstan menunjukkan nilai tambah barang dan jasa tersebut yang dihitung menggunakan harga yang berlaku pada satu tahun tertentu sebagai dasar. PDRB atas dasar harga berlaku dapat digunakan untuk melihat pergeseran dan struktur ekonomi, sedangkan harga konstan digunakan untuk mengetahui pertumbuhan ekonomi dari tahun ke tahun. (BPS, 2023)

## **2.8. Pengeluaran perkapita**

Pengeluaran perkapita digunakan untuk mengukur standar hidup manusia. Hal ini juga dipengaruhi oleh pengetahuan serta peluang yang ada untuk

merealisasikan pengetahuan dalam berbagai kegiatan produktif sehingga menghasilkan output baik berupa barang maupun jasa sebagai pendapatan. Kemudian pendapatan yang ada menciptakan pengeluaran atau konsumsi. Pengeluaran perkapita memberikan gambaran tingkat daya beli PPP (*Purchasing Power Parity*) masyarakat, dan sebagai salah satu komponen yang digunakan dalam melihat status pembangunan manusia di suatu wilayah.

## **BAB 3**

### **METODE PENELITIAN**

#### **3.1. Dataset**

Data yang digunakan dalam analisis ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Maluku menggunakan data tahun 2020-2023. Dataset ini mencakup berbagai variabel berdasarkan Indeks Pembangunan Manusia (IPM) Provinsi Maluku dengan variabel yang digunakan dalam analisis ini adalah sebagai berikut:

x1 = Umur Harapan Hidup saat lahir

x2 = Jumlah Sarana Kesehatan

x3 = Jumlah Siswa (SMA)

x4 = Harapan lama sekolah

x5 = Persentase Penduduk Miskin

x6 = PDRB atas harga yang berlaku

x7 = Pengeluaran Perkapita

#### **3.2. Metodologi**

Analisis ini menggunakan pendekatan analisis data retrospektif untuk mengeksplorasi faktor-faktor yang mempengaruhi Indeks Pembangunan Manusia (IPM) di Provinsi Maluku dari tahun 2020 hingga 2023. Analisis dilakukan dengan menerapkan tiga metode yaitu, metode regresi linear, Ridge dan Lasso pada dataset yang mencakup berbagai variabel independen yang relevan dengan IPM, seperti jumlah sarana kesehatan, angka partisipasi sekolah, pengeluaran perkapita, PDRB dan lain-lain.

#### **3.3. Algoritma**

##### **3.3.1. Regresi Linear**

Regresi Linear adalah suatu model regresi linear yang menggunakan metode untuk mendapatkan taksiran modelnya. Metode OLS adalah suatu metode penaksiran parameter model regresi yang meminimumkan jumlah kuadrat *error*. Model regresi OLS dengan satu variabel prediktor dapat ditulis seperti pada persamaan berikut (Kutner, dkk. (2004), Montgomery dkk (2008)):

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, t = 1, 2, \dots, T \quad (1)$$

dengan  $X_t$  adalah variabel prediktor pada pengamatan ke- $t$ ,  $Y_t$  adalah variabel respon pada pengamatan ke- $t$ ,  $\beta_0, \beta_1$  adalah parameter model dan  $\varepsilon_t$  adalah variabel error pada pengamatan ke- $t$ .

Penaksir OLS untuk parameter  $\beta_0$  dan  $\beta_1$  adalah seperti pada persamaan berikut (Sembiring, (2003), Draper dan Smith (1992)).

$$\hat{\beta}_0 = \frac{\sum_{t=1}^T Y_t}{T} - \hat{\beta}_1 \frac{\sum_{t=1}^T X_t}{T} = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (2)$$

dengan

$$\hat{\beta}_1 = \frac{\sum_{t=1}^T (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^T (X_t - \bar{X})^2} \quad (3)$$

Berdasarkan Persamaan (2) dan Persamaan (3) diperoleh taksiran model regresi OLS seperti pada persamaan berikut (Sembiring, 2003).

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t = \bar{Y} + \hat{\beta}_1 (X_t - \bar{X}) \quad (4)$$

### 3.3.2. Regresi Ridge

Regresi Ridge adalah suatu teknik yang dikembangkan untuk menstabilkan nilai koefisien regresi karena adanya multikolinieritas. Metode Regresi Ridge pertama kali diperkenalkan oleh A. E. Hoerl pada tahun 1962. Regresi Ridge merupakan modifikasi dari metode OLS yang menghasilkan penduga bias dari koefisien regresi (Kutner, et al, 2005). Regresi Ridge mengurangi dampak multikolinieritas dengan menentukan penduga yang bias tetapi mempunyai varians yang lebih kecil dari varians penduga OLS. Menurut Ohlyver (2011), meskipun metode ini menghasilkan penduga koefisien regresi bias, penduga ini bisa mendekati nilai parameter yang sebenarnya. Hal ini dapat diketahui dari perbandingan *mean square error* (MSE) antara penduga Ridge dengan penduga kuadrat terkecil (*least square*), dimana MSE penduga Ridge lebih kecil daripada MSE penduga OLS.

Regresi Ridge serupa dengan OLS yang meminimumkan *sum of squares error* (SSE) atau jumlah kuadrat sisaan pada pendugaan koefisien regresi. Namun, dalam metode ini menambahkan penalti penyusutan dalam meminimumkan SSE. Sehingga persamaan yang terbentuk adalah sebagai berikut:

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Dimana  $\lambda \geq 0$  adalah parameter penyusutan dan  $\sum_{j=1}^p \beta_j^2 = 1$  adalah penalti penyusutan. Jika  $\lambda = 0$ , maka penalti penyusutan tidak memberikan pengaruh apapun sehingga regresi ridge akan menghasilkan penduga koefisien regresi yang sama dengan OLS. Namun, jika  $\lambda \rightarrow \infty$  akan berdampak pada penalti penyusutan yang semakin besar dan koefisien pendugaan yang semakin mendekati nol. Pada OLS hanya dihasilkan satu set dugaan koefisien, sedangkan pada regresi ridge dihasilkan set dugaan koefisien berbeda-beda untuk setiap nilai  $\lambda$ . Parameter penyusutan ( $\lambda$ ) yang optimal ditentukan dengan metode validasi silang.

### 3.3.3. Regresi Lasso

LASSO (*Least Absolute Shrinkage and Selection Operator*) adalah salah satu metode penyusutan seperti regresi ridge yang dapat mengatasi permasalahan multikolinearitas yang diperkenalkan oleh Tibshirani tahun 1996. Dalam regresi ridge, penduga koefisien regresi disusutkan ke arah nol seiring dengan peningkatan nilai  $\lambda$ . Satu hal yang tidak dapat dilakukan oleh regresi ridge adalah melakukan seleksi peubah secara otomatis.

Menurut Tibshirani (1996), dua teknik standar untuk meningkatkan pendugaan OLS adalah pemilihan subset (*subset selection*) dan regresi ridge, tapi keduanya memiliki kelemahan. Pemilihan subset menyediakan model yang dapat diinterpretasikan tetapi bisa sangat bervariasi karena merupakan proses diskrit yang dipertahankan atau dikeluarkan dari model. Perubahan kecil dalam data dapat menghasilkan model terpilih yang sangat berbeda dan ini dapat mengurangi akurasi prediksi. Regresi Ridge adalah proses kontinu yang mengecilkan koefisien dan karenanya lebih stabil. Namun, tidak menetapkan koefisien apapun menjadi 0 dan karenanya tidak memberikan model yang mudah diinterpretasikan. Sedangkan LASSO dapat menyusutkan beberapa koefisien dan menetapkan yang lain ke 0. Oleh karena itu LASSO dapat mempertahankan fitur-fitur yang baik dari pemilihan subset dan regresi ridge.

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

LASSO dan Regresi Ridge memiliki persamaan yang hampir sama, hanya penaltinya saja yang berbeda. Penalti penyusutan pada regresi ridge  $\sum_{j=1}^p \beta_j^2$

diganti dengan  $\sum_{j=1}^p |\beta_j|$  sebagai penalti penyusutan LASSO. Oleh karena itu, LASSO dapat menyusutkan koefisien menjadi nol. Hal tersebut menyebabkan LASSO dapat menghasilkan model dengan peubah penjelas yang lebih sedikit dan parsimoni. Menurut Prabowo, Wilandari, dan Rusgiyono (2015), penalti penyusutan pada metode LASSO menyebabkan nilai penduga koefisien parameter menyusut sehingga variabel prediktor yang penting atau berpengaruh terhadap model tetap dimasukkan ke dalam model, sedangkan variabel prediktor yang kurang penting akan disusutkan sampai nol dan terseleksi dari model sehingga model menjadi lebih efisien.

## BAB 4

### HASIL PENELITIAN DAN LUARAN YANG DICAPAI

#### 4.1. Membaca dataset

```
df = pd.read_excel("/content/drive/MyDrive/Ruppu (1)/Data_Maluku.xlsx")
df.head()
```

	Kota	UHH	Sarana Kesehatan	Jumlah siswa	lama sekolah	Miskin	PDRB	Pengeluaran	IPM
0	Maluku Tenggara Barat	65.1975	16.75	5386.00	12.3100	25.6825	3031.7950	6456.75	63.3600
1	Maluku Tenggara	66.5950	22.75	3632.75	12.8550	22.0700	3646.8200	7773.50	66.7450
2	Maluku Tengah	67.6075	38.50	17463.75	14.2450	18.7425	9643.9675	10411.75	71.7075
3	Buru	67.4400	13.00	4609.25	13.1300	16.5675	2673.2375	10489.00	69.6950
4	Kepulauan Aru	64.8475	31.75	3115.75	12.3250	25.1275	3913.5700	7768.75	64.1900
5	Seram Bagian Barat	64.1675	23.50	9222.75	13.5475	23.7550	3372.7100	8906.00	66.4200
6	Seram Bagian Timur	61.8550	24.00	4321.00	12.8350	22.0250	3309.8200	9636.00	64.8225
7	Maluku Barat Daya	64.3025	22.00	3716.50	12.4600	28.7800	2198.0025	7010.25	62.8175
8	Buru Selatan	67.3575	15.00	2715.75	12.7650	15.4175	1576.3325	7747.50	65.2725
9	Ambon	71.2525	32.50	13534.50	16.0550	4.8650	16496.6125	14340.25	81.4400
10	Tual	67.1550	16.00	3379.50	13.9875	21.7525	2806.5075	7525.00	68.4700

Tabel 1. Dataset Provinsi Maluku

#### 4.2. Nilai statistik deskriptif

```
df.describe()
```

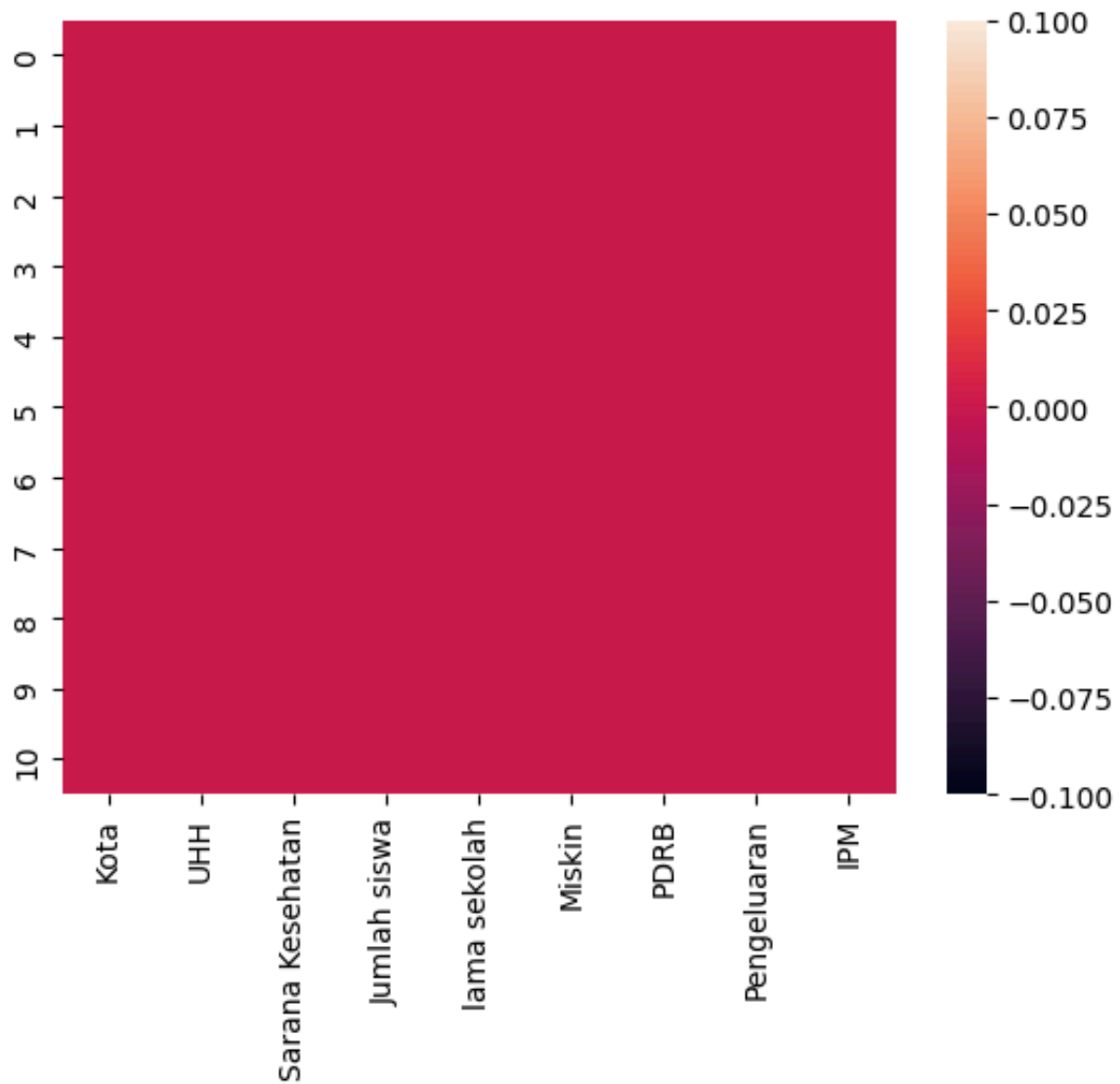
	UHH	Sarana Kesehatan	Jumlah siswa	lama sekolah	Miskin	PDRB	Pengeluaran	IPM
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	66.161591	23.250000	6463.409091	13.319545	20.435000	4788.125000	8914.977273	67.721818
std	2.467446	8.128653	4879.865037	1.112585	6.50059	4420.776996	2241.743279	5.302198
min	61.855000	13.000000	2715.750000	12.310000	4.865000	1576.332500	6456.750000	62.817500
25%	64.575000	16.375000	3506.125000	12.612500	17.655000	2739.872500	7636.250000	64.506250
50%	66.595000	22.750000	4321.000000	12.855000	22.025000	3309.820000	7773.500000	66.420000
75%	67.398750	27.875000	7304.375000	13.767500	24.44125	3780.195000	10023.875000	69.082500
max	71.252500	38.500000	17463.750000	16.055000	28.780000	16496.612500	14340.250000	81.440000

Tabel 2. Nilai statistik deskriptif

Berdasarkan hasil output di atas diperoleh nilai ringkasan statistik deskriptif dari setiap variabel yang terdiri dari count, nilai rata-rata (mean), standart deviasi (std), nilai minimum, nilai maksimum, kuartil 1, kuartil 2, dan kuartil 3.

### 4.3. Visualisasi

```
sns.heatmap(df.isnull())
```

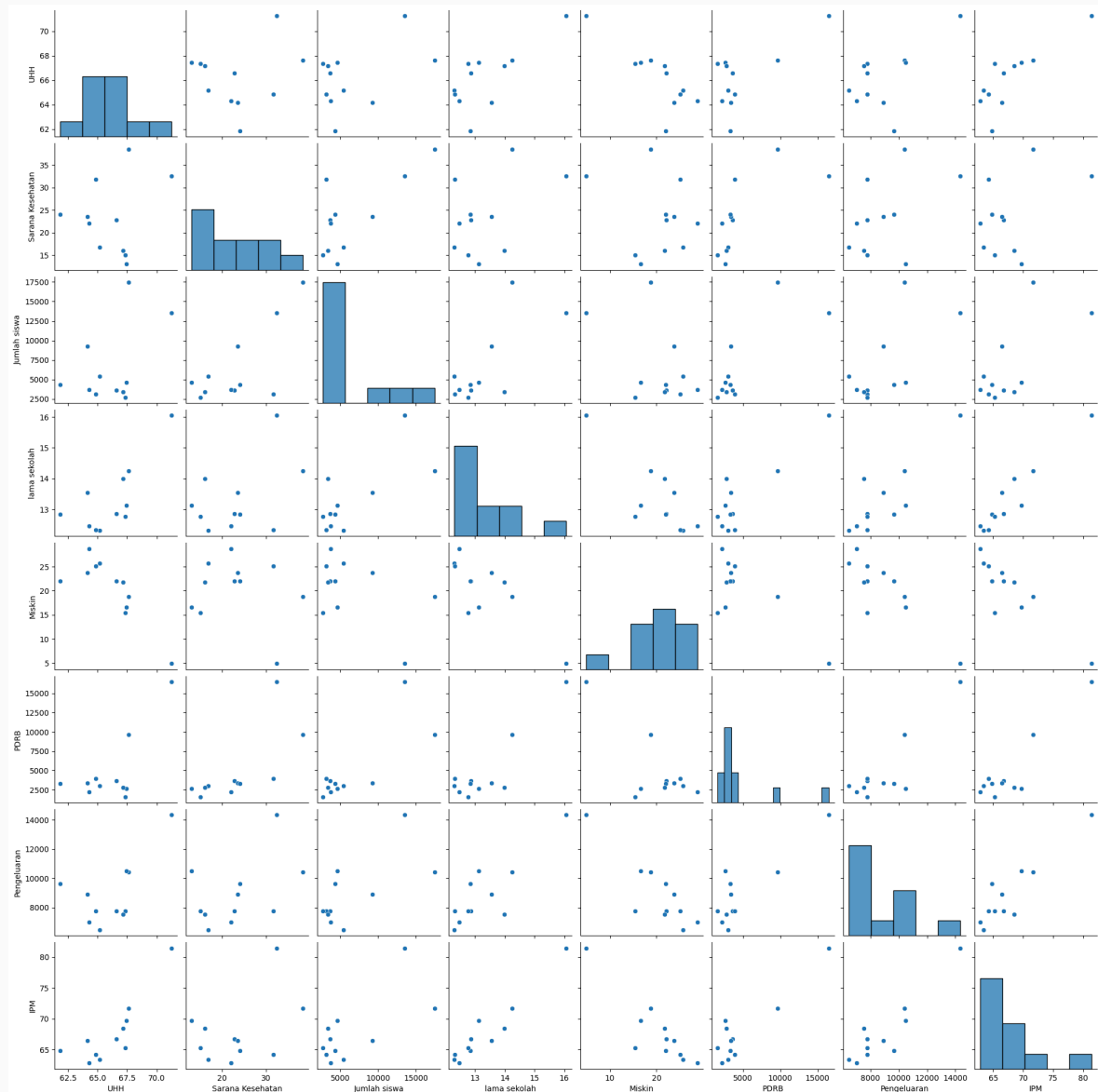


Gambar 1. Visualisasi Heatmap



# Memvisualisasikan hubungan antar variabel dengan scatter plot untuk melihat pola dan mengidentifikasi hubungan antar variabel dalam garis besar

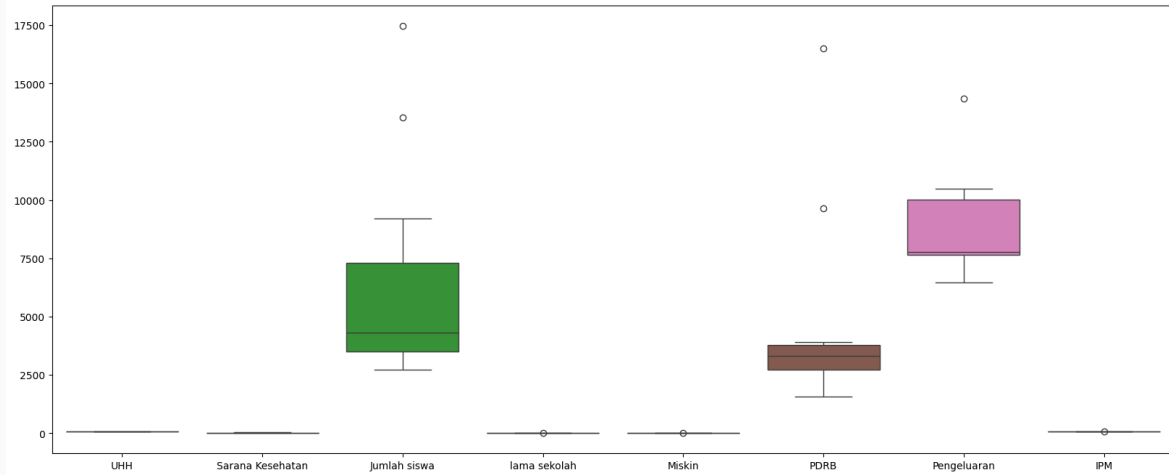
```
import matplotlib.pyplot as plt
sns.pairplot(df)
plt.show()
```



Gambar 2. Visualisasi Scatter plot

```
# Looking for outliers using box plot
```

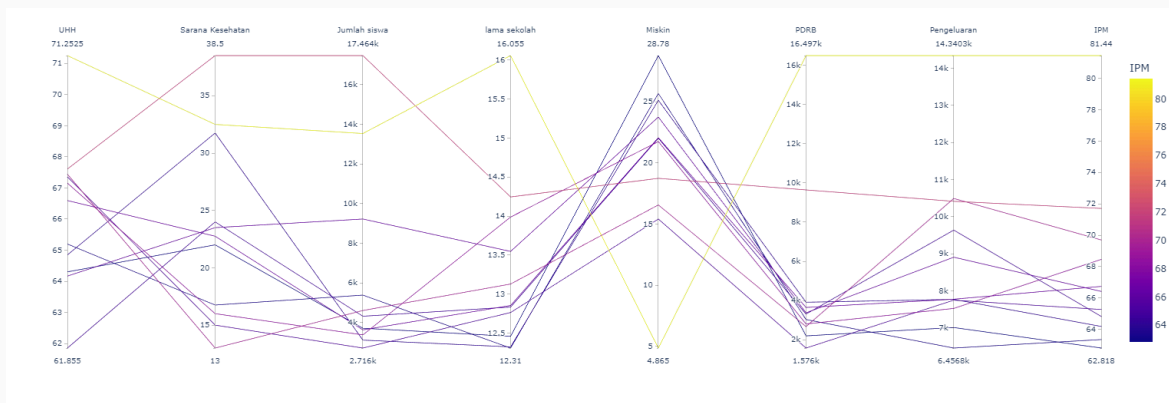
```
plt.figure(figsize = (20, 8))
sns.boxplot(data = df, width = 0.8)
plt.show()
```



Gambar 3. Visualisasi boxplot

Hasil analisis boxplot menunjukkan adanya sedikit outlier di feature 'lama sekolah', 'miskin', dan 'IPM'. Selain itu, pada jumlah siswa dan PDRB juga memiliki outlier yang cukup jauh dari batas rata-rata. Meskipun begitu disini outlier tidak diterapkan perubahan, dikarenakan outlier disini menunjukkan ketimpangan antar kabupaten/kota, sehingga ketimpangan tersebut bisa diatasi dengan lebih menyamaratakan infrastruktur dan pembangunan.

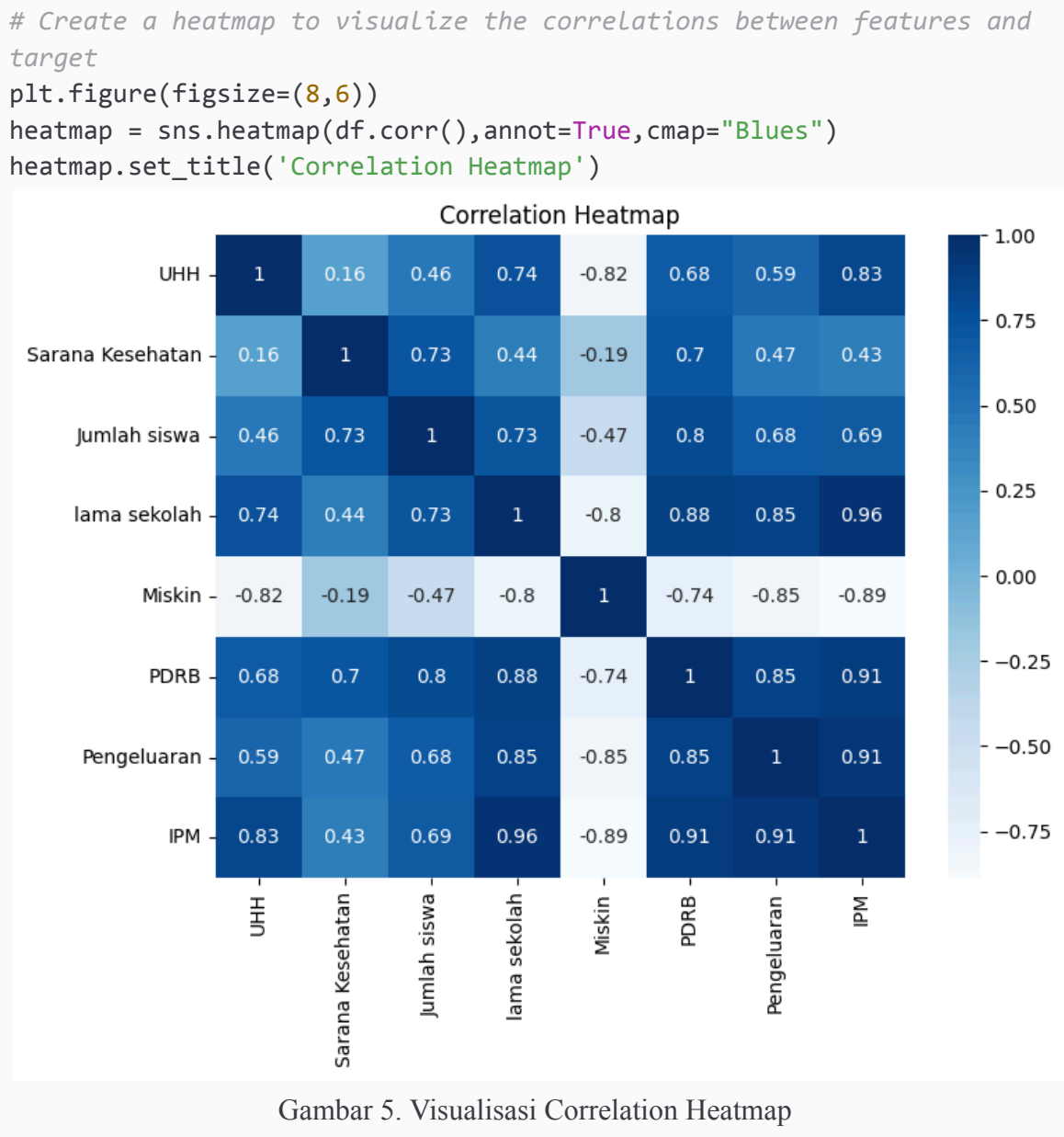
```
fig = px.parallel_coordinates(df, color=df['IPM'])
fig.show()
```



Gambar 4. Visualisasi Parallel Coordinate

Hasil plot menampilkan hubungan antar variabel. Ini memungkinkan pengguna untuk melihat bagaimana setiap variabel berinteraksi satu sama lain dalam satu plot. Untuk warna

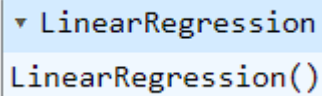
yang cerah menggambarkan bagaimana tiap variabel mempengaruhi IPM di atas rata-rata, sedangkan warna yang gelap untuk tiap variabel dengan IPM di bawah rata-rata.



Plot di atas menggambarkan kecenderungan korelasi tiap variabel. Diketahui bahwa keseluruhan variabel saling berkorelasi tinggi kecuali variabel 'miskin' yang memiliki korelasi negatif. Korelasi negatif antara variabel 'miskin' dan variabel lain menunjukkan bahwa kedua variabel tersebut bergerak dalam arah yang berlawanan. Dalam konteks ini, bisa berarti bahwa ketika satu variabel meningkat, variabel lainnya cenderung menurun, dan sebaliknya.

#### 4.4. Linear Regresi

```
lr = LinearRegression()
```



```
# Tanpa mempertimbangkan data train dan data test
```

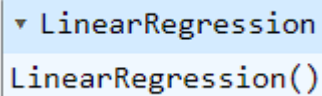
```
lr.fit(X_scaled, y)  
y_pred = lr.predict(X_scaled)
```

```
r2 = r2_score(y, y_pred)  
print(r2)
```

```
0.9956101552857205
```

```
lr = LinearRegression()
```

```
lr.fit(X_train, y_train)
```



```
train_score_lr = lr.score(X_train, y_train)  
test_score_lr = lr.score(X_test, y_test)
```

```
print("The train score for lr model is {}".format(train_score_lr))  
print("The test score for lr model is {}".format(test_score_lr))
```

```
lr_yprid = lr.predict(X_test)  
rmse = mean_squared_error(y_test, lr_yprid)
```

```
print('Slope:', lr.coef_)  
print('Intercept:', lr.intercept_)  
print('Root mean squared error: ', rmse)
```

```
The train score for lr model is 1.0
```

```
The test score for lr model is 0.3414731765307035
```

```
Slope: [2.05042156 0.36539121 0.41607394 0.00601105 0.11883322
```

```
3.11267016
```

```
2.68311028]
```

```
Intercept: 68.89096887427462
```

```
Root mean squared error: 26.58051200560009
```

## 4.5. Regresi Lasso

```
lasso_pred = lasso.predict(X_test)

r2 = r2_score(lasso_pred, y_test)
r2

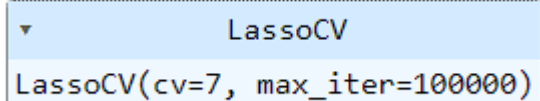
num_predictors = X_test.shape[1]
num_observations = len(y_test)
adjusted_r2 = 1 - (1 - r2) * ((num_observations - 1) / (num_observations
- num_predictors - 1))

print(adjusted_r2)
```

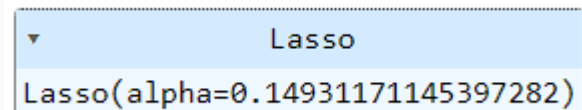
1.4993221103961942

```
from sklearn.linear_model import Lasso, LassoCV

lasso_cv = LassoCV(alphas = None, cv = 7, max_iter = 100000)
lasso_cv.fit(X_train, y_train)
```



```
lasso = Lasso(alpha = lasso_cv.alpha_)
lasso.fit(X_train, y_train)
```



Disini kami melakukan hyper parameter menggunakan LassoCV untuk menemukan parameter yang optimal yang kemudian diterapkan ke dalam algoritma lasso regression. Kemudian, objek Lasso CV dibuat dengan parameter `alfas=None` (untuk penaksiran otomatis), `cv=7` (jumlah lipatan validasi silang, berdasarkan jumlah variabel maksimum dari si feature), dan `max_iter=100000` (jumlah iterasi maksimum untuk mendapatkan hasil yang lebih stabil).

```
Lasso(alpha=0.14931171145397282)
lasso.score(X_train, y_train)
```

0.8949410581801919

Didapat sebuah parameter dari hasil Lasso Cross validation, kemudian diterapkan ke dalam untuk model ridge.

```

lasso_pred = lasso.predict(X_test)
r2 = r2_score(lasso_pred, y_test)
r2

num_predictors = X_test.shape[1]
num_observations = len(y_test)
adjusted_r2 = 1 - (1 - r2) * ((num_observations - 1) / (num_observations
- num_predictors - 1))

print(adjusted_r2)
1.4993221103961942

```

```

r2_train_lasso = r2_score(y_train, lasso.predict(X_train))
r2_test_lasso = r2_score(y_test, lasso.predict(X_test))

print("The train R-squared for lasso model is
{}".format(r2_train_lasso))
print("The test R-squared for lasso model is {}".format(r2_test_lasso))

lasso_y_pred = lasso.predict(X_test)
rmse = mean_squared_error(y_test, lasso_y_pred)

print('Root mean squared error: ', rmse)

print('Slope:', lasso.coef_)
print('Intercept:', lasso.intercept_)

The train R-squared for lasso model is 0.8949410581801919
The test R-squared for lasso model is 0.7052205133056311
Root mean squared error: 11.898360713395641
Slope: [ 1.23128394  0.          0.          0.08027036 -0.          0.
        2.34122519]
Intercept: 66.77745268142239

```

## 4.6. Regresi Ridge

```

alphas = np.random.uniform(0, 10, 50)
ridge_cv = RidgeCV(alphas = alphas, cv = 7)
ridge_cv.fit(X_train, y_train)

```

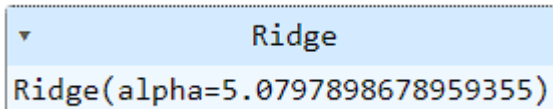
Disini kami melakukan hyper parameter menggunakan Ridge CV untuk menemukan parameter yang optimal yang kemudian diterapkan ke dalam algoritma ridge regression. Ridge Regression dengan Cross-Validation (Ridge CV) dapat membantu meningkatkan

kinerja model regresi dengan mengatasi masalah multikolinearitas dan memberikan estimasi yang lebih baik terhadap parameter model. Dasi hasil output di atas menunjukkan hasil dari array alpha yang optimal untuk setiap data yang akan kita terapkan model ridge.

```
# best alpha parameter
alpha = ridge_cv.alpha_
alpha

5.0797898678959355

ridge = Ridge(alpha = ridge_cv.alpha_)
ridge.fit(X_train, y_train)
```



```
▼ Ridge
Ridge(alpha=5.0797898678959355)
```

Didapat sebuah parameter dari hasil Ridge Cross validation yang kemudian diterapkan ke dalam untuk model ridge regression

```
ridge_pred = ridge.predict(X_test)

r2 = r2_score(ridge_pred, y_test)
print(r2)

num_predictors = X_test.shape[1]
num_observations = len(y_test)
adjusted_r2 = 1 - (1 - r2) * ((num_observations - 1) / (num_observations
- num_predictors - 1))

print(adjusted_r2)

-2.1640311793258418
3.373023384494381
```

```
r2_train_ridge = r2_score(y_train, ridge.predict(X_train))
r2_test_ridge = r2_score(y_test, ridge.predict(X_test))

print("The train R-squared for ridge model is {}".format(r2_train_ridge))
print("The test R-squared for ridge model is {}".format(r2_test_ridge))

ridge_y_pred = ridge.predict(X_test)
rmse = mean_squared_error(y_test, ridge_y_pred)

print('Root mean squared error: ', rmse)
```

```
print('Slope:', ridge.coef_)  
print('Intercept:', ridge.intercept_)
```

The train R-squared for ridge model is 0.8391371512853627

The test R-squared for ridge model is 0.7825218554674407

Root mean squared error: 8.77820040989241

Slope: [ 0.81596284 -0.07299519 0.31316735 0.65359682 -0.57222795  
0.30481011

1.22383076]

Intercept: 66.9178794532619

Untuk mempertimbangkan model lasso, dengan mempertimbangkan nilai koefisien determinasi dari data train dan data test, didapat bahwa model menggambarkan cukup baik.



## **BAB 5**

### **KESIMPULAN DAN SARAN**

#### **5.1. Kesimpulan**

Berdasarkan hasil analisis yang telah dilakukan menggunakan metode regresi linear, ridge dan lasso dapat disimpulkan bahwa ketiga metode analisis regresi dilakukan untuk memprediksi variabel target. Model linear regresi menampilkan akurasi yang tinggi, baik OLS maupun tanpa OLS, menghasilkan koefisien determinasi sebesar 0,95 yang modelnya adalah sebagai berikut:

$$Y = 67.721818 + 1.697755 X_1 + -0.186662 X_2 + -0.294841 X_3 + 1.820446 X_4 + 0.697222 X_5 + 0.750340 X_6 + 2.307310 X_7$$

Berdasarkan OLS, indikator-indikator yang mempengaruhi IPM di wilayah Maluku dengan menggunakan model regresi terbaik, dengan model regresi kuantil 0,95 adalah indikator sarana kesehatan (X2), lama sekolah (X4), dan pengeluaran rumah tangga (X7). Nilai koefisien determinasi adalah 0,95 telah menunjukkan bahwa indikator IPM mampu menjelaskan variabel-variabel X1, X2, X3, X4, X5, X6, X7, sedangkan 5% dijelaskan oleh variabel lainnya. Meskipun begitu nilai koefisien determinasi yang tinggi menunjukkan adanya overfitting karena setelah dilakukan prediksi terhadap data baru nilai koefisien determinasi menjadi menurun sebesar 70%, sehingga dilakukan regularisasi L1 dan L2 menggunakan Ridge dan Lasso untuk mengatasi overfitting.

Kedua model tersebut menunjukkan nilai yang bagus dalam menangani overfitting, dimana keduanya mampu menangani overfitting dengan baik. Model Lasso dengan koefisien determinasi data train dan data test sebesar 89% dan 70%, dan untuk model Ridge menangani sedikit lebih baik dengan koefisien determinasi data train dan data test sebesar 83% dan 78%.

#### **5.2. Saran**

Rekomendasi yang dapat diberikan berdasarkan analisis yang telah dilakukan adalah:

1. Pendidikan (Jumlah siswa) dan Ekonomi (PDRB tiap kabupaten) terdapat ketimpangan di beberapa kabupaten Maluku, sehingga Pemerintah diharapkan mampu mengambil langkah-langkah untuk meningkatkan akses pendidikan dan mengembangkan potensi ekonomi di wilayah-wilayah yang terpinggirkan.
2. Empat variabel yang signifikan pada penelitian ini yakni variabel Tingkat Pengangguran Terbuka (X1), Kemiskinan (X3), Rata Lama Sekolah (X4) dan Angka Partisipasi Sekolah (X6). Dari variabel tersebut dapat memberikan pertimbangan

pemerintahan Provinsi Maluku dalam menentukan kebijakan Pemerintah dalam meningkatkan Indeks Pembangunan Manusia di Maluku.

3. Bagi penelitian selanjutnya dapat menambahkan variabel penelitian baik itu variabel independen, yang dapat berhubungan dengan indeks pembangunan manusia, karena di dalam penelitian ini kemampuan variabel independen dalam menjelaskan variabel dependen (nilai perusahaan) masih terbatas.

## DAFTAR PUSTAKA

- Erly Nofriyanti M., dan Francis Hutabarat. (2021, November). *Pengaruh Angka Harapan Lama Sekolah, Rata-Rata Lama Sekolah, Pengeluaran per Kapita Terhadap Indeks Pembangunan Manusia*. Retrieved from Jurnal Ilmiah Akuntansi Manajemen: <https://www.jurnal-umbuton.ac.id/index.php/jiam/article/view/1718/1067>
- Fathurahman M. (2012, Mei). *Metode Cochran-Orcutt untuk Mengatasi Autokorelasi*. Retrieved from Jurnal Eksponensial: <https://fmipa.unmul.ac.id/files/docs/5.Jurnal%20Pak%20Fathur%20.pdf>
- Kusuma Guntur W., dan Wulansari Ika Y. (2020, Mei 12). *Analisis kemiskinan dan kerentanan kemiskinan dengan Regresi Ridge, LASSO, dan Elastic-Net di Provinsi Jawa Tengah tahun 2017*. Retrieved from Seminar Nasional Official Statistics: <https://prosiding.stis.ac.id/index.php/semnasoffstat/article/view/189>
- Mahrany Yunita. (2012). *Pengaruh indikator komposit indeks pembangunan manusia terhadap pertumbuhan ekonomi di Sulawesi Selatan*. Retrieved from Skripsi: Sarjana Fakultas Ekonomi dan Bisnis Universitas Hasanuddin Makassar: <https://core.ac.uk/download/pdf/25487666.pdf>
- Ningsih Sri R., Damanik Irfan S., Windarto Agus P., Tambunan Heru S., Jalaluddin, dan Wanto Anjar. (2019, September). *Analisis K-Medoids Dalam Pengelompokan*. Retrieved from PROSIDING SEMINAR NASIONAL RISET INFORMATION SCIENCE: <https://tunasbangsa.ac.id/seminar/index.php/senaris/article/view/78/79>
- Pritha Bose. (2023, August 25). *What are Lasso and Ridge Techniques?* Retrieved from Analytixlabs: <https://www.analytixlabs.co.in/blog/lasso-and-ridge-regression/>
- Raehani Rahmawati. (2023, November 27). *Perbandingan Regresi Ridge dan Principal Component Regression dalam Mengatasi Multikolinieritas pada Faktor-Faktor yang Mempengaruhi Kemiskinan di Indonesia*. Retrieved from Universitas Mataram Repository: <http://eprints.unram.ac.id/id/eprint/43629>
- Sumarah Jati, dan Wulandari Ajeng Tiara . (2021, Oktober). *Pemanfaatan Algoritma K-Means untuk Pengelompokan Angka* . Retrieved from Jurnal Media Informatika Budidarma: <http://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/3277/224>