

Big Data Assignment

Multiple Linear Regression:

Life Expectancy



MULTIPLE LINEAR REGRESSION

Multiple Linear Regression (MLR) merupakan suatu metode yang digunakan untuk memodelkan hubungan linear antara variabel dependen dengan satu atau lebih variabel independen. Variabel dependen adalah variabel yang dipengaruhi atau variabel yang menjadi akibat karena adanya variabel independen. Sedangkan variabel independen adalah variabel yang menjadi penyebab adanya perubahan pada variabel dependen atau variabel yang mempengaruhi variabel lain.

The diagram illustrates the Multiple Linear Regression equation:
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
 Each term is labeled with an arrow:

- Y_i is labeled "Dependent Variable".
- β_0 is labeled "Population Y intercept".
- β_1 is labeled "Population Slope Coefficient".
- X_i is labeled "Independent Variable".
- ϵ_i is labeled "Random Error term".

 Brackets at the bottom group the terms into two components:

- A blue bracket under $\beta_0 + \beta_1 X_i$ is labeled "Linear component".
- A blue bracket under ϵ_i is labeled "Random Error component".



Life Expectancy Dataset

Intro

Life expectancy atau harapan hidup merupakan ukuran statistik atau waktu rata-rata suatu organisme diharapkan untuk hidup berdasarkan faktor demografis yang telah ditentukan.

Info

Data dikumpulkan dari situs WHO dan Perserikatan Bangsa-Bangsa dengan bantuan Deeksha Russell dan Duan Wang.

Sumber Dataset :

<https://www.kaggle.com/kumarajarshi/life-expectancy-who>



Business Understanding

Ruang Lingkup Bisnis : Kesehatan

Secara ringkas penelitian ini memfokuskan pada faktor imunisasi, faktor mortalitas, faktor ekonomi, faktor sosial dan faktor lain yang berhubungan dengan kesehatan dimana faktor tersebut dapat mempengaruhi harapan hidup seseorang. Karena pengamatan dataset ini didasarkan pada negara yang berbeda, maka akan lebih mudah bagi suatu negara untuk menentukan faktor yang berkontribusi terhadap nilai harapan hidup yang lebih rendah. Ini akan membantu dalam menyarankan suatu negara, area penting mana yang harus diperhatikan untuk meningkatkan harapan hidup penduduknya secara efisien.



Data Understanding (I)

Field Name	Description	Data Type
Country	<i>Negara</i>	object
Year	<i>Tahun</i>	int64
Status	<i>Negara maju atau berkembang</i>	object
Life expectancy	<i>Harapan hidup dalam usia (tahun)</i>	float64
Adult Mortality	<i>Tingkat kematian orang dewasa (ribuan)</i>	float64
infant deaths	<i>Jumlah kematian bayi (ribuan)</i>	int64
Alcohol	<i>Konsumsi alkohol per-kapita (liter)</i>	float64
percentage expenditure	<i>Persentase belanja terkait kesehatan (%)</i>	float64



Data Understanding (II)

Field Name	Description	Data Type
Hepatitis B	<i>Imunisasi Hepatitis B (%)</i>	float64
Measles	<i>Jumlah terkena campak (ribuan)</i>	int64
BMI	<i>Nilai rata-rata Body Mass Index</i>	float64
under-five deaths	<i>Tingkat kematian bawah usia 5 tahun (ribuan)</i>	int64
Polio	<i>Imunisasi polio (%)</i>	float64
Total expenditure	<i>Belanja pemerintah terkait kesehatan (%)</i>	float64
Diphtheria	<i>Imunisasi difteri (%)</i>	float64
HIV/AIDS	<i>Tingkat kematian akibat HIV/AIDS (ribuan)</i>	float64

Data Understanding (III)

Field Name	Description	Data Type
GDP	Gross Domestic Product per kapita (USD)	float64
Population	Total populasi suatu negara	float64
thinnes 1-19 years	Prevalensi kurus pada anak usia 10-19 thn (%)	float64
thinnes 5-9 years	Prevalensi kurus pada anak usia 1-9 thn (%)	float64
Income composition of resources	Komposisi pendapatan	float64
Schooling	Lama mengikuti pendidikan formal (tahun)	float64

Data Understanding (IV)

Life Expectancy Dataset

Terdapat 2.938 baris data dan 22 kolom

Terdapat nilai kosong (null value) pada beberapa kolom, sehingga perlu ditangani agar kualitas data baik untuk dilakukan pemodelan



Data Preparation

◆ Menangani null value

Pada kolom yang memiliki nilai kosong (null value) diisi dengan nilai tengah (median) dari keseluruhan data yang terisi pada kolom tersebut.

◆ Menghapus kolom tertentu

Kolom yang tidak digunakan dalam pemodelan dihapus dari DataFrame. Kolom dimaksud merupakan kolom dengan tipe data Object (String) dan kolom yang memiliki nilai p value > 0.05 .

◆ Mengecek ringkasan statistik dan hubungan antar variabel

Pengecekan ringkasan data statistik pada dataset yang digunakan. Disamping itu, dilakukan pengecekan hubungan linear antara variabel independen dengan variabel dependen

Modeling (Training)

- ✓ **Menentukan Variabel**
Variabel X sebagai variabel independen terdiri dari semua variabel kolom selain kolom "Life Expectancy". Sementara variabel Y sebagai variabel dependen merupakan variabel kolom "Life Expectancy".
- ✓ **Membagi Data**
Dari keseluruhan jumlah data dibagi menjadi dua bagian, yaitu data training dan data testing dengan komposisi 50:50.
- ✓ **Menghitung Nilai MLR**
Nilai MLR diperoleh dengan menggunakan function `LinearRegression()` pada data training, sehingga menghasilkan nilai koefisien setiap variabel independen dan nilai intercept. Pada tahap ini menghasilkan variabel pemodelan `lin_reg`.





Evaluation (Testing and Accuracy)

Pengetesan (Testing)

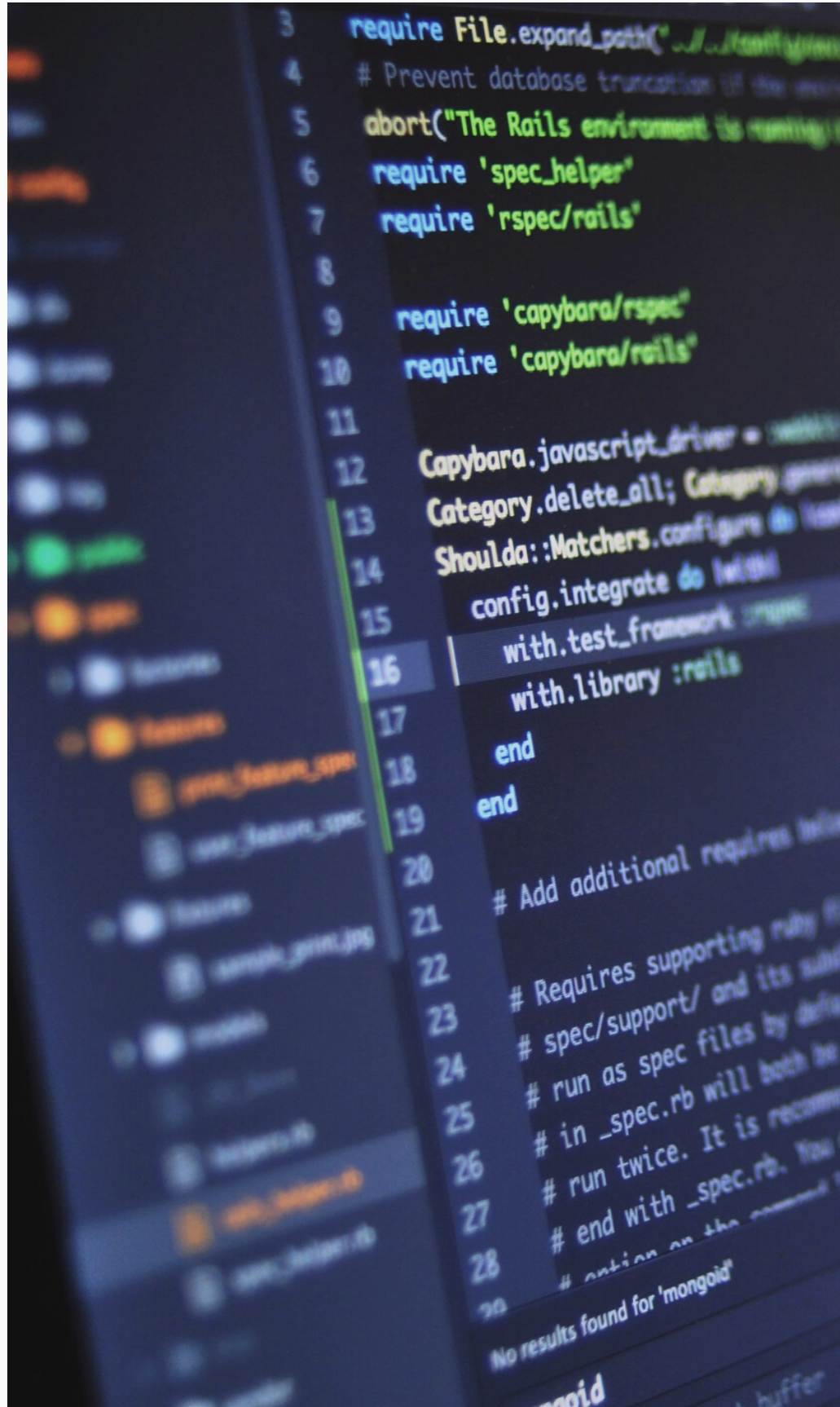
Testing dilakukan dengan menjalankan hasil pemodelan (lin_reg) terhadap data testing.

Pengecekan Akurasi (I)

Pengecekan akurasi dengan menggunakan function score() pada pemodelan MLR (lin_reg) memiliki nilai R-squared 0.81955376217453. Nilai ini menunjukkan bahwa akurasi pemodelan yang telah dibuat sebesar 81,95%.

Pengecekan Akurasi (II)

Pengecekan akurasi selanjutnya menggunakan distribution plot dengan mengkombinasikan nilai prediksi dengan nilai sebenarnya. Pada distribution plot menunjukkan bahwa nilai prediksi hampir mendekati nilai sebenarnya.



Deployment

Tahap ini dilakukan menggunakan bahasa pemrograman Python dengan IDE Visual Studio Code.

Code secara keseluruhan terdapat pada link Github :

<https://github.com/dyanaagustina/Learn-MLR-Python>

Thank You

