BAB 2
LANDASAN TEORI

Berikut merupakan teori-teori yang mendasari penelitian untuk melakukan prediksi peningkatan kasus kematian COVID-19 di Indonesia dengan menggunakan model algoritma *Linear Regression* dan melakukan evaluasi terhadap akurasi model yang telah dibuat dengan menggunakan perhitungan *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE) dan *R Squared* (R^2).

2.1 Linear Regression

Linear Regression menyesuaikan model linear dengan koefisien w = (w1,...,wp), w adalah koefisien untuk meminimalkan jumlah sisa kuadrat antara target yang diminati dalam dataset , dan target yang diprediksi dengan pendekatan linear[11]. Model regresi memiliki banyak varian, seperti regresi linier, regresi ridge, regresi bertahap, dan regresi polinomial. Regresi linier adalah model sederhana yang digunakan untuk mencari hubungan antara variabel dependen dan variabel independen. Persamaan 1 menunjukkan hubungan antara dependen (kasus baru Covid-19 dari awal penelitan dimulai hingga akhir tahun 2020) dan variabel independen. Setiap analisis univariat dalam model regresi linier digunakan untuk menunjukkan seberapa besar setiap variabel independen akan diprediksi oleh variabel dependen. Analisis multivariat juga digunakan untuk menentukan variabel prediktor terbanyak untuk jumlah total kasus baru kematian COVID-19.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_p X_p + \varepsilon \tag{2.1}$$

- Y = prediksi
- β_p = intersep variabel
- X_p = p-idependen
- ε = Error

Y adalah jumlah total kasus baru kematian Covid-19 dan X_1,X_2,..., dan X_p adalah p-idependen yang berarti nilai p untuk setiap variabel independen menguji

hipotesis nol bahwa variabel tersebut tidak memiliki korelasi dengan variabel dependen. β_0,β_1,β_2,..., dan β_p masing-masing adalah intersep dan koefisien variabel sedangkan ε sebagai istilah kesalahan yang terjadi dalam model. Berikut merupakan persamaan yang dipakai dalam mencari nilai prediksi garis linear :

$$Y(pred) = b_0 + b_1.x \tag{2.2}$$

- Y = prediksi
- b_0 = intersep
- b_1x = slope

Jika nilai $b_1 > 0$, maka x (preditor) dan y (target) memiliki hubungan positif yang berarti peningkatan x akan meningkatkan y dan jika nilai $b_1 < 0$, maka x (preditor) dan y (target) memiliki hubungan negatif yang berarti peningkatan x akan menurunkan y. Berikut merupakan persamaan dari b_1:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.3}$$

- b_1 = koefisien regresi
- x_i = nilai sebenarnya x
- y_i = nilai sebenarnya y
- $\bar{x}$ = nilai rata-rata x
- $\bar{y}$ = nilai rata-rata y

Untuk nilai b_0 jika model tidak terdapat nilai x = 0 maka prediksi akan menjadi tidak dapat digunakan.Jika model menyertakan nilai 0 , maka b_0 akan menjadi rata-rata dari semua nilai yang diprediksi saat x = 0 .Namun menetapkan nilai 0 untuk semua variabel seringkali tidak dapat dilakukan.Dikarenakan nilai b_0 menjamin nilai residual mean 0,jika dalam perhitungan tidak menggunakan b_0 maka regresi akan melewati paksa data asal sehingga koefisien regresi dan prediksi akan menjadi bias. Berikut merupakan persamaan dari b_0 :

$$b_0 = \bar{y} - b_1\bar{x} \tag{2.4}$$

- b_0 = intersep
- b_1 = koefisien regresi
- $\bar{y}$ = nilai rata-rata y
- $\bar{x}$ = nilai rata-rata x

untuk nilai b_0 dan b_1 harus dipilih sehingga dapat meminimalkan kesalahan jika terdapat kesalahan pada kuadrat yang diambil sebagai metrik untuk mengevaluasi model maka akan mendapatkan hasil garis yang dapat mengurangi *error* atau kesalahan.

$$Error = \sqrt{\sum_{i=1}^{n} a - b} \tag{2.5}$$

- a = actual output
- b = predicted output

Berikut merupakan *pseudocode* sederhana dari algoritma *Linear Regression*.

Algorithm 1 Linear Regression

```
Read number n :
for  i=1 in n : do
    Read X_i and Y_i
end for
.
.
Initialize :
X = 0
X_1 = 0
Y = 0
Y_1 = 0
.
.
for i=1 in n do
    X = X + X_i
    X_1 = X_1 + X_i * X_i
    Y = Y + X_i
    Y_1 = Y_1 + X_i * X_i
end for
.
.
Calculate Required constant a and b of y = a + bx :
b = (n* XY - X * Y)/(n*X_1 - X * X)
a = (Y - b*X)/n
.
.
Display value a and b :
print(a)
print(b)
```

2.2 *Polynomial Features*

Polynomial Features adalah fitur yang dibuat dengan menaikkan fitur yang ada ke eksponen. Contohnya jika ada dataset yang memiliki fitur input X, maka fitur polinomial akan membuat fitur baru dengan mengkuadratkan nilai X. Proses pengkuadratan ini dapat diulang untuk setiap variabel input yang ada dalam dataset, sehingga membuat versi transformasi dari masing-masing variabel. Sehingga *Poly-*

nomial Features juga dapat dikatakan sebagai jenis rekayasa fitur dikarenakan membuat fitur input baru berdasarkan fitur input yang sudah ada. Penggunaan derajat dalam polinomial digunakan untuk mengontrol jumlah fitur yang ditambahkan, contohnya derajat 3 akan menambahkan dua variabel baru untuk setiap variabel input, pada umumnya derajat kecil yang digunakan adalah 2 atau 3. *"Secara umum, tidak pada biasanya menggunakan d lebih besar dari 3 atau 4 dikarenakan untuk nilai d yang terlalu besar kurva polinomial dapat menjadi terlalu fleksibel dan dapat mengambil beberapa bentuk yang sangat aneh"* [12].

Pada umum untuk menambahkan variabel baru yang mewakili interaksi antar fitur, misalnya kolom baru yang mewakili satu variabel dikalikan dengan variabel lainnya. Sebuah versi kuadrat atau pangkat tiga dari variabel input akan mengubah distribusi probabilitas, memisahkan nilai-nilai kecil dan besar, pemisahan yang meningkat dengan ukuran eksponen. Sehingga penggunaan *Polynomial Features* pada *Linear Regression* dapat direspon dengan sangat baik dan efektif untuk mengidentifikasi pola nonlinear [13]. Penggunaan *polynomial feature* dalam penelitian berutujuan untuk menghasilkan prediksi dengan hasil yang maksimum sehingga dapat menghasilkan prediksi dengan tingkat akurasi yang tinggi, dikarenakan *Polynomial Features* memiliki kecocokan algoritma perhitungan data terhadap dataset yang digunakan didalam penelitian.

2.3 *Coronavirus Disease 2019* (COVID-19)

Pada bulan Desember 2019, telah terjadi penularan virus baru yang dikenal sebagai SARS-CoV-2 atau yang dikenal sebagai COVID-19 sebutan tersebut secara resmi diberikan oleh WHO, COVID-19 merupakan virus yang dapat menyebar secara pesat [1].COVID-19 pertama kali ditemukan di Wuhan China. COVID-19 diidentifikasi sebagai penyakit pneumonia.COVID-19 memiliki efek yang beragam pada tubuh manusia termasuk sindrom pernafasan akut yang parah dan kegagalan pada fungsi organ dan pada akhirnya dapat menyebabkan kematian yang sangat singkat [2].Sejak terjadinya penyebaran yang cukup berbahaya sehingga hampir seluruh negara mendeklarasikan penutupan wilayah dan kota yang terkena dampak salah satunya Indonesia.Berdasarkan data Covid19.go.id telah mencatat sebanyak 1.583.182 terkonfirmasi COVID-19, 1.431.892 dinyatakan sembuh dan 42.906 lainya dinyatakan meninggal pada tanggal 14 April 2021[14].

2.4 Teknik Evaluasi

Dalam melakukan evaluasi model terdapat beberapa perhitungan yang dapat dilakukan yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE) dan *R Squared* (R^2). Perhitungan *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE) dapat dikatakan diterima apabila hasil yang dihasilkan dari perhitungan mendekati angka 0, hal ini berdasarkan tabel berikut :

<10	Tingkat Prediksi Tinggi
10-20	Prediksi dapat diterima
20-50	Prediksi dapat ditoleransi
>50	Tidak dapat diterima
Source:	Lewis 1982,p.40

Dan dalam perhitungan evaluasi R^2 dapat dikatakan diterima apabila hasil yang dihasilkan mendekati angka 1, hl ini berdasarkan tabel berikut :

R=0	Tidak ada korelasi antara dua variabel
$0 < R \leq 0,25$	Korelasi sangat lemah
$0,25 < R \leq 0,50$	Korelasi cukup
$0,50 < R \leq 0,75$	Korelasi kuat
$0,75 < R \leq 0,99$	Korelasi sangat kuat
R = 1,00	Korelasi sempurna

Dengan menggunakan perhitungan tersebut diharapkan dapat menghasilkan model yang cukup akurat terhadap hasil prediksi yang dihasilkan.

2.4.1 *Mean Absolute Error* (MAE)

Mean Absolute Error (MAE) adalah metode yang digunakan dalam pengukuran tingkat keakuratan model prediksi.Nilai MAE menunjukan rata-rata kesalahan absolut antara hasil prediksi dengan nilai riil [15].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{2.6}$$

- MAE = Mean Absolute Error
- n = jumlah data
- x_i = nilai hasil prediksi

- x = nilai sebenarnya

2.4.2 *Root Mean Square Error* (RMSE)

Root Mean Square Root (RMSE) adalah metode penjumlahan kuadrat error atau selisih antara nilai riil dan nilai prediksi RMSE dihitung dengan mengkuadratkan error , dibagi dengan jumlah data rata-rata, lalu diakarkan [16].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{2.7}$$

- RMSE = Root Mean Square Error
- n = jumlah data
- x_i = nilai sebenarnya
- $\hat{x}_i$ = nilai prediksi

2.4.3 *Mean Absolute Percentage Error* (MAPE)

Mean Absolute percentage Error (MAPE) adalah metode yang menggunakan kesalahan absolute pada setiap periode yang dibagi dengan nilai observasi riil.Penggunaan MAPE mengindikasi seberapa besar kesalahan dalam prediksi yang membandingkan nilai riil dalam deret[16].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{\hat{x}_i}\right| \tag{2.8}$$

- MAPE = Mean Absolute percentage Error
- n = jumlah data
- x_i = nilai sebenarnya
- $\hat{x}_i$ = nilai prediksi

2.4.4 R2 Score

R2 Score adalah Koefisien determinasi yang digunakan untuk mengevaluasi kinerja model regresi linier. Ini adalah jumlah variasi dalam atribut dependen output yang dapat diprediksi dari variabel independen input [17].

$$SSR = \sum(x_i - \bar{x}) \tag{2.9}$$

- SSR = *Sum Squared Regression*
- x_i = nilai hasil prediksi
- $\bar{x}$ = nilai rata-rata x

$$SST = \sum(x - \bar{x}) \tag{2.10}$$

- SST = *Sum Squared Total*
- x = nilai sebenarnya
- $\bar{x}$ = nilai rata-rata

$$R^2 = 1 - \frac{SSR}{SST} \tag{2.11}$$

- R^2 = R Squared
- SSR = SSR = *Sum Squared Regression*
- SST = *Sum Squared Total*