

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Terdapat beberapa penelitian terdahulu yang melakukan implementasi menggunakan Support Vector Regression (SVR) dan ARIMA yang menjadi referensi bagi penulis untuk melakukan penelitian ini.

1. **Comparison of Time Series ARIMA Model and Support Vector Regression** [14]. Pada penelitian ini, dilakukan perbandingan metode SVR dan ARIMA dengan menggunakan tiga *dataset* urut waktu yang berbeda dan tidak saling terkait yang bertujuan untuk memprediksi kejadian pada masa mendatang dengan membandingkan kedua model tersebut. Dari hasil penelitian ini, SVR menghasilkan prediksi yang tepat dalam analisis prediksi waktu dengan menggunakan pengujian RMSE dengan hasil masing-masing pada tiga *dataset* yaitu SVR sebesar 0.88, 1.23, 0.36 dan ARIMA sebesar 355, 128, 45. RMSE dinilai memberikan hasil pengujian yang tepat untuk data urut waktu atau *time series*. Kemampuan SVR dalam mengatasi *overfitting* dan data *non-linear* dinilai menggeneralisasi data dengan baik. Berdasar hal itu, penulis menggunakan SVR untuk melakukan prediksi menggunakan data terkonfirmasi COVID-19 di Indonesia dengan *dataset* dengan urut waktu pada Maret 2020 - September 2023.
2. **Support Vector Regression (SVR) Model for Forecasting Number of Passengers on Domestic Flights at Sultan Hasanuddin Airport Makassar** [15]. Dalam penelitian ini, dilakukan penelitian untuk memprediksi jumlah penumpang

penerbangan dengan membandingkan metode ARIMA dan SVR. Dalam pemodelan ini didapatkan parameter terbaik untuk pemodelan ARIMA serta dengan kernel dan parameter terbaik untuk pemodelan SVR. Pemodelan SVR dinilai lebih baik dengan nilai pengujian MAPE yaitu 11,3812 yang lebih kecil dibandingkan dengan pemodelan ARIMA yaitu 12,3207.

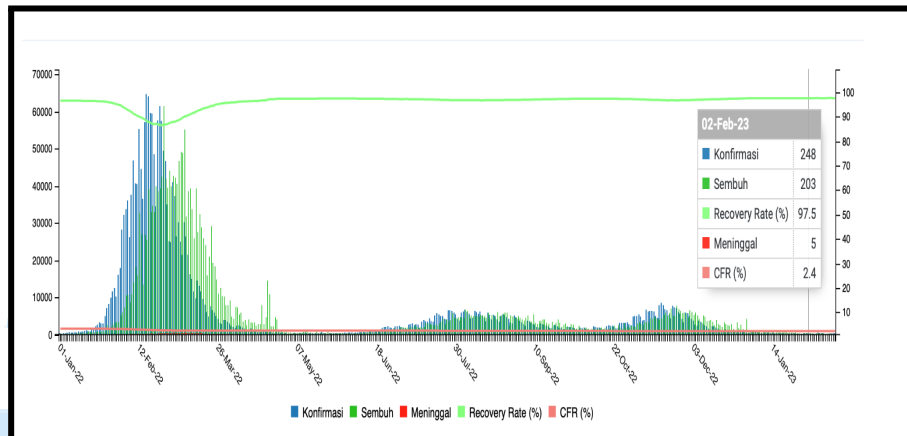
- 3. Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models [16]** . Dalam penelitian ini dilakukan perbandingan beberapa metode *time series* yaitu ARIMA, SARIMA dan Prophet Model. Dari hasil penelitian ini, ketiga metode diatas dinilai cukup baik dalam memprediksi secara akurat perkembangan suatu wabah atau virus yang menjadi endemi. Penelitian ini menggunakan *dataset* akumulatif COVID-19 dari negara USA , Brazil dan India dengan berdasar data total kasus baru. Dalam penelitian ini digunakan pengujian RMSE, MAE dan MAPE untuk pengujian dan didapatkan jika ARIMA mendapat nilai *error* yang lebih kecil dibanding pemodelan SARIMA dan Prophet Model. Berdasarkan hasil tersebut, penelitian ini menilai jika ARIMA cocok digunakan dalam memprediksi *trend* kasus aktif dari COVID-19.
- 4. A network analysis and support vector regression approaches for visualising and predicting the COVID-19 outbreak in Malaysia [17]**. Penelitian ini menggunakan metode SVR dalam memprediksi kasus aktif COVID-19 di Malaysia menggunakan *dataset* dari Desember 2020-Maret 2021. Dilakukan pengujian menggunakan MSE dengan hasil 0.151 dan RMSE 0.388 untuk *daily confirmed cases* dan dapat dinilai jika SVR menghasilkan pemodelan yang baik dalam memprediksi kasus terkonfirmasi

COVID-19 . Dalam kesimpulan penelitian ini, disarankan untuk melakukan perbandingan dengan pemodelan regresi logistik seperti pemodelan ARIMA dan LSTM khususnya dalam memprediksi kasus COVID-19. Berdasarkan hasil tersebut, peneliti menggunakan data terbaru dari Maret 2020 - September 2022 menggunakan pemodelan ARIMA dan membandingkan dengan hasil pemodelan SVR pada data yang sama

2.2 Tinjauan Teori

1. *Coronavirus Disease 2019 (COVID-19)*

Coronavirus Disease 2019 (COVID-19) merupakan kelompok virus pada hewan atau manusia yang menyebabkan infeksi pada saluran pernapasan dan biasanya dimulai dengan gejala pilek, batuk sampai dengan sesak napas. COVID-19 ditemukan pada bulan Desember 2019 di Wuhan, Tiongkok, China dan menjadi pandemi yang mempengaruhi banyak kegiatan serta perubahan di masyarakat seluruh dunia. Virus COVID dapat tersebar melalui percikan dari hidung atau mulut serta membuat manusia harus menjaga jarak untuk berbicara dan berinteraksi [18]. Sampai pada bulan Februari 2023, masih terdapat 6.733.697 kasus terkonfirmasi COVID di Indonesia [19]. Oleh karena itu , masih dibutuhkan pemodelan untuk prediksi kasus terkonfirmasi COVID-19 di Indonesia.



Gambar 2.1 Grafik Kasus COVID-19 di Indonesia

2. Time series

Pada penelitian ini, penulis menggunakan *dataset time series* yaitu pengukuran berulang dari waktu ke waktu dan diukur secara konsisten pada jarak interval yang sama. [20] *Time series* terjadi secara berurutan dalam beberapa periode waktu dan memungkinkan data dikumpulkan sehingga memberikan informasi serta variabel dalam *time series* dapat berubah sewaktu-waktu. Menggunakan *dataset time series* sebagai prediksi dapat menggunakan data terdahulu dan melakukan pemodelan atau kalkulasi untuk prediksi di masa yang akan datang terkait dengan variabel yang diteliti. *Time series* banyak digunakan untuk menganalisis sesuatu yang sedang *trend* atau yang baru saja terjadi sehingga dilakukan uji stasioner untuk mendapatkan mengetahui jika data bersifat *linear* atau konstan. Hal ini dikarenakan pemodelan ARIMA dapat dilakukan jika data bersifat *linear*.

3. Uji Stasioner

Dalam analisis *time series* [21], dibutuhkan data yang stasioner. Data dikatakan stasioner jika nilai rata-ratanya tidak

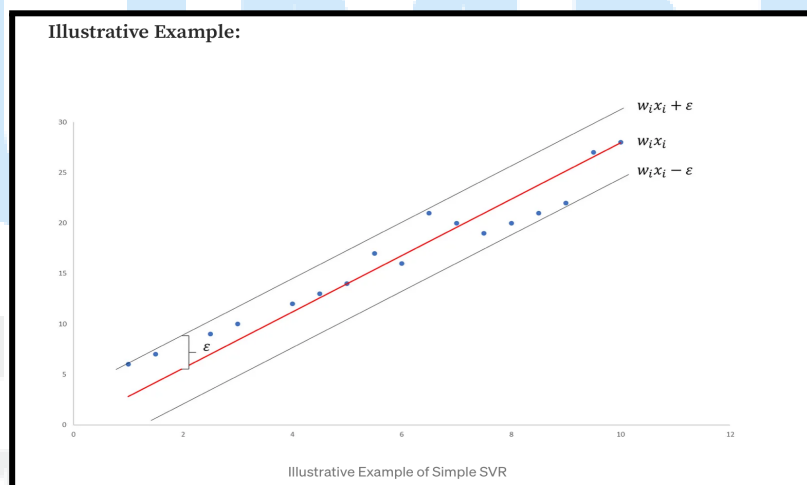
memiliki perubahan yang sistematis sepanjang waktu atau memiliki nilai yang konstan atau *linear*. Salah satu pengujian stasioner yang umum digunakan adalah metode *Augmented Dickey-Fuller Test* (*ADF*). Hipotesis pengujian dari *Augmented Dickey-Fuller* adalah:

- Jika nilai $-p > 0,05$ maka data tidak stasioner
- Jika nilai $-p \leq 0,05$ maka data stasioner
- Jika tes statistik $ADF > critical\ value$ maka data tidak stasioner
- Jika tes statistik $ADF < critical\ value$ maka data stasioner

Jika data memenuhi hipotesis diatas, maka data telah bersifat stasioner.

4. Support Vector Regression (SVR)

Support Vector Regression merupakan metode Support Vector Machine (SVM) pada metode regresi dengan tujuan untuk mencari fungsi sebagai suatu *hyperplane* (garis pemisah) dengan membuat *error* sekecil mungkin [22]. Dalam SVR kesalahan absolut yang kurang atau sama dengan margin disebut dengan kesalahan maksimum, ϵ (epsilon).



Gambar 2.2 Persamaan SVR

Tujuan nya ialah menemukan nilai fungsi $f(x)$ yang memiliki deviasi paling tinggi ϵ yang berasal dari semua *data training* .

$$f(x) = (w, x) + b \text{ with } w \in X, b \in R \quad (2-1)$$

Yang dimana $(,)$ menunjukkan perkalian dalam X dan mencari w paling kecil dengan melakukan optimisasi dengan :

$$\text{minimize } \frac{1}{2} \|w\|^2 \quad (2-2)$$

$$\text{subject to } \begin{cases} y_i - (w, x_i) - b \leq \epsilon \\ (w, x_i) + b - y_i \leq \epsilon \end{cases} \quad (2-3)$$

Melakukan asumsi bahwa fungsi f mendekati x_i, y_i dengan presisi pada ϵ yang dimana berarti optimisasi dapat dilakukan . Kendalanya ialah untuk meminimalisir kesalahan antara prediksi fungsi untuk *input* dan *output* dengan nilai sebenarnya.

$$\text{Min } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^t (\xi_i + \xi_i) \quad (2-4)$$

Fungsi untuk meminimalisir risiko pada pemodelan SVR dan akurasi pada *training data set* , parameter C digunakan untuk mengukur generalisasi dan akurasi pada *data training* dan parameter e untuk menentukan tingkat toleransi terhadap *error*.

5. Kernel

Kernel merupakan metode sederhana untuk memetakan data *non-linear* dan memudahkan klasifikasi data dengan menemukan

hyperplane yang dapat memisahkan dataset dengan baik dan dapat mencegah atau meminimalisir *overfitting* [23]. Nilai dari kernel dapat berupa kernel dari *linear*, RBF (*Radial Basic Function*) dan *polynomial* yang penjelasannya yaitu :

- a. Kernel Linier yang merupakan kernel yang digunakan untuk analisis data *linear* yang memiliki persamaan :

$$K(x, xi) = \text{sum}(x * xi) \quad (2-5)$$

- b. RBF (*Radial Basic Function*) merupakan kernel Gaussian yang digunakan untuk analisis dan pemodelan data yang bersifat *non-linear* dan memiliki persamaan :

$$K(x,xi) = \exp(-\text{gamma} * \text{sum}((x - xi)^2)) \quad (2-6)$$

- c. *Polynomial* merupakan *dataset* yang dinormalisasi dengan mencari nilai optimal dengan persamaan :

$$K(x,xi) = 1 + \text{sum}(x * xi)^d \quad (2-7)$$

6. Autoregressive Integrated Moving Average (ARIMA)

Autoregressive Integrated Moving Average (ARIMA) merupakan metode *regresi linear* yang digunakan untuk peramalan data deret waktu atau *time series forecasting* yang diambil dalam kurun waktu tertentu dengan menggunakan data masa lalu [24].

ARIMA terdiri dari *Autoregressive (AR)*, *Integrated (I)*, *Moving Average (MA)* dengan detail sebagai berikut :

- *Autoregressive (AR)* dilakukan dengan meregresikan variabel target dengan nilai masa lalunya dan dinamakan variabel X dan fungsi nilai Y yang diamati adalah nilai n

masa lalu dimana n adalah parameter yang dipilih dan B_0 , B_1 adalah regresi untuk melatih pemodelannya dengan persamaan berikut:

$$Y = B_0 + B_1 * Y_{lag1} + B_2 * Y_{lag2} + \dots + B_n * Y_{lag} \quad (2-8)$$

- *Integrated (I)* dilakukan dengan menerapkan *differencing* pada data untuk mendapatkan kestasioneran data menggunakan variabel Y dengan persamaan sebagai berikut.

$$Y_{forward1} - Y = B_0 + B_1 * (Y - Y_{lag1}) + B_2 * (Y_{lag1} - Y_{lag2}) + \dots \quad (2-9)$$

- *Moving Average (MA)* dilakukan dengan menghitung sama seperti *Autoregressive (AR)* tetapi dengan nilai E yang merupakan kesalahan dalam penjelasan model MA dan menampilkan nilai residual antara model dan variabel dengan persamaan :

$$Y = B_0 + B_1 * E_{lag1} + B_2 * E_{lag2} + \dots + B_n * E_{lag} \quad (2-10)$$

7. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) merupakan pengujian statistik dengan mengukur rata-rata selisih dari akar kuadrat antara nilai aktual dan nilai prediksi. *Root Mean Square Error (RMSE)* menghitung nilai *error* dari hasil prediksi Berikut persamaan RMSE .

$$\text{RMSE} = \left(\frac{\sum (Y_i - \hat{y}_i)}{n} \right)^{1/2} \quad (2-11)$$

Dengan :

n = Jumlah data

y = nilai hasil observasi

\hat{y} = nilai hasil prediksi

i = urutan pada data

8. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) adalah salah satu metode untuk mengukur tingkat keakuratan suatu model peramalan. Nilai MAE menghasilkan rata-rata *error* absolut antara hasil prediksi dengan nilai aktual dengan persamaan berikut.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |f_i - Y_i| \quad (2-12)$$

dimana :

f_i = nilai hasil peramalan

Y_i = nilai hasil aktual

n = jumlah data

UNIVERSITAS
MULTIMEDIA
NUSANTARA

9. Mean Squared Error (MSE)

Mean Squared Error (MSE) adalah kuadrat dari rata-rata kesalahan selisih antara nilai aktual dan peramalan dengan melihat selisih. MSE menghasilkan estimasi nilai *error* pada peramalan. Perhitungan *Mean Squared Error (MSE)* dilakukan dengan mengurangkan nilai data aktual dan data peramalan kemudian hasilnya dijumlahkan secara keseluruhan dan membagi dengan jumlah data. Berikut rumus persamaan *MSE* :

$$MSE = \left(\frac{1}{n}\right) * \Sigma(y_i - \hat{y}_i)^2 \quad (2-13)$$

dimana:

n = Jumlah data

y_i = Nilai sebenarnya

\hat{y}_i = Nilai prediksi

Σ = Jumlah dari seluruh data

10. R-Squared

R-Squared merupakan ukuran statistik dalam pemodelan yang menentukan besar nilai kecocokan dengan hasil sebuah pemodelan. Kualitas ukuran statistik tergantung banyak faktor seperti variabel yang digunakan tergantung beberapa faktor diantaranya model yang digunakan, ukuran variabel dan transformasi data yang digunakan. R-Squared memiliki persamaan sebagai berikut.

$$\begin{aligned} R_1^2 &= 1 - \left[(1 - R^2) \left(\frac{n-1}{n-p-1} \right) \right] = \left[1 - \frac{p-1}{n-1} \left(\frac{SSE}{SST} \right) \right] \quad (2-14) \\ &= 1 - \frac{MSE}{SST / p - 1} \end{aligned}$$

Keterangan:

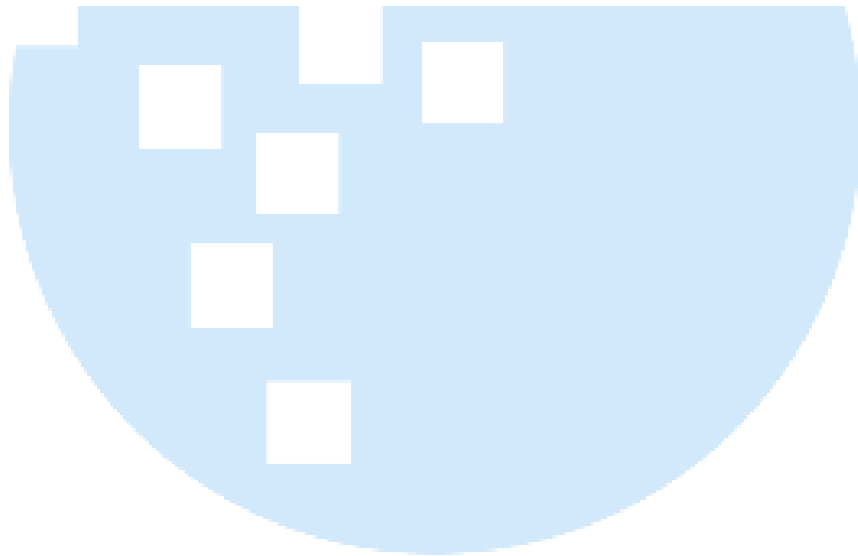
n : jumlah observasi

p : jumlah variabel

MSE : Mean Squared Error

SST : Sum Squared Total

SSE : Sum Squared Error



UMN

UNIVERSITAS

MULTIMEDIA

NUSANTARA