DECODE: Jurnal Pendidikan Teknologi Informasi
ISSN Cetak : 2775-2984 | ISSN Online : 2775-1813
Vol. 2, No. 2, September 2022, Hal: 84-95, Doi: http://dx.doi.org/10.51454/decode.v2i2.48
http://journal.umkendari.ac.id/index.php/decode

MODEL RANDOM FOREST REGRESSION UNTUK PERAMALAN PENYEBARAN COVID-19 DI INDONESIA

Diana Tri Susetianingtias[1)], Eka Patriya[1)]*, Rodiah[1)]
[1)]Universitas Gunadarma, Depok, Indonesia
Email: ekapatriya@staff.gunadarma.ac.id

Abstrak

Penyebaran COVID-19 sangat cepat yang membuat pada tanggal 27 Februrari 2020, sudah menginfeksi 78630 orang di China dan 2747 orang lainnya meninggal dunia. Keberadaan COVID-19 di Indonesia sendiri pertama kali terkonfirmasi pada tanggal 2 Maret 2020. Pada penelitian ini, peneliti akan melakukan peramalan penyebaran COVID-19 di Indonesia menggunakan metode *Random Forest Regression*. *Raw Dataset* yang digunakan adalah *dataset* yang di dapat dari situs *www.kaggle.com* yang berisikan *record* sebanyak 10695 *record* yang dirangkum dari tanggal 1 Maret 2020 hingga 21 Januari 2021. Jumlah fitur yang dimiliki *raw dataset* sebanyak 37 fitur. Proses *preprocessing* pada penelitian ini terdiri dari konversi fitur, seleksi fitur dan mendapatkan fitur untuk model. Metode seleksi fitur yang digunakan adalah *Recursive Feature Elimination* yang berhasil menyeleksi fitur dari *dataset* yang tadinya berjumlah 37 menjadi 20 fitur. Pelatihan model menggunakan *training set* yang berjumlah 8555 *record*. Peramalan menggunakan model *Random Forest Regression* akan menggunakan *validation set* yang berjumlah 2139 *record.* Hasil perhitungan *error* pada model *Random Forest Regression* tidak besar, yaitu sebesar 6.477 untuk peramalan *New Cases*, dan 0.2469 untuk peramalan *New Deaths* yang artinya hasil nilai yang diramalkan dengan nilai aktual tidak berbeda jauh.

Kata kunci: COVID-19, Fitur, Peramalan, *Random Forest Regression*, Seleksi.

RANDOM FOREST REGRESSION MODEL FOR FORECAST OF COVID-19 SPREAD IN INDONESIA

Abstract

The spread of COVID-19 is so fast that on February 27, 2020, it has infected 78630 people in China and another 2747 people have died. The existence of COVID-19 in Indonesia itself was first confirmed on March 2, 2020. In this study, researchers will forecast the spread of COVID-19 in Indonesia using the Random Forest Regression method. The used Raw Dataset is the dataset obtained from the website www.kaggle.com which contains 10,695 records summarized from March 1, 2020 to January 21, 2021. The number of features of the raw dataset is 37 features. The preprocessing process in this study consists of feature conversion, feature selection and getting features for the model. The used feature selection method is the Recursive Feature Elimination which successfully selects features from the dataset from 37 to 20 features. Model training using training set totaling 8555 records. Forecasting using the Random Forest Regression model will use a validation set totaling 2139 records. The results of the calculation of errors for the Random Forest Regression model are not large, namely 6,477 for New Cases forecasting, and 0.2469 for New Deaths forecasting, which means that the predicted value results with the actual value are not much different.

***Keywords**: COVID-19, Feature, Forecasting, Random Forest Regression, Selection.*

Submitted:	Reviewed:	Accepted:	Published:
29 Juni 2022	16 Juli 2022	29 Juli 2022	29 September 2022

PENDAHULUAN

Penyebaran wabah *Corona Virus Disease 2019* (COVID-19) sangat cepat sehingga pada tanggal 27 Februari 2020, telah tercatat 78630 orang di China terinfeksi dan 2747 diantaranya meninggal dunia (Cortegiani A, Einav S, Giarranto A, Ingoglia G., Ippolito M, 2020). Menurut data yang diperoleh (ASEAN, 2020), kasus COVID-19 yang telah terkonfirmasi tercatat sebanyak 306.644 dengan angka kematian berjumlah 7.900 di wilayah ASEAN yang merupakan himpunan negara-negara Asia Tenggara, termasuk Indonesia. Keberadaan COVID-19 di Indonesia sendiri pertama kali terkonfirmasi pada tanggal 2 Maret 2020. Penyebaran COVID-19 di Indonesia sangat cepat, sehingga pada tanggal 26 Maret 2020, telah tercatat 790 orang yang terinfeksi dan sudah menyebar ke lebih dari 17 provinsi yang menjadikan indonesia sebagai negara dengan kasus COVID-19 tertinggi di Asia Tenggara (Ryalino, 2020).

Salah satu cara melihat perkembangan COVID-19 salah satunya di Indonesia melalui metode peramalan yang dikenal dengan istilah *forecasting* (Ashadi, Asriadi, Dewi, Lamasitudju, Setialaksana, dan Sulaiman, 2020). *Forecasting* adalah suatu teknik analisa perhitungan dengan pendekatan kualitatif maupun kuantitatif untuk memperkirakan kejadian di masa depan dengan mereferensi terhadap data-data historis (Riadi, 2017). Peramalan penyebaran COVID-19 di tiap provinsi di Indonesia menggunakan *history data* yang memiliki kandungan informasi seperti *growth rate, fatality rate*, dan *case per million* perhari yang tercatat di suatu provinsi (Kaggle Dataset, 2021). Informasi yang ada pada *history data* dapat dijadikan fitur dalam melakukan peramalan penyebaran COVID-19 di setiap provinsi.

Salah satu metode dalam melakukan peramalan adalah *Random Forest Regression*. *Random Forest Regression* merupakan salah satu model regresi berdasarkan teknik *supervised learning* menggunakan fitur-fitur dari suatu *history data*. Hubungan antara fitur dan target akan diwakili oleh serangkaian kondisi yang terkait dan diatur dalam struktur seperti pohon dari atas kebawah (Baron, Reis & Shahaf, 2018). Penelitian terkait peramalan data menggunakan *Random Forest Regression* dilakukan peneliti terdahulu. Penelitian (Darst dan Malecki, 2017) menggunakan *feature selection* dengan metode *recursive feature elimination* pada model *Random Forest* dalam melakukan peramalan terhadap. Penelitian ini menggunakan kurang lebih 324 fitur yang digunakan, menggunakan model *Random Forest*. Hasil dari penelitian menunjukkan nilai RMSE 0.04. Penelitian (Luong dan Dokuchaev, 2018) menggunakan *dataset* yang didapat dari Reuters. Dataset ini merupakan data yang diambil dari 1 Januari 2008 hingga 31 Desember 2014. Langkah-langkah yang digunakan pada penelitian ini adalah menentukan model untuk melakukan *forecasting*, pengukuran volatilitas, menentukan model untuk volatilitas dan melakukan *forecasting* nilai volatilitas dengan model yang sudah ditentukan. Hasil error yang dihasilkan didalam penelitian ini diukur dengan meggunakan RMSE yang bernilai 0.0996. Penelitian (Pavlyshenko, 2018) menerapkan *Random Forest* untuk melakukan peramalan. *Dataset* yang digunakan dalam penelitian ini adalah *dataset* Rossman Store Sales yang bersumber dari Kaggle. Langkah pertama yang dilakukan pada penelitian ini adalah membuat analisis yang deskriptif seperti memvisualisasikan data dengan menggunakan beberapa tipe diagram. Pada penelitian ini dihasilkan perhitungan *error* dengan nilai sebesar 13.6%.

Penelitian ini mengimplementasikan penggunaan *Random Forest Regression* dalam melakukan peramalan penyebaran COVID-19 di Indonesia dengan menggunakan menggunakan 37 fitur yang diambil dari dataset Kaggle mulai periode 1 Maret 2020 hingga 21 Januari 2021 sejumlah 10695 *record*. Eliminasi fitur dilakukan menggunakan RFE menghasilkan 20 fitur yang akan dimasukkan ke model Random Forest *Regression*. Penelitian ini melakukan pembagian dataset dengan rasio 8 : 2 untuk data pelatihan dan data validasi. Perhitungan error untuk melihat hasil data aktual dengan data hasil peramalan menggunakan *Mean Absolute Error* (MAE). Hasil penelitian diharapkan dapat menghasilkan peramalan

penyebaran COVID-19 khususnya di Indonesia sehingga pemerintah dapat mengambil kebijakan terkait provinsi dengan angka penyebaran tertinggi.

METODE

Pada penelitian ini dilakukan beberapa tahapan-tahapan untuk membuat sistem peramalan penyebaran COVID-19 di Indonesia menggunakan metode *Random Forets.* Penelitian ini menggunakan kaggle dataset penyebaran COVID-19 (sumber : www.kaggle.com, 2022) dari beberapa negara, salah satunya negara Indonesia. *Preprocessing* terhadap *dataset* terdiri dari tiga tahapan yaitu seleksi fitur untuk mendapatkan fitur fitur yang diperlukan dan juga konversi beberapa tipe data dari suatu fitur. Setelah itu visualisasi perkembangan COVID-19 di setiap provinsi Indonesia.Kumpulan *dataset* kemudian dibagi menjadi dua jenis yaitu *training set* untuk digunakan pada pelatihan model *Random Forest Regression* dan *validation set* untuk keperluan perbandingan agar tidak tejadi *overfitting,* ujicoba peramalan, dan evaluasi kinerja model. Model yang sudah dirancang digunakan untuk melakukan peramalan Langkah dan dievaluasi menggunakan *Mean Absolute Error* (MAE).

Dataset COVID-19 sangat perlu diuraikan dalam penelitian ini. *History dataset* merupakan data perkembangan per hari di Indonesia yang digunakan dari 1 Maret 2020 – 21 Januari 2021, sejumlah 11 bulan dengan jumlah data yang diambil sebanyak 10694 *record* (sumber : www.kaggle.com, 2022). *Dataset t*erdiri dari fitur seperti Total Kematian, Kasus Baru, Kematian Baru, *Latitude*, *Longitude* dan yang lainnya. Sebagai contoh, Tabel 1 contoh bagian dari *history data* 20 hari pertama (1 Maret – 20 Maret 2020) untuk daerah DKI Jakarta dengan beberapa fitur seperti *Date, New_Cases, New_Deaths, Total_Cases, dan Total_Deaths*.

Tabel 1. Data 1 Maret – 20 Maret 2020

Date	New_Cases	New_Deaths	Total_Cases	Total_Deaths
3/1/2020	2	0	489	20
3/2/2020	2	0	491	20
3/3/2020	2	0	493	20
3/4/2020	2	0	495	20
3/5/2020	0	1	495	21
3/6/2020	0	0	495	21
3/7/2020	0	2	495	23
3/8/2020	0	0	495	23
3/9/2020	0	1	495	24
3/10/2020	0	0	495	24
3/11/2020	0	0	495	24
3/12/2020	0	1	495	25
3/13/2020	0	1	495	26
3/14/2020	1	1	496	27
3/15/2020	6	3	502	30
3/16/2020	0	3	502	33
3/17/2020	1	1	503	34
3/18/2020	0	1	503	35
3/19/2020	1	0	504	35
3/20/2020	1	3	505	38

Preprocessing Peramalan Data COVID-19 digunakan untuk mengolah data. Tahap awal pengolahan data yang dilakukan terhadap *history dataset* adalah melakukan fitur *history dataset* menggunakan metode *Recursive Feature Elimination* (RFE) dengan mengeliminasi fitur yang berlebihan yang tidak mempunyai perngaruh terhadap peramalan. Tujuan eliminasi fitur yang dilakukan secara berulang adalah untuk mendapatkan fitur yang sangat berpengaruh saat melakukan peramalan. Fitur diurutkan untuk mengukur seberapa besar fitur tersebut terhadap hasil dari peramalan. Dalam menentukan nilai peringkat fitur pada model ini, dapat dihitung menggunakan persamaan 1 seperti sebuah *tree* hanya memiliki dua *child node.*

$$ni_j = w_j c_j - w_{left(j)} c_{left(j)} - w_{right(j)} c_{right(j)} \quad (1)$$

Variabel *ni sub(j)* merupakan nilai *importance* pada *node* j, *w sub(j)* adalah nilai *weight* pada sampel saat j, *c sub(j)* merupakan nilai ketidakmurnian pada *node* j, *left*(j) adalah *child node* sebelah kiri pada *node* j dan variabel *right*(j) merupakan *child node* sebelah kanan pada *node* j Metode RFE akan melakukan pelatihan pada data latih dengan mencoba setiap fitur yang ada menggunakan Persamaan 2.

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad (2)$$

Dimana *Fi sub(i)* adalah nilai *importance* pada fitur ke I dan *Ni sub(j)* merupakan nilai *importance* pada *node* j. Pertama melatih model menggunakan *Random Forest Regression* dengan membentuk *training set*, sehingga mendapatkan *importance value* dari setiap fitur terhadap peramalan. Setelah itu fitur akan diurutkan berdasarkan *importance value* terbesar hingga terkecil. Lalu mengeliminasi fitur dengan *importance value* terkecil dan menggunakan fitur-fitur terpilih untuk melatih ulang model dan mengevaluasi kinerja dari model. Proses ini bersifat mengulang hingga subset fitur telah di habis.

Tahap seleksi fitur akan menghasilkan fitur terpilih dari proses seleksi fitur yang digunakan untuk melakukan peramalan. Fitur yang terpilih adalah fitur yang mempunyai *importance value* atau yang menduduki rangking teratas. Setelah melakukan seleksi fitur menggunakan RFE dan model *Random Forest Regression*, berhasil mendapatkan 20 fitur yang digunakan untuk melakukan peramalan. Pada penelitian ini akan dilakukan peramalan terhadap *New Cases* dan *New Deaths.* Hasil dari rangking dan seleksi fitur pada kedua *dataset* dapat dilihat pada Tabel 2 untuk *dataset* peramalan *New Cases* dan Tabel 3 untuk *dataset* peramalan *New Deaths.* Pada Tabel 2 dan Tabel 3 memiliki kriteria fitur yang terpilih untuk proses peramalan yaitu fitur dengan rangking bernilai 1 dan kriteria terpilih Ya, dimana untuk fitur terpilih akan diberi warna.

Tabel 2. Fitur Hasil Seleksi Fitur *Dataset* Peramalan *New Cases*

Nama Fitur	Rangking	Terpilih
Date	1	Ya
Location_ISO_Code	1	Ya
New_Deaths	1	Ya
New_Recovered	1	Ya
New_Active_Cases	1	Ya
Total_Cases	1	Ya
Total_Deaths	1	Ya
Total_Recovered	1	Ya
Total_Active_Cases	1	Ya
Location_Level	16	Tidak
City_or_Regency	17	Tidak

Nama Fitur	Rangking	Terpilih
Province	12	Tidak
Country	18	Tidak
Continent	19	Tidak
Island	10	Tidak
Time_Zone	15	Tidak
Special_Status	14	Tidak
Total_Regencies	11	Tidak
Total_Cities	6	Tidak
Total_Districts	5	Tidak
Total_Urban_Villages	9	Tidak
Total_Rural_Villages	13	Tidak
Area_(km2)	7	Tidak
Population	1	Ya
Population_Density	3	Tidak
Longitude	1	Ya
Latitude	1	Ya
New_Cases_per_Million	1	Ya
Total_Cases_per_Million	1	Ya
New_Deaths_per_Million	1	Ya
Total_Deaths_per_Million	1	Ya
Case_Fatality_Rate	1	Ya
Case_Recovered_Rate	1	Ya
Growth_Factor_of_New_Cases	1	Ya
Growth_Factor_of_New_Deaths	1	Ya

Tabel 3. Fitur Hasil Seleksi Fitur *Dataset* Peramalan *New Deaths*

Nama Fitur	Rangking	Terpilih
Date	2	Tidak
Location_ISO_Code	1	Ya
New_Cases	1	Ya
New_Recovered	1	Ya
New_Active_Cases	6	Tidak
Total_Cases	1	Ya
Total_Deaths	1	Ya
Total_Recovered	1	Ya
Total_Active_Cases	1	Ya
Location_Level	13	Tidak
City_or_Regency	18	Tidak
Province	7	Tidak
Country	17	Tidak
Continent	19	Tidak
Island	12	Tidak
Time_Zone	14	Tidak
Special_Status	15	Tidak
Total_Regencies	1	Ya
Total_Cities	6	Tidak
Total_Districts	1	Ya
Total_Urban_Villages	9	Tidak
Total_Rural_Villages	1	Ya
Area_(km2)	7	Tidak
Population	1	Ya
Population_Density	11	Tidak

Nama Fitur	Rangking	Terpilih
Longitude	1	Ya
Latitude	1	Ya
New_Cases_per_Million	1	Ya
Total_Cases_per_Million	1	Ya
New_Deaths_per_Million	1	Ya
Total_Deaths_per_Million	1	Ya
Case_Fatality_Rate	1	Ya
Case_Recovered_Rate	1	Ya
Growth_Factor_of_New_Cases	4	Tidak
Growth_Factor_of_New_Deaths	1	Ya

Kedua Tabel yang berisikan fitur dari kedua dataset untuk melakukan peramalan, baik peramalan *New Cases* dan *New Deaths* memiliki jumlah fitur terpilih yag sama, namun fitur-fiturnya berbeda. Sebagai contoh pada Tabel 3 menggunakan fitur *Growth_Factor_of_New_Cases* untuk melakukan peramalan *New Cases* tetapi untuk peramalan *New Deaths,* fitur *Growth_Factor_of_New_Cases* tidak digunakan.

Pembagian Dataset Menjadi Data Latih dan Validasi: Pada model *Random Forest Regression* digunakan *training set* untuk membentuk melakukan regresi yang nantinya akan digunakan untuk peramalan. Data *validation set* harus berbeda dengan data *training set* agar memiliki generalisasi yang baik dalam sebuah regresi. Pada penelitian ini telah dilakukan eksperimen denga beberapa rasio *training set* dan *validation set* untuk kedua *dataset.* Hasil dari eksperimen dapat dilihat pada Tabel 4 untuk *dataset New Cases* dan Tabel 5 untuk *dataset New Deaths.*

Tabel 4. Hasil Eksperimen *Training & Validation Set Dataset New Cases*

Rasio	Nilai Error MAE
6:4	8.2777
7:3	7.2447
8:2	**6.4770**

Seperti yang dapat dilihat pada Tabel 4 hasil eksperimen *training set* dan *validation set* pada *dataset New Cases* didapatkan nilai *error* MAE terkecil pada rasio 8:2, sehingga rasio ini digunakan untuk peramalan penyebaran COVID-19 di Indonesia.

Pada Tabel 5 memiliki nilai *error* MAE terkecil untuk melakukan peramalan penyebaran COVID-19 di Indonesia pada rasio 8:2, sehingga rasio ini digunakan pada model *Random Forest Regression.*

Tabel 5. Hasil Eksperimen *Training & Validation Set Dataset New Deaths*

Rasio	Nilai Error MAE
6:4	0.3718
7:3	0.314
8:2	0.2469

Tahapan berikutnya adalah Implementasi Metode *Random Forest Regression* untuk Peramalan Penyebaran COVID-19 di Indonesia. Pembentukan model Regresi dilakukan dengan menentukan fitur dari data *input* yang dijadikan sebagai parameter dalam proses peramalan dan variabel yang akan diramal dengan *base learner,* yaitu *decision tree*. Pembentukan model regresi dimulai dengan pemilihan *subset training set* kemudian membentuk *node-node* pada *tree*. Pencarian *node* akan terus dilakukan hingga semua *node* pada setiap *tree* terbentuk. Selanjutnya melakukan pelatihan model *Random Forest Regression,* Pembentukan model *Random Forest Regression* dapat mempengaruhi kinerja model.

Selanjutnya pada setiap *decision tree* dikombinasikan untuk dilakukan pencarian nilai keluaran yang bertipe kontinu. Keluaran setiap *decision tree* akan dirata-ratakan berdasarkan jumlah *decision tree* yang dibuat dan dijadikan sebagai keluaran dari *Random Forest Regression.* Pembentukan model RFR menggunakan pseudocode berikut:

```
rf2=RandomForestRegressor(n_estimators=100,criterion='mse',rando
m_state=42, max_features='auto')
```

Pada penelitian in di rancang model *Random Forest Regression* yang terdiri dari *decision tree* sebanyak 100 nilai 100 adalah nilai *default* yang disediakan oleh pustaka sklearn, semakin nilai n_estimators nya tinggi maka akan menghasilkan kinerja model yang lebih baik, akan tetapi membutuhkan memori yang besar dan waktu yang lama. Model ini akan menggunakan metode MSE dalam penentuan *splitting point* pada sebuah *tree. Random State* adalah sebuah parameter dari model ini yang berfungsi untuk mengacak saat melakukan pembentukan dari setiap *tree* yang dibuat. *Maximum features* adalah parameter yang digunakan untuk menentukan varian fitur yang akan digunakan. Pada penelitian ini parameter ini diatur secara *auto.*

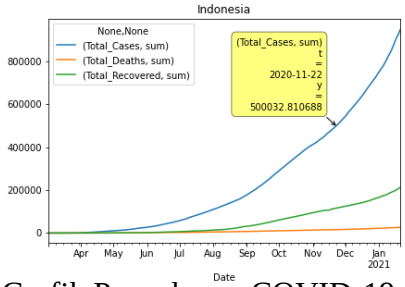
Pada penelitian ini perhitungan nilai *error* yang digunakan adalah *Mean Absolute Error* dimana menggunakan Persamaan :

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - Y'_t|}{Y_t} \qquad (3)$$

Dimana Y_t adalah nilai asli pada waktu ke-t, Y'_t merupakan nilai peramalan pada waktu ke-t dan n merupakan jumlah data.

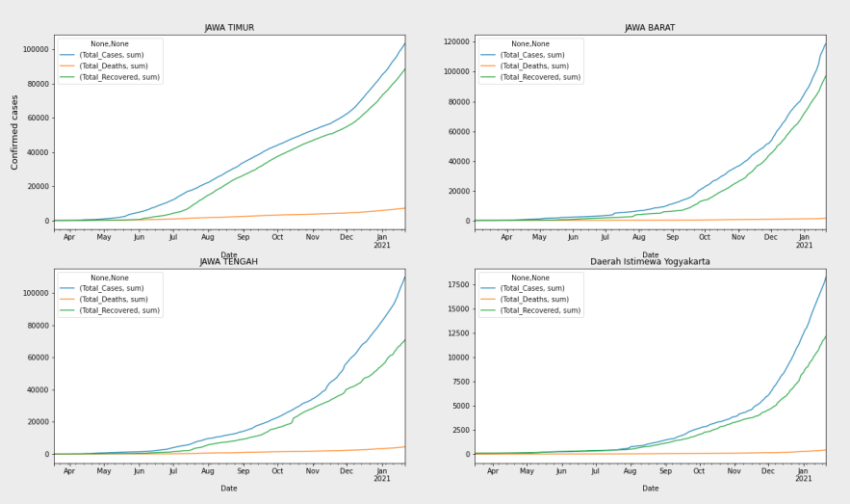
HASIL DAN PEMBAHASAN

Pada Gambar 1 menunjukan penyebaran COVID-19 yang terjadi di Indonesia. Garis yang berwarna biru menunjukan total orang yang terjangkit COVID-19, pada garis yang berwarna orange menunjukan total orang yang meninggal akibat COVID-19 dan pada garis yang berwarna hijau menunjukan total orang yang sembuh dari COVID-19.



Gambar 1. Grafik Penyebaran COVID-19 di Indonesia

Gambar 2 menunjukkan grafik Penyebaran COVID-19 di Jawa Timur, Jawa Barat, Jawa Tengah, dan DIY.



Gambar 2. Grafik Penyebaran COVID-19 di Jawa Timur, Jawa Barat, Jawa Tengah, dan DIY

Pada Tabel 6 menampilkan hasil dari nilai-nilai MAE yang merupakan nilai *error* dari perhitungan peramalan.

Tabel 6. Tabel Hasil MAE Peramalan Covid-19 di Indonesia

Peramalan	**Hasil MAE**
New Cases	6.4770
New Deaths	0.2469

Pada Tabel 7 menampilkan *score* dari peramalan *New Cases* yang telah diujikan terhadap *validation set.*

Tabel 7. *Score* Uji Coba *Validation Set* Peramalan *New Cases*

Uji Coba	***Score***
Validation Set	0.99819

Pada Gambar 3 menampilkan 5 *record* pertama pada *dataset* hasil peramalan *New Cases* di masing-masing provinsi. *Score* yang dimilki *validation set* pada *dataset* peramalan *New Cases* adalah 0.99819 yang dapat dikatakan *score* yang dimiliki pada peramalan ini adalah sejumlah 99.8%. Persentase ini menunjukan bahwa antara data aktual dan data hasil peramalan tidak berbeda jauh.

	Location	New_Cases	Date	Forecasted
0	Kalimantan Timur	44	2020-08-24	43.0
1	Banten	171	2020-10-21	142.0
2	Maluku	48	2020-10-07	51.0
3	Nusa Tenggara Barat	38	2020-11-27	37.0
4	Kepulauan Bangka Belitung	30	2020-12-03	29.0

Gambar 3. *Record Dataset* Hasil Peramalan *New Cases*

Dapat dilihat pada baris yang di beri persegi panjang berwarna biru hasil peramalan yang dilakukan oleh model *Random Forest* tidak berbeda jauh dengan nilai aktual. Nilai *New Cases* aslinya yang terdapat pada *dataset* adalah 44 sementara nilai hasil peramalan menggunakan MAE menghasilkan nilai 0.0227.

Pada Tabel 8 menampilkan *score* dari peramalan *New Deaths* yang telah diujikan terhadap *validation set.*

Tabel 8. *Score* Uji Coba *Validation Set* Peramalan *New Deaths*

Uji Coba	*Score*
Validation Set	0.99355

Pada Gambar 4 menampilkan 5 *record* pertama pada *dataset* hasil peramalan *New Deaths* di masing-masing provinsi. *Score* yang dimilki *validation set* pada *dataset* peramalan *New Deaths* adalah 0.99355 yang dapat dikatakan *score* yang dimiliki pada peramalan ini adalah sebesar 99.3%. Persentase menunjukan bahwa antara data aktual dan data peramalan tidak berbeda jauh.

	Location	New_Deaths	Forecasted
0	Kalimantan Timur	3	3.0
1	Banten	1	1.0
2	Maluku	0	0.0
3	Nusa Tenggara Barat	1	1.0
4	Kepulauan Bangka Belitung	1	1.0

Gambar 4. *Record Dataset* Hasil Peramalan *New Deaths*

Dapat dilihat pada baris yang di beri persegi panjang berwarna oren hasil peramalan yang dilakukan oleh model *Random Forest* sama dengan nilai aktual. Nilai *New Deaths* aslinya yang terdapat pada *dataset* adalah 3 dan juga nilai hasil peramalan yang dilakukan oleh model adalah 3, sehingga nilai MAE pada baris ini adalah 0 yang artinya tidak ada kesalahan dalam peramalannya.

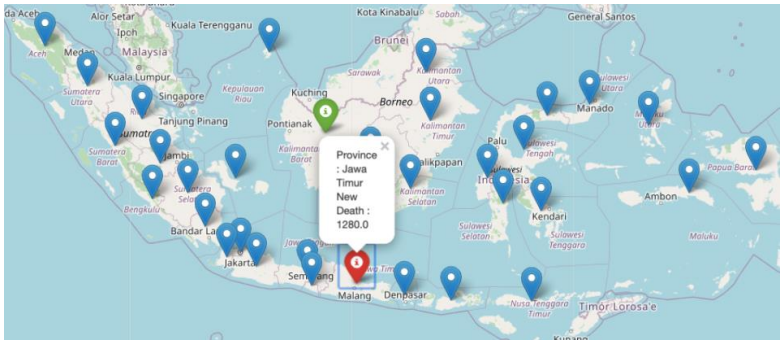
Berikut hasil visualisasi penyebaran covid-19 ke dalam peta. Hasil dari peramalan akan di gambarkan ke dalam peta untuk setiap provinsinya. Peta yang akan digambarkan mempunya 3 warna *marker* yang berbeda yaitu merah, biru, hijau. Daerah provinsi yang diberi *marker* berwarna merah menandakan bahwa yang terdampak COVID-19 sangat tinggi dibandingkan provinsi lainnya. Daerah provinsi yang diberi *marker* berwarna hijau menandakan bahwa yang terdampak COVID-19 sangat rendah dibandingkan provinsi lainnya. Gambar 5 adalah visualisasi untuk peta penyebaran COVID-19 hasil peramalan dengan *New Cases* tertinggi.

Gambar 5. Hasil Peramalan *New Cases* dengan Nilai Tertinggi

Hasil peramalan *New Cases* menunjukan Provinsi DKI Jakarta lah yang mempunyai nilai tertinggi dengan nilai 60595,oleh karena itu *marker* di Provinsi DKI Jakarta diberikan berwarna merah, sedangkan Gambar 6 merupakan hasil visualisasi dengan *New Cases* terendah.

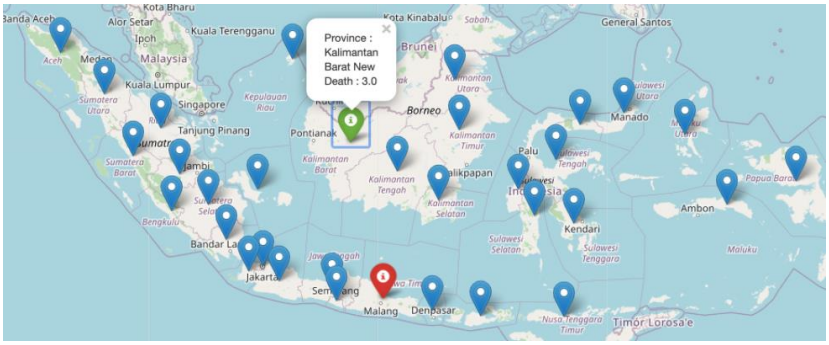
Gambar 6. Hasil Peramalan *New Cases* dengan Nilai Terendah

Marker yang berwarna hijau merupakan daerah provinsi dengan nilai *New Cases* terendah,provinsi tersebut adalah Sulawesi Barat dengan nilai 48, sedangkan pada Gambar 6 adalah hasil peramalan *New Deaths* yang sudah divisualisasi dengan peta.



Gambar 7. Hasil Peramalan *New Deaths* dengan Nilai Tertinggi

Marker berwarna merah terdapat pada Provinsi Jawa Timur dengan jumlah *New Deaths* sebanyak 1280. Kasus *New Deaths* dengan nilai terendah akan di beri *marker* berwarna hijau. Provinsi yang mendapat marker berwarna hijau adalah Kalimantan Barat.Visualisasi nilai terendah dapat dilihat pada Gambar 7.



Gambar 8. Hasil Peramalan *New Deaths* dengan Nilai Terendah

KESIMPULAN DAN SARAN

Berdasarkan hasil ujicoba peramalan dan implementasi menggunakan model *Random Forest Regressor* terhadap *history data* pada penelitian ini, dapat diambil kesimpulan antara lain: Pelatihan model berhasil membentuk *training set* dan *validation set*, dengan rasio sebesar 8:2 dimana 80% (8555 data) masuk ke dalam *testing set* dan 20% (2139 data) masuk ke dalam *validation set*. Penentuan rasio ini didasarkan dari perhitungan nilai terendah nilai *error* MAE

dengan dilakukan ujicoba dari beberapa rasio.Nilai perhitungan *error* dari hasil peramalan yang dilakukan terhadap *validation set* adalah 6.4770, untuk peramalan *New Cases* dan 0.2469 untuk peramalan *New Deaths*, dari hasil perhitungan *error* kedua ujicoba menunjukan bahwa hasil peramalan dan data aktual tidak berbeda jauh.

Pengembangan lebih lanjut dapat dilakukan dengan meningkatkan *performance* dari model Hal yang dapat dilakukan adalah dengan membuat mengubah nilai parameter yang dalam metode ini adalah pengubahan nilai *n_estimator* yang berfungsi untuk menentukan banyaknya *tree* pada saat melakukan regresi untuk peramalan, dan juga dapat mengubah nilai *random state* yang berfungsi mengacak data-data untuk dibagikan ke dalam masing masing *tree* dan melakukan pelatihan menggunakan *machine learning* lainnya seperti *XGBoost* dan *Long Short-Term Memory* (LSTM).

DAFTAR PUSTAKA

Alfiyatin, A. N., Mahmudy, W. F., Ananda, C. F., & Anggodo, Y. P. (2019). Penerapan Extreme Learning Machine (ELM) untuk Peramalan Laju Inflasi di Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *6*(2), 179-186. http://dx.doi.org/10.25126/jtiik.201962900

ASEAN. (2020). *Risk Assesment for International Dissemination of* COVID-19 *to the* ASEAN *Region.*

Browniee J. (2021). *An Introduction to Feature Selection*. Available on: https://machinelearningmastery.com/an-introduction-to-feature-selection/

Cortegiani, A., Ingoglia, G., Ippolito, M., Giarratano, A., & Einav, S. (2020). A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. *Journal of critical care, 57*, 279-283. https://doi.org/10.1016/j.jcrc.2020.03.005

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using Recursive Feature Elimination in Random Forest To Account For Correlated Variables in High Dimensional Data. *BMC genetics*, *19*(1), 1-6.

Han, J., Kamber M, Pei J. (2012). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*, Elsevier, 382-383.

Hayes A. (2021). *Chi Squre Statistic Definition*. Available on: https://www.investopedia.com/terms/c/chi-square-statistic.asp

He, F., Deng, Y., & Li, W. (2020). Coronavirus disease 2019: What we know?. *Journal of Medical Virology*, *92*(7), 719-725. https://doi.org/10.1002/jmv.25766

Heizer J., & Barry R. (2009). *Operation Management*. Buku 1 edisi 9. Jakarta: Salemba Empat

Liebeskind M. (2021). *Machine Learning Techniques for Salses Forecasting*. Available on: https://towardsdatascience.com/5-machine-learning-techniques-for-sales-forecasting-598e4984b109

Luong, C., & Dokuchaev, N. (2018). Forecasting Of Realised Volatility with The Random Forests Algorithm. *Journal of Risk and Financial Management*, *11*(4), 1-15. https://doi.org/10.3390/jrfm11040061

Malik, S., Harode, R., & Kunwar, A. S. (2020). XGBoost: A Deep Dive into Boosting (Introduction Documentation). *Simon Fraser University: Burnaby, BC, Canada*.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 15. https://doi.org/10.3390/data4010015

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic Random Forest: A Machine Learning Algorithm For Noisy Data Sets. *The Astronomical Journal*, *157*(1), 1-12.

Riadi M. (2021). *Pengertian, Fungsi dan Jenis-Jenis Peramalan*. Available on: https://www.kajianpustaka.com/2017/11/pengertian-fungsi-dan-jenis-peramalan-forecasting.html. Tanggal akses: 4 April 2021

Rustam, Z., & Maghfirah, N. (2018). Correlated Based Svm-Rfe as Feature Selection For Cancer Classification Using Microarray Databases. In *AIP Conference Proceedings* (Vol. 2023, No. 1, p. 020235). AIP Publishing LLC.

Ryalino, C. (2020). How Indonesia copes with coronavirus disease 2019 so far (part one): The country, the government, and the society. *Bali Journal of Anesthesiology*, *4*(2), 33-34.

Scikit-learn. (2021). *Scikit Learn Documentation Python*. Available on: https://scikit-learn.org/stable/

Shalev-Shwartz., Ben-David. (2013). *Understanding Machine Learning:From Theory to Algorithm (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019*

Wibawa, M. S., & Novianti, K. D. P. (2017). Reduksi fitur untuk optimalisasi klasifikasi tumor payudara berdasarkan data citra FNA. *E-Proceedings KNS&I STIKOM Bali*, 73-78.

Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, *212*, 353-363. https://doi.org/10.1016/j.snb.2015.02.025

How to cite:
Susetianingtias, D. T., Patriya, E., & Rodiah, R. (2022). Implementasi Random Forest Regression Pada Peramalan Penyebaran Covid-19 di Indonesia. *DECODE: Jurnal Pendidikan Teknologi Informasi, 2(2),* 84-95. DOI: http://dx.doi.org/10.51454/decode.v2i2.48