Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY

Agung Supoyo[1], Putri Taqwa Prasetyaningrum[2]
Universitas Mercu Buana Yogyakarta
agungsupoyo@gmail.com[1], putri@mercubuana-yogya.ac.id[2]

Abstrak - Masih tingginya kasus Covid-19 di DIY pada awal tahun 2021 ditambah dengan sulitnya mencari ruang perawatan rumah sakit, sehingga diperlukan analisis prediksi waktu perawatan. Hasil analisis sebagai pendukung keputusan Pemerintah dalam mengambil kebijakan ketersediaan kamar rumah sakit dan penerapan PPKM. Selain itu juga diperlukan analisis terhadap atribut-atribut yang paling mempengaruhi lama perawatan pasien. Penelitian menggunakan dataset yang diperoleh dari Dinas Kominfo DIY untuk kasus periode Maret sampai dengan September 2020. Diperlukan preprocessing (*data reduction, data cleaning* dan *data integration*) sebelum dilakukan analisis *data mining*. *Preprocessing* menghasilkan dataset sejumlah 271 record data dengan 31 kolom. Analisis data mining menggunakan algoritma Random Forest, k-NN dan Deep Learning menghasilkan performance model dengan RMSE masing-masing sebesar 4,949; 6,349 dan 5,436. Setelah dilakukan seleksi atribut untuk optimalisasi dihasilkan nilai RMSE sebesar 4.817 pada algoritma Random Forest dengan menggunakan 23 atribut. Hasil analisis belum cukup baik jika dibandingkan dengan rata-rata lama perawatan sebesar 15.339 hari karena menghasilkan NRMSE sebesar 31,40%. Nilai performance model dipengaruhi oleh pemilihan atribut yang digunakan. Lima atribut yang paling berpengaruh terhadap lama perawatan pasien adalah usia, jenis kelamin, kecamatan, batuk. Untuk meningkatkan performance model diperlukan penelitian lanjutan menggunakan record data yang lebih banyak dengan tambahan atribut lain seperti rumah sakit perawatan dan tindakan medis.
Kata Kunci : prediksi, lama perawatan covid-19, data mining, random forest, k-NN, deep learning

Abstract *- Covid-19 is a dangerous disease that can cause death in patients with comorbidities. In early 2021, cases of Covid-19 transmission in DIY were high, and they were coupled with the difficulty of finding hospital treatment rooms, data-based research is needed to predict patient care-time. In addition, it is also necessary to analyze the factors that most influence the length of patient care.*
The study uses a dataset of the DIY's positive covid epidemiology investigation, with a total of 3,029 lines consisting of 55 attributes, and a laboratory examination report dataset which consists of 2,823 rows of data with 15 attributes. The dataset is obtained from the Department of Communication and Informatics, DIY, for the case-period of March to September 2020. Preprocessing (data reduction, data cleaning data integration) is required before data mining analysis is carried out. Preprocessing produces a dataset of 271 rows with 32 attributes. Data mining analysis is conducted using Random Forest, k-NN and Deep Learning algorithms, which results in a performance model with RMSE of 4.949; 6,359 and 5,436. After selecting the attributes for optimization, an RMSE value of 4.853 is generated in the Random Forest algorithm using 25 attributes.The modeling results are not good enough to be used in predicting the length of treatment, because when compared with the average length of treatment of 15,339 days, the NRMSE is 31.63%. The performance model's value is influenced by the selection of the attributes used. The five attributes that most influence the length of patient care are age, gender, district, whether the patient is a health worker, and coughing. To improve the model's performance, further research is needed using more data records with the addition of other attributes that affect the length of treatment.
Keywords: prediction, length of treatment for Covid-19, Random Forest, k-NN, Deep Learning

I. PENDAHULUAN

COVID-19 adalah penyakit yang disebabkan oleh jenis baru coronavirus yang disebut SARS-CoV-2. WHO pertama kali mengetahui virus baru ini pada 31 Desember 2019, dilaporkan sebagai kelompok kasus "virus pneumonia" di Wuhan, Republik Rakyat Cina. Berdasarkan data dari website WHO, disebutkan bahwa gejala utama SARS-CoV-2 adalah demam, batuk, dan sesak napas yang dalam banyak kasus tampak mirip dengan penyakit flu.
Centers for Disease Control and Prevention (CDC) atau Pusat Pengendalian dan Pencegahan Penyakit Amerika Serikat, menyatakan bahwa 94% kematian terjadi pada pasien Covid-19 yang disertain dengan penyakit penyerta atau kondisi kesehatan bawaan (comorbidities). Sisanya, sekitar 6% kematian benar-benar disebabkan oleh virus corona SARS-CoV-2. CDC mencantumkan beberapa penyakit penyerta yang menyebabkan kematian pasien Covid-19, seperti influenza, pneumonia, gagal napas, tekanan darah tinggi, diabetes, demensia vaskular, gagal jantung, dan gagal ginjal.

Berdasarkan data yang diolah dari corona.jogjaprov.go.id, di DIY pada Maret 2021 terdapat 5.649 kasus baru atau 182 kasus infeksi per hari. Jumlah tersebut sedikit turun dari jumlah kasus pada Februari sebanyak 5.998 kasus atau setara 214 kasus per hari. Jumlah kasus pada Januari 2021 menunjukkan jumlah tertinggi selama setahun pandemi di DIY dengan mencapai 9.670 kasus atau rata-rata 312 kasus per hari. Meskipun terjadi penurunan kasus dari Januari hingga Maret 2021, hal ini juga karena jumlah tes PCR menurun. Mengutip dari portal berita tirto.id pada artikel yang berjudul "Titik Lemah Penanganan Corona di Yogyakarta: Data Bed Tak Sesuai" (Syambudi, 2021), disebutkan bahwa Rumah sakit rujukan COVID-19 semakin penuh dan pasien sulit mencari ruang perawatan, akan tetapi di sisi lain saban hari dilaporkan ada puluhan bed kosong.

Pada penelitian terkait yang berjudul *Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery* dilakukan prediksi kesembuhan pasien Covid-19 menggunakan dataset epidemiologi pasien Covid-19 di Korea Selatan. *Support Vector Machine*, *Naives Bayes*, *Logistic Regression*, *Random Forest*, dan *K-Nearest Neighbor* diaplikasikan langsung pada dataset dengan menggunakan bahasa pemrograman python. Dataset terdiri dari 5 atribut yang meliputi jenis kelamin, usia, kasus infeksi dan waktu Perawatan dan keadaan pasien. Diperoleh bahwa Decission Tree mendapatkan akurasi tertinggi dengan nilai 99,85% (Muhammad et al., 2020).

Fokus penelitian ini adalah analisis prediksi lama perawatan pasien covid-19 di DIY untuk kemudian digunakan sebagai pendukung keputusan Pemerintah dalam mengambil kebijakan terkait dengan ketersediaan kamar rumah sakit dan penerapan PPKM (Pemberlakuan Pembatasan Kegiatan Masyarakat). Selain itu diperlukan juga analisis terhadap faktor-faktor yang mempengaruhi lama perawatan pasien sebagai upaya antisipasi dalam perawatan pasien di rumah sakit.

1. Datamining

Secara sederhana dapat dipahami bahwa data mining atau dikenal juga dengan istilah knowledge discovery in database (KDD) adalah serangkaian proses yang bertujuan untuk mengekstraksi pola-pola penting atau menarik dari sejumlah data berukuran sangat besar yang tidak dapat dikenali secara manual.

Data mining merupakan bagian yang terintegrasi dari penemuan pengetahuan dalam database yang merupakan proses dengan urutan sebagai berikut.

a. Data Selection (Seleksi Data)
Data yang relevan akan diambil dari database, kemudian dilakukan analisis korelasi untuk memperoleh karakteristik data.
b. Data Cleaning (Pembersihan Data)
Dilakukan penghapusan data-data yang tidak lengkap, mengandung error, dan tidak konsisten dari dataset.
c. Data Integration (Integrasi Data)
Data dari berbagai macam sumber data dan disimpan dalam penyimpanan data yang koheren.
d. Data Transformation (Transformasi Data)
Data ditransformasikan kedalam format yang sesuai agar dapat dioleh pada proses selanjutnya.
e. Data Mining (Penambangan Data)
Implementasi algoritma mengekstrak pola data.
f. Pattern Evaluation (Evaluasi Pola)
Dilakukan identifikasi pola yang menarik untuk merepresentasikan pengetahuan berdasarkan beberapa pengukuran yang telah dilakukan.
g. Knowledge Presentation (Presentasi Pengetahuan)
Teknik visualisasi dalam merepresentasikan pengetahuan yang diperoleh.

2. Prediksi

Prediksi adalah proses memperkirakan secara sistematis apa yang paling mungkin terjadi di masa depan berdasarkan informasi yang dimiliki pada masa lalu dan sekarang, sehingga dapat meminimalkan kesalahan (perbedaan antara apa yang terjadi dan hasil yang diprediksi). Hasil prediksi tidak harus berupa hal yang jelas akan terjadi secara tepat, melainkan menemukan jawaban yang sedekat mungkin dengan apa yang akan terjadi. Pada penelitian yang dilakukan (Virdaus & Prasetyaningrum, 2020) dilakukan empat skenario atau kondisi yang dijadikan sasaran data dalam pengolahan datanya agar diperoleh hasil yang lebih variatif dan lebih presisi dengan tingkat akurasi prediksi tertinggi.

3. Random Forest

Random Forests adalah metode pembelajaran *ensemble* yang digunakan untuk klasifikasi, regresi, dan tugas lainnya yang dijalankan dengan membangun banyak pohon keputusan selama pelatihan (Primajaya & Sari, 2018). Untuk tugas klasifikasi, output dari Random Forest adalah kelas yang dipilih oleh sebagian besar trees. Untuk tugas regresi dihasilkan dari prediksi rata-rata atau rata-rata dari masing-masing tree.

Algoritma Random Forest mempunyai tiga aspek kerja utama, yaitu: (1) melakukan *bootstrap sampling* untuk membangun pohon

prediksi; (2) menggunakan prediktor acak untuk setiap pohon keputusan; (3) kemudian Random Forest menggabungkan hasil dari setiap pohon keputusan untuk membuat prediksi melalui pemungutan suara mayoritas untuk klasifikasi atau prediksi rata-rata (Gading Sadewo et al., 2017). Beberapa metode klasifikasi telah dicoba dan menghasilkan hasil yang cukup jauh berbeda, hasil terbaik adalah dengan metode Random dengan skor akurasi 88% (Prasetyaningrum et al., 2021).

4. k-Nearest Neighbour

K-NN mengimplementasikan konsep berdasarkan pada asumsi lokalitas dalam ruang data. Dalam lingkungan lokal pola x diharapkan memiliki nilai output yang sama y (atau label kelas) untuk f(x). Akibatnya, untuk x′ diketahui label harus mirip dengan label dari pola terdekat, yang dimodelkan dengan rata-rata nilai output dari sampel terdekat K (Kramer, 2011). Pada Penelitian (Haspriyanti & Prasetyaningrum, 2020), Algoritma K-Nearest Neighbor untuk penjualan produk layanan terlaris, guna untuk mempermudah pihak perusahaan dalam perencanaan penyediaan layanan.

KNN bekerja dengan prinsip mencari jarak terpendek antara data yang akan dievaluasi dengan k tetangga terdekatnya pada data training. Data training diproyeksikan ke dalam ruang multi-dimensi dengan masing-masing dimensi menggambarkan fitur data. Ruang dibagi menjadi beberapa bagian berdasarkan klasifikasi data training. Sebuah titik dalam ruang ini ditandai sebagai kategori c, jika kategori c adalah kategori yang paling umum di antara k tetangga terdekat dari titik tersebut (Whidhiasih et al., 2013).

Rumus regresi KNN adalah sebagai berikut.

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad \text{.....(1)}$$

Keterangan:
x = *data training*
y = *data testing*
D = Jarak

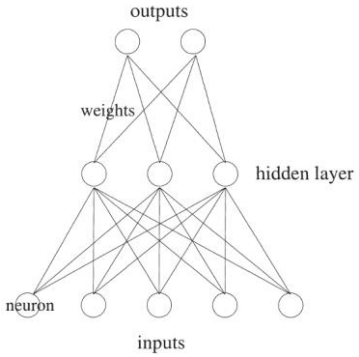
$$f(x') = \frac{1}{k} \sum_{x_1 \in N_k(x')} y_i \quad \text{.....(2)}$$

Keterangan:
x' = prediksi
k = jumlah tetangga terdekat
Nk(x') = tetangga terdekat
y_i = output tetangga terdekat

5. Neural Network

Neural Network adalah bagian dari pembelajaran mesin dan merupakan inti dari algoritma pembelajaran yang mendalam. Terinspirasi oleh fungsi canggih otak manusia tentang bagaimana ratusan miliar neuron yang saling memberi sinyal satu sama lain dan memproses informasi secara paralel (Khan, 2018)Algoritma Neural Network dapat digunakan untuk berbagai macam analisis data, meliputi: prediksi/regresi, pengenalan pola, *clustering*/pengelompokan dan optimasi (Kumar & Haynes, 2003)

Neural Network terdiri atas lapisan input neuron (*node*), beberapa lapisan neuron yang tersembunyi (*hidden layer*) dan lapisan terakhir sebagai neuron output. Antar neuron saling terhubung membentuk koneksi yang diasosiasikan dalam nomor numerik yang disebut sebagai weight. Arsitektur neural network digambarkan seperti pada gambar 1.



Sumber: (Khan, 2018)

Gambar 1. Arsitektur Neural Networks

Input dari neuron dilakukan operasi pembobotan (*weight*), menjumlahkannya (*weighted sum*) dan menambahkan bias. Hasil operasi tersebut dijadikan parameter dari *activation function* yang akan dijadikan output neuron. *Activation function* menentukan suatu neuron harus aktif berdasarkan *weighted sum* dari input (Khan, 2018).

Berikut merupakan rumus perhitungan Neural Networks

$$h_i = \sigma \sum_{j=1}^{N} V_{ij} x_j + T_i^{hid} \quad \text{.....(3)}$$

Keterangan:
h_i = output
σ = *activation fuction*
N = jumlah input neuron
V_{ij} = *weight*
X_j = input neuron
T_i^{hid} = bias *hidden neuron*

6. Root mean square error

Root mean square error (RMSE) adalah ukuran yang umum digunakan untuk menghitung perbedaan antara nilai regresi yang dihasilkan oleh model dan nilai sebenarnya (Hyndman & Koehler, 2006). RMSE merupakan tingkat kesalahan hasil regresi, artinya semakin

kecil nilai RMSE (mendekati 0), maka hasil regresi akan semakin akurat.Nilai RMSE dapat dihitung dengan persamaan berikut.

$$RMSE=\sqrt{\frac{\sum(X-Y)^2}{n}} \quad \text{.....(4)}$$

Keterangan:
X = nilai sebenarnya
Y = nilai prediksi
n = jumlah data

Normalisasi RMSE (NRMSE) memfasilitasi perbandingan nilai regresi terhadap dataset dengan skala yang berbeda. Penghitungan NRMSE adalah dengan membangi RMSE menggunakan rata-rata atau rentang data yang diukur (nilai maximum dikurangi nilai minumum. NRMSE dinyatakan dalam persentase, nilai yang lebih rendah menunjukkan *varians residual* yang lebih sedikit. Rumus NRMSE seperti berikut.

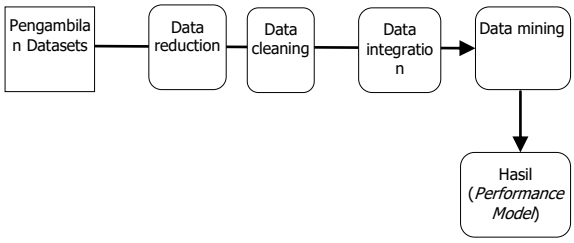
$$NRMSE = \frac{RMSE}{max_i Y_i - min_i Y_i} \quad or \quad NRMSE = \frac{RMSE}{\bar{y}} \quad \text{....(5)}$$

7. Covid-19

Covid-19 adalah penyakit yang disebabkan oleh jenis baru coronavirus yang disebut SARS-CoV-2. Pandemi Covid-19 telah mempengaruhi hampir seluruh dunia dan menyebabkan kematian lebih dari 315.131 pasien (Albahri et al., 2020). Biasanya pasien Covid-19 mengalami gejala berupa demam, batuk kering, dan kelelahan. Sekitar 80 persen pasien yang mengalami gejala sembuh tanpa perlu perawatan rumah sakit. Sementara 15 persen menderita sakit parah dan membutuhkan bantuan tabung oksigen dan 5 persen mengalami kritis serta membutuhkan perawatan intensif. Komplikasi penyakit yang dapat menyebabkan kematian antara lain gagal napas, sindrom gangguan pernapasan akut, sepsis dan syok septik, tromboemboli, dan/atau gagal organ multipel (termasuk kerusakan jantung, hati, atau ginjal).

II. METODOLOGI PENELITIAN

Alur penelitian ditampilkan seperti gambar 3 berikut ini .



Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 2. Alur Penelitian

1. Pengambilan Datasets

Pengambilan data dilakukan untuk memenuhi kebutuhan informasi data yang digunakan dalam penelitian. Data yang penulis terima dari Dinas Kominfo DIY adalah data dari bulan Maret sampai dengan September 2020 yang terdiri dari dua buah datasets, Dataset PE (Penyelidikan Epidemiologi) Covid Positif DIY dan Dataset Laporan Hasil Pemeriksaan Lab.

2. Data Reduction

Dari dua datasets yang diperoleh dilakukan seleksi attribute (kolom) sehingga hanya data yang berpengaruh dalam analisis yang akan digunakan.

3. Data Cleaning

Dilakukan untuk menghasilkan dataset yang bersih sehingga siap untuk dapat digunakan padad tahap data mining. Dataset dibersihkan untuk memperolah data berisi nilai-nilai yang relevan dan tidak terdapat *missing value* serta data *redudant*.

4. Data Integration

Tahap ini dilakukan untuk mendapatkan sebuah dataset yang siap dilakukan analisis data mining. Dataset tersebut diperloleh dari penggabungan Dataset PE (Penyelidikan Epidemiologi) Covid Positif DIY dan Dataset Laporan Hasil Pemeriksaan Lab.

5. Data mining

Analisis data mining dilakukan dengan menggunakan aplikasi Rapidminer. Dalam tahap ini dilakukan data training dan data testing untuk memperoleh model. Algoritma *regression* yang digunakan adalah Random Forest, k-NN dan Deep Learning.

6. Hasil

Berdasarkan hasil yang diperoleh dari masing-masing algoritma dalam proses data mining, performance model dari ketiga algoritma dibandingkan. Kemudian dicari algoritma dengan nilai *performance model* terbaik dari tiga algoritma tersebut dan dilakukan optimalisasi agar memperoleh hasil yang lebih baik.

III. HASIL DAN PEMBAHASAN

1. Pemrosesan Data Penelitian

Pada tahap *Pre-Processing* terdapat beberapa langkah yang dilakukan untuk mempersiapkan data mentah menjadi data yang siap digunakan untuk proses *data mining* (Pratama & Prasetyaningrum, 2021).

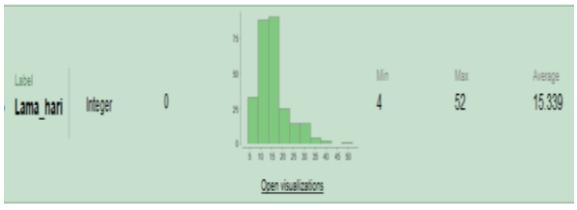
Setelah dilakukan *preprocessing* (*data reduction, data cleaning* dan *data integration*) diperoleh dataset yang berisi 31 atribut data sejumlah 271 *record data*. Format dataset seperti pada Tabel 1.

Tabel 1. Format data penelitian

No	Nama kolom	Tipe	Keterangan
1.	Lama_hari	numeric	dalam satuan hari
2.	jenis kelamin	binominal	L/P
3.	Usia	numeric	dalam satuan tahun
4.	kecamatan	polinominal	
5.	petugas kesehatan	numeric	1= ya; 0 = tidak
6.	demam	numeric	1= ya; 0 = tidak
7.	batuk	numeric	1= ya; 0 = tidak
8.	pilek	numeric	1= ya; 0 = tidak
9.	sakit tenggorokan	numeric	1= ya; 0 = tidak
10.	sesak napas	numeric	1= ya; 0 = tidak
11.	menggigil	numeric	1= ya; 0 = tidak
12.	sakit kepala	numeric	1= ya; 0 = tidak
13.	lemah	numeric	1= ya; 0 = tidak
14.	sakit otot	numeric	1= ya; 0 = tidak
15.	mual/ muntah	numeric	1= ya; 0 = tidak
16.	sakit perut	numeric	1= ya; 0 = tidak
17.	diare	numeric	1= ya; 0 = tidak
18.	hamil	numeric	1= ya; 0 = tidak
19.	diabetes	numeric	1= ya; 0 = tidak
20.	jantung	numeric	1= ya; 0 = tidak
21.	hipertensi	numeric	1= ya; 0 = tidak
22.	keganasan	numeric	1= ya; 0 = tidak
23.	gangguan imunologi	numeric	1= ya; 0 = tidak
24.	ginjal	numeric	1= ya; 0 = tidak
25.	hati	numeric	1= ya; 0 = tidak
26.	PPOK	numeric	1= ya; 0 = tidak
27.	asma	numeric	1= ya; 0 = tidak
28.	TBC	numeric	1= ya; 0 = tidak
29.	penyakit pernafasan lainnya	numeric	1= ya; 0 = tidak
30.	pneumonia	numeric	1= ya; 0 = tidak
31.	diagnosis lain	numeric	1= ya; 0 = tidak

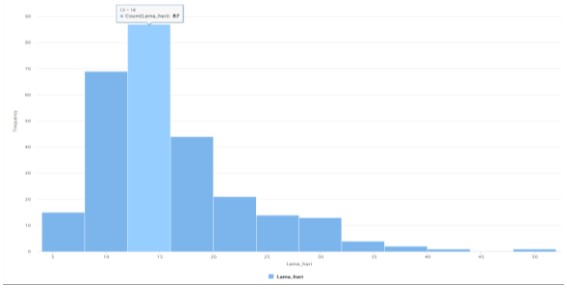
Sumber: (Supoyo & Prasetyaningrum, 2022)

Dataset diimport ke aplikasi Rapidminer untuk kemudian diolah menggunakan algoritma *data mining*. Setelah diimport dapat ditampilkan statistik data penelitian. Gambar menunjukan bahwa lama perawatan pasien covid-19 berkisar antara 4 sampai 52 hari dengan rata rata 15,339.



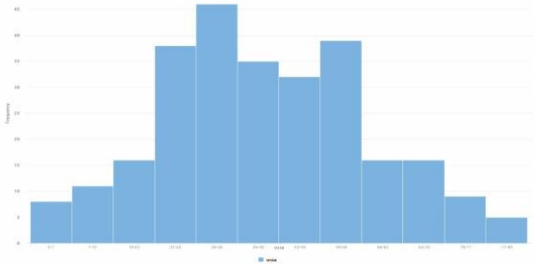
Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 3. Statistik Atribut Lama_Hari

Gambar menampilkan sebaran data bahwa kebanyakan pasien dirawat selama 12-16 hari.



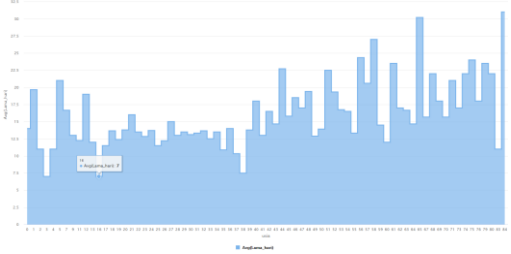
Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 4. Sebaran Lama Perawatan Pasien

Usia pasien didominasi oleh usia produktif dengan rentang nilai 21 sampai dengan 56 tahun seperti yang ditampilkan pada gambar 5.



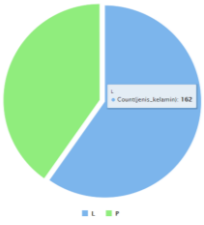
Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 5. Sebaran data usia pasien

Tampak bahwa terdapat korelasi antara usia pasien dan lama perawatan, pasien berusia 16 sampai 39 tahun membutuhkan rata-rata perawatan lebih sedikit dibanding usia lebih tua. Sedangkan grafik pasien usia 0 sampai 15 tahun lama perawatannya naik turun seperi ditampilkan pada gambar 6.



Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 6. Korelasi Usia Terhadap Lama Perawatan

Berdasarkan data yang diperoleh, jumlah pasien laki-laki lebih banyak dengan prosentase sekitar 60% berbanding dengan pasien perempuan 40%. Perbanding jenis kelamin seperti pada gambar 7.

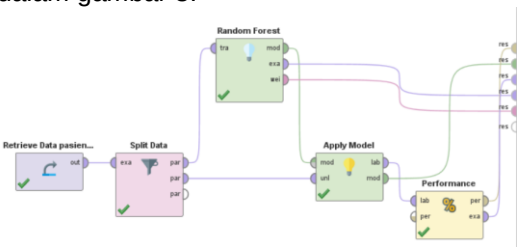


Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 7. Perbandingan jenis kelamin pasien

Untuk dapat lakukan pemrosesan dalam tahap data mining diperlukan pembagian dataset menjadi dua bagian yaitu data trining sebanyak 80% dan data testing 20%.

2. Implementasi Algoritma Random Forest

Implementasi algoritma Random Forest menggunakan aplikasi Rapidminer ditampilkan dalam gambar 8.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 8. Implementasi Algoritma Random Forest

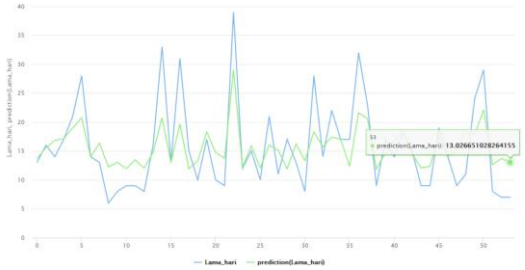
Parameter yang perlu diatur dalam algoritma Random forest adalah *number of tree* dan *maximal dept*. *Number of tree* yang diujikan adalah sebesar 100, 200, 300 dan 400, serta dikombinasikan dengan *maximal of dept* sebesar 8, 10 dan 12. *Performance modeling* ditampilkan dalam tabel berikut.

Tabel 2. Performance Algoritma Random Forest

No	number of tree	maximal depth	RMSE
1.	100	8	5.068
2.	100	10	5.060
3.	100	12	5.060
4.	200	8	5.016
5.	200	10	4.996
6.	200	12	4.996
7.	300	8	4.983
8.	300	10	4.951
9.	**300**	**12**	**4.949**
10.	400	8	4.995
11.	400	10	4.966
12.	400	12	4.965

Sumber: (Supoyo & Prasetyaningrum, 2022)

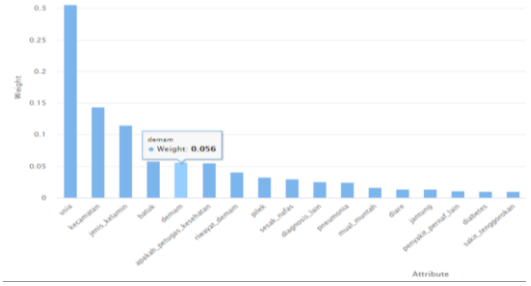
Berdasarkan analisis diperoleh performance terbaik dengan RMSE sebesar 4,949. Grafik perbandingan prediksi lama perawatan terhadap lama perawatan sebenarnya ditampilkan dalam gambar 9.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 9. Grafik Prediksi Lama Perawatan Algoritma Random Forest

Dalam analisis diperoleh pula pembobotan atribut yang berpengaruh terhadap label. Usia, kecamatan dan jenis kelamin merupakan atribut yang paling berpengaruh terhadap lama perawatan. Pembobotan atribut seperti pada gambar 10.

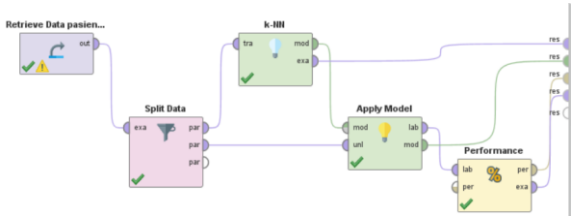


Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 10. Pembobotan Atribut Algoritma Random Forest

3. Implementasi Algoritma K-NN

Pengaturan analisis seperti pada gambar 11.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 11. Implementasi Algoritma k-NN

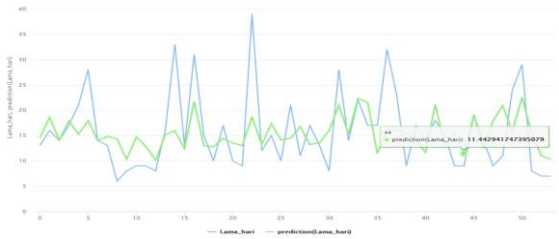
Parameter yang perlu disesuaikan dalam analisis ini adalah nilai k. Tidak ada ketentuan tentang nilai k ideal yang sebaiknya digunakan. Untuk mendapatkan akurasi terbaik maka dalam analisis ini digunakan k = 3, 5, 7, 9 dan 11. Diperoleh hasil analisis seperti tabel berikut.

Tabel 3. Performance analisis Algoritma k-NN

No	k	RMSE
1.	3	7.191
2.	5	6.473
3.	7	6.377
4.	9	6.359
5.	11	6.769

Sumber: (Supoyo & Prasetyaningrum, 2022)

Performance terbaik yang dihasilkan mempunyai nilai RMSE 6,359. Grafik perbandingan prediksi lama perawatan terhadap lama perawatan sebenarnya ditampilkan pada gambar 12.

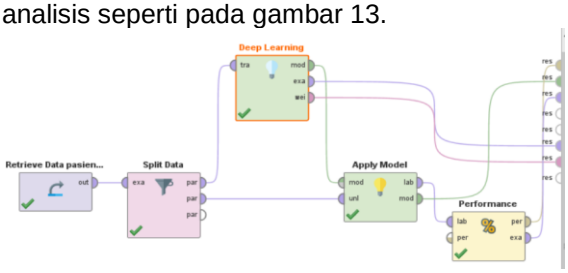


Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 12. Grafik Prediksi Lama Perawatan Algoritma k-NN

4. Implementasi Algoritma Deep Learning

Parameter yang perlu diatur dalam analisis Algoritma Deep Learning adalah *Activation Fucntion (activation)* dan *hidden layer size.* Dalam analisis ini menggunakan jenis jenis *activation* yaitu *Tanh* dan *Reclifier*. Digunakan *hidden layer size* default yaitu dua buah hidden layer dengan ukuran 50 dan 50. Pengaturan analisis seperti pada gambar 13.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 13. Implementasi Algoritma Deep Learning

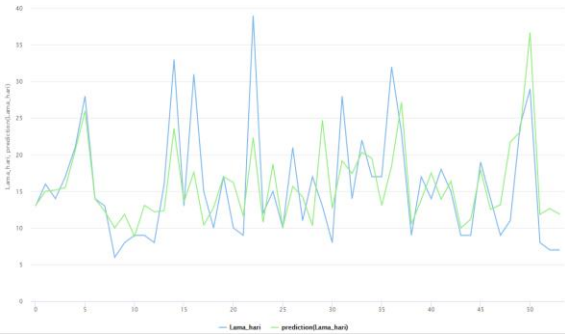
Dari implementasi algoritma didapatkan hasil *performance* yang tidak konsisten. Dengan menggunakan pengaturan parameter yang sama, dihasilkan performance RMSE yang berbeda-beda setiap kali *running.* Performance hasil analisis ditampikan pada tabel 4.

Tabel 4. Performance Analisis Algoritma Deep Learning

Activation	Running ke	RMSE
Tanh	1	5.928
	2	5.436
	3	5.753
	4	6.054
Reclifier	1	6.278
	2	6.315
	3	6.126
	4	6.422

Sumber: (Supoyo & Prasetyaningrum, 2022)

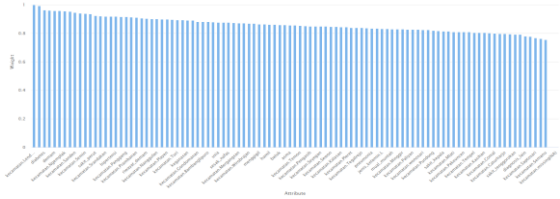
Performance terbaik yang dihasilkan mempunyai nilai RMSE 5,436. Grafik perbandingan prediksi lama perawatan terhadap lama perawatan sebenarnya ditampilkan pada gambar 14.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 14. Grafik Prediksi Lama Perawatan Algoritma Deep Learning

Diperoleh pula pembobotan atribut yang berpengaruh terhadap label. Pembobotan atribut seperti pada gambar 15.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 15. Pembobotan Atribut Algoritma Deep Learning

5. Perbandingan Hasil Analisis

Dari hasil analisis tiga algoritma yang digunakan diperoleh performance yang nilainya berbeda-beda, algoritma Random Forest mempunyai nilai performance terbaik dengan nilai 4,949. Perbandingan Performance hasil analisis algoritma ditampilkan pada tabel 5.

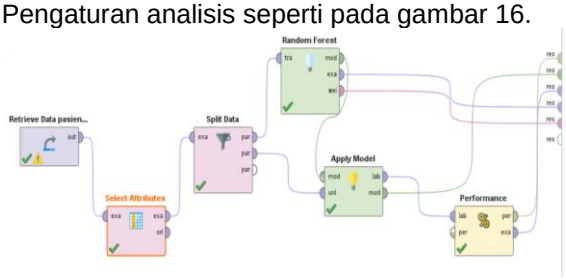
Tabel 5. Perbandingan Performance Algoritma

No	Algoritma	RMSE
1.	Random Forest	4.949
2.	k-NN	6.359
3.	Deep Learning Neural Network	5.436

Sumber: (Supoyo & Prasetyaningrum, 2022)

6. Optimalisasi Analisis

Pada algoritma Random forest yang mempunyai performance terbaik dilakukan optimalisasi dengan cara reduksi atribut. Operator *select atribut* digunakan dalam pengaturan pada aplikasi Rapidminer. Pengaturan analisis seperti pada gambar 16.



Sumber: (Supoyo & Prasetyaningrum, 2022)

Gambar 16. Implementasi Optimalisasi Algoritma Random Forest

Atribut yang nilai pembobotannya rendah dikurangi agar diperoleh nilai performance yang lebih baik. Dalam optimalisasi ini reduksi atribut yang diujicobakan sejumlah 4, 6, 8, 10, 12 dan 14. Hasil analisis ditampilkan pada tabel 6.

Tabel 6. Performance Analisis Optimalisasi Algoritma Random Forest

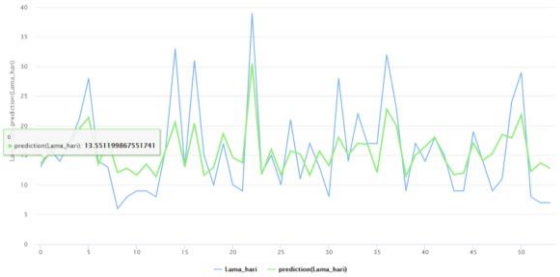
No	Reduksi Atribut	Jumlah Atribut yang digunakan	RMSE
1.	4	27	4.939
2.	6	25	4.873
3.	8	23	4.817
4.	10	21	4.830
5.	12	19	4.905
6.	14	17	4.953

Sumber: (Supoyo & Prasetyaningrum, 2022)

Dihasilkan performance model dengan nilai RMSE terbaik sebesar 4,817 dalam analisis yang menggunakan 23 atribut. Berdasarkan nilai RMSE diperoleh nilai NRMSE sebagai berikut.

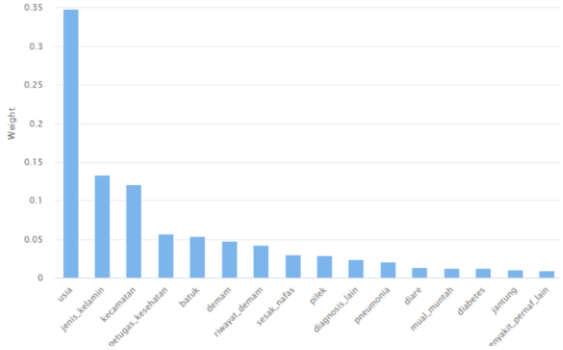
NRMSE = RMSE / mean x 100%
NRMSE = 4,817 / 15,339 x 100%
NRMSE = 31,40%

Grafik perbandingan prediksi lama perawatan terhadap lama perawatan sebenarnya ditampilkan pada gambar 17.



Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 17. Grafik Prediksi Lama Perawatan Optimalisasi Algoritma Random Forest

Setelah dilakukan optimalisasi, terdapat perubahan urutan atribut yang paling berpengaruh terhadap label jika dibandingkan dengan pembobotan sebelum dilakukan optimalisasi. Usia, jenis kelamin dan kecamatan merupakan atribut yang paling berpengaruh pada pembobotan setelah optimalisasi sedangkan sebelumnya tiga atribut teratasa adalah usia, kecamatan dan jenis kelamin. Grafik perbandingan lama perawatan terhadap lama perawatan sebenarnya ditampilkan pada gambar berikut.



Sumber: (Supoyo & Prasetyaningrum, 2022)
Gambar 18. Pembobotan Atribut Optimalisasi Algoritma Random Forest

IV. KESIMPULAN

Dari hasil penelitian dapat disimpulkan bahwa dari tiga algoritma yang digunakan, Random Forest memiliki performance model terbaik sebesar 4,817. Namun *performance model* tersebut belum cukup baik untuk digunakan dalam memprediksi lama perawatan pasien Covid-19 di DIY karena jika dibandingkan dengan rata-rata lama perawatan sebesar 15.339 hari maka diperoleh nilai NRMSE yang masih terlalu besar yakni 31,40%. Berdasarkan pengujian, seleksi atribut yang digunakan mempengaruhi *performance* hasil prediksi. Lima atribut/faktor yang paling berpengaruh terhadap lama perawatan pasien adalah usia, jenis kelamin, kecamatan, apakah petugas kesehatan dan batuk. Diperlukan penelitian lanjutan untuk meningkat hasil prediksi dengan menggunakan lebih banyak *record data* serta tambahan atribute lain yang berpengaruh terhadap lama perawatan, seperti RS perawatan dan tindakan medis.

V. REFERENSI

Albahri, A. S., Hamid, R. A., Alwan, J. k., Al-qays, Z. T., Zaidan, A. A., Zaidan, B. B., Albahri, A. O. S., AlAmoodi, A. H., Khlaf, J. M., Almahdi, E. M., Thabet, E., Hadi, S. M., Mohammed, K. I., Alsalem, M. A., Al-Obaidi, J. R., & Madhloom, H. T. (2020). Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. *Journal of Medical Systems*, *44*(7). https://doi.org/10.1007/s10916-020-01582-x

Gading Sadewo, M., Perdana Windarto, A., Hartama, D., (2017). Penerapan Datamining pada Populasi Daging Ayam Ras Pedaging di Indonesia Berdasarkan Provinsi Menggunkanan K-Means Clustering. *InfoTekJar : Jurnal Nasional Informatika Dan Teknologi Jaringan*, *2*(1), 60–67.

Haspriyanti, A. U., & Prasetyaningrum, P. T. (2020). Penerapan Data Mining Untuk Prediksi Layanan Produk Indihome Menggunakan Metode K-Nearst Neighbor The Data Mining Application for IndiHome Product Service Prediction by Using K-Nearest Neighbor Method. *JISAI*.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/J.IJFORECAST.2006.03.001

Khan, G. M. (2018). Artificial neural network (ANNs). *Studies in Computational Intelligence*, *725*, 39–55. https://doi.org/10.1007/978-3-319-67466-7_4

Kramer, O. (2011). *Unsupervised K-Nearest Neighbor Regression*.

Kumar, K. ;, & Haynes, J. D. (2003). Forecasting credit ratings Using ANN and statistical techniques. *General Rights INTERNATIONAL JOURNAL OF BUSINESS STUDIES*, *11*(1), 91–108.

Muhammad, L. J., Islam, M. M., Usman, S. S., & Ayon, S. I. (2020). Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery. *SN Computer Science*, *1*(4), 1–7. https://doi.org/10.1007/s42979-020-00216-w

Prasetyaningrum, P. T., Mercu, U., Yogyakarta, B., Pratama, I., Mercu, U., Yogyakarta, B., Chandra, A. Y., Mercu, U., & Yogyakarta, B. (2021). Implementation Of Machine Learning To Determine The Best Employees Using Random Forest Method. *International Journal of Computer, Network Security and Information System (IJCONSIST)*, *2*(March), 53–59.

Pratama, I., & Prasetyaningrum, P. T. (2021). Pemetaan Profil Mahasiswa Untuk Peningkatan Strategi Promosi Perguruan Tinggi Menggunakan Predictive Apriori. *Jurnal Eksplora Informatika*, *10*(2), 159–166. https://doi.org/10.30864/eksplora.v10i2.505

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Supoyo, A., & Prasetyaningrum, P. T. (2022). *Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY*.

Virdaus, D., & Prasetyaningrum, P. T. (2020). Penerapan Data Mining Untuk Memprediksi Harga Bawang Merah Di Yogyakarta Menggunakan Metode K-Nearest Neighbor. *Journal Of …*, *84*, 1–8.

Whidhiasih, R. N., Wahanani, N. A., & Supriyanto. (2013). *Klasifikasi Buah Belimbing Berdasarkan Citra Red-Green-Blue Menggunakan KNN Dan LDA*. Jurnal Penelitian Ilmu Komputer, System Embedded& Logic.