Jurnal Teknologi Sistem Informasi dan Aplikasi **ISSN: 2654-3788**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2654-4229**
Vol. 7, No. 4, Oktober 2024 (1548-1556) **DOI: 10.32493/jtsi.v7i4.40755**

Penerapan Seleksi Fitur pada Deteksi Coronavirus Disease 19 (COVID-19) berbasis Random Forest

Aries Saifudin[1], Endar Nirmala[2], Irpan Kusyadi[3]

[1,2]Teknik Informatika, Universitas Pamulang, Jl. Raya Puspitek No. 46 Buaran, Serpong, Tangerang Selatan, Banten, Indonesia, 15417
[3]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Terbuka, Jl. Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan, Banten, Indonesia, 15418
e-mail: [1]aries.saifudin@unpam.ac.id, [2]dosen00216@unpam.ac.id, [3]irpan.kusyadi@ecampus.ut.ac.id

Submitted Date: October 02nd, 2024 Reviewed Date: October 28th, 2024
Revised Date: October 31st, 2024 Accepted Date: October 31st, 2024

Abstract

Data mining using machine learning algorithms can be used to help analyze historical data to predict COVID-19. The dataset used for predicting COVID-19 has many features, but those features have the possibility of redundancy or irrelevance that can cause a decrease in classifier performance. This research proposes a model that implements feature selection to select relevant features and can provide improved performance predictions for diagnose COVID-19. Some proposed feature selection techniques are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), Sequential Forward Floating Selection (SBFS), Sequential Backward Floating Selection (SBFS), and selectKBest. The classification algorithm used to classify is Random Forest. The model that gives the best performance value is the model that applies the SFS dan SFFS as feature selection.

Keywords: Feature Selection; Prediction; COVID-19

Abstrak

Data Mining menggunakan algoritma pembelajaran mesin (machine learning) dapat digunakan untuk membantu menganalisis data historis untuk memprediksi COVID-19. Dataset yang digunakan untuk memprediksi COVID-19 memiliki banyak fitur, namun fitur tersebut memiliki kemungkinan redundansi atau tidak relevan yang dapat menyebabkan penurunan kinerja pengklasifikasi. Penelitian ini mengusulkan model yang menerapkan pemilihan fitur (feature selection) untuk memilih fitur yang relevan dan dapat memberikan prediksi kinerja yang lebih baik untuk diagnosa/prediksi COVID-19. Beberapa teknik pemilihan fitur yang diusulkan adalah Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), Sequential Forward Floating Selection (SBFS), Sequential Backward Floating Selection (SBFS), dan selectKBest. Algoritma klasifikasi yang digunakan untuk mengklasifikasikan adalah Random Forest. Model yang memberikan nilai kinerja terbaik adalah model yang menerapkan SFS dan SFFS sebagai seleksi fitur.

Kata kunci: Seleksi Fitur; Prediksi; COVID-19

1 Introduction

Coronavirus disease 2019 (COVID-19) adalah penyakit menular yang disebabkan oleh virus SARS-CoV-2. Kebanyakan orang yang jatuh sakit COVID-19 akan mengalami gejala ringan hingga sedang dan sembuh tanpa pengobatan khusus. Namun, beberapa akan menjadi sakit parah dan memerlukan perhatian medis. Pada tahun 2019, virus corona baru diidentifikasi sebagai penyebab wabah penyakit yang berasal dari China. Virus ini merupakan patogen zoonotik yang memiliki tingkat mutasi tinggi, dan dapat menetap pada manusia dan binatang dengan presentasi klinis beragam, mulai dari asimptomatik, gejala ringan

sampai berat, sampai kematian (Cascella et al., 2022)(Sahin et al., 2020)(Guo et al., 2020).

Coronavirus disease 2019 (COVID-19) merupakan penyakit infeksi saluran pernapasan yang disebabkan oleh severe acute respiratory syndrome virus corona 2 (SARS-CoV-2), atau sering disebut virus Corona. Coronavirus adalah keluarga virus yang dapat menyebabkan penyakit seperti flu biasa, sindrom pernapasan akut parah (SARS) dan sindrom pernapasan Timur Tengah (MERS).

Virus dapat menyebar dari mulut atau hidung orang yang terinfeksi dalam partikel cairan kecil ketika mereka batuk, bersin, berbicara, bernyanyi, atau bernapas. Partikel-partikel ini berkisar dari tetesan pernapasan yang lebih besar hingga aerosol yang lebih kecil. Anda dapat terinfeksi dengan menghirup virus jika Anda berada di dekat seseorang yang memiliki COVID-19, atau dengan menyentuh permukaan yang terkontaminasi dan kemudian mata, hidung, atau mulut Anda. Virus menyebar lebih mudah di dalam ruangan dan di tempat ramai.

Virus ini dikenal sebagai sindrom pernafasan akut yang parah coronavirus 2 (SARS-CoV-2). Penyakit yang ditimbulkannya disebut coronavirus disease 2019 (COVID-19). Pada Maret 2020, Organisasi Kesehatan Dunia (WHO) menyatakan wabah COVID-19 sebagai pandemi.

Hasil pemeriksaan laboratorium pada pasien COVID-19 tidak spesifik, tetapi sering ditemukan limfopenia, peningkatan laktat dehidrogenase, dan peningkatan aminotransferase. Sedangkan pemeriksaan pencitraan toraks dapat menunjukkan gambaran pneumonia (Cascella et al., 2022)(World Health Organization, 2020).

Karakteristik gambaran COVID-19 pada CT scan toraks nonkontras adalah ground glass opacification (GGO) bilateral, multilobar, dengan distribusi periferal atau posterior. Walaupun kurang spesifik, ultrasonography (USG) dan rontgen toraks juga dapat membantu menegakkan diagnosis COVID-19.Diagnosis baku emas COVID-19 adalah mendeteksi virus RNA dengan pemeriksaan nucleic acid amplification test (NAAT) dengan metode real time polymerase chain reaction (RT-PCR) (Cascella et al., 2022)(World Health Organization, 2020).

Sampai saat ini, belum terdapat terapi spesifik dalam penanganan COVID-19. Terdapat dua studi terbesar tentang terapi COVID-19 yang hingga saat ini masih berjalan secara global. Studi menunjukkan bahwa antivirus favipiravir, remdesivir, dan tocilizumab mungkin memiliki beberapa manfaat untuk penanganan COVID-19, dan sudah diperbolehkan penggunaannya di Indonesia (Yanti et al., 2020)(Mendez, 2020)(Sindi et al., 2020)

Pasien COVID-19 tanpa gejala dan derajat ringan umumnya hanya disarankan isolasi di rumah dan menggunakan obat simptomatik. Pasien dengan gejala derajat sedang sampai berat membutuhkan terapi oksigen, sehingga disarankan untuk dirawat inap dan terkadang diperlukan tindakan intubasi dan ventilasi mekanik apabila terjadi gagal napas atau acute respiratory distress syndrome (Cennimo, 2024)(Burhan et al., 2020)

Adanya infeksi COVID-19 yang meluas telah mendorong upaya di seluruh dunia untuk mengendalikan dan mengelola virus, dan diharapkan dapat mengekangnya sepenuhnya. Salah satu penelitian penting adalah penggunaan machine learning (ML) untuk memahami dan melawan COVID-19. Ini saat ini merupakan bidang penelitian aktif. Meskipun sudah ada banyak survei dalam literatur, ada kebutuhan untuk mengikuti jumlah publikasi yang berkembang pesat tentang aplikasi ML terkait COVID-19. Makalah ini menyajikan ulasan laporan terbaru tentang algoritma ML yang digunakan dalam kaitannya dengan COVID-19. Kami fokus pada potensi ML untuk dua aplikasi utama: diagnosis COVID-19 dan prediksi risiko dan tingkat keparahan kematian, menggunakan data klinis dan laboratorium yang tersedia. Aspek yang terkait dengan tipe algoritme, kumpulan data pelatihan, dan pemilihan fitur dibahas. Saat kami meliput karya yang diterbitkan antara Januari 2020 dan Januari 2021, beberapa poin penting telah terungkap. Sebagian besar algoritme pembelajaran mesin yang digunakan dalam dua aplikasi ini adalah algoritme pembelajaran yang diawasi. Model yang sudah mapan belum digunakan dalam implementasi dunia nyata, dan sebagian besar penelitian terkait bersifat eksperimental. Fitur diagnostik dan prognostik yang ditemukan oleh model ML konsisten dengan hasil yang disajikan dalam literatur medis. Keterbatasan aplikasi yang ada adalah penggunaan set data yang tidak seimbang yang rentan terhadap bias seleksi.

Negara-negara di seluruh dunia telah terkena dampak virus, mengakibatkan berbagai tindakan

diberlakukan, termasuk penguncian negara, jam malam, dan pembatasan perjalanan. Meskipun gejala umum infeksi COVID-19 biasanya ringan, bagi beberapa pasien infeksi dapat menyebabkan komplikasi serius, dan terkadang mematikan. Mengelola jumlah kasus COVID-19 yang melonjak adalah tantangan besar yang membuat fasilitas perawatan kesehatan di seluruh dunia kewalahan; namun, masih ada informasi yang cukup tentang virus tersebut. Sejak munculnya infeksi COVID-19, para peneliti dari berbagai disiplin ilmu telah mengeksplorasi virus baru ini. Machine learning (ML) adalah cabang dari kecerdasan buatan (AI) yang berfokus pada produksi sistem yang mampu belajar dari contoh dan meningkatkan tanpa diprogram secara eksplisit(Burhan et al., 2020). Machine learning (ML) telah berhasil diterapkan di banyak bidang, termasuk perawatan kesehatan (Mendez, 2020) dan informatika medis (Oxford University, 2022). Satu arah penelitian penting memanfaatkan Machine learning (ML) untuk memahami dan melawan COVID-19. Banyak lini penelitian telah dimulai untuk penerapan dan pengembangan algoritme Machine learning (ML) terkait COVID-19.

Telah banyak metode diagnosa yang digunakan untuk mendeteksi seseorang terkena COVID-19 atau tidak. Namun, keakuratan tes dapat bervariasi tergantung pada saat sampel Anda diambil selama perjalanan penyakit Anda. Jika Anda dites terlalu cepat setelah terpapar COVID-19, mungkin tidak ada cukup virus di tubuh Anda untuk mendapatkan hasil yang akurat. Jika ini masalahnya pada saat tes, tes Anda mungkin kembali negatif, bahkan jika Anda benar-benar memiliki virus. Ini akan dianggap sebagai tes 'negatif palsu'. Penting untuk dipahami bahwa profesional perawatan kesehatan mempertimbangkan sejumlah faktor dalam membuat diagnosis COVID-19.

Dataset COVID-19 berisi banyak fitur/atribut, tetapi belum tentu semuanya relevan dan dapat meningkatkan kinerja model. Pada penelitian ini diusulkan penerapan pemilihan fitur menggunakan algoritma random forest agar diperoleh fitur yang relevan saja yang digunakan untuk melatih model. Dengan penerapan pemilihan fitur dapat mengurangi kompleksitas komputasi.

Pemilihan fitur selalu menjadi masalah besar dalam pembelajaran mesin. Pemilihan fitur merupakan bagian terpenting dari proyek ilmu data, karena ini membantu kami mengurangi dimensi kumpulan data dan menghapus variabel yang tidak berguna. Untungnya, ada beberapa model yang membantu kami menghitung pentingnya fitur, yang membantu kami mengabaikan yang kurang berguna. Random Forest adalah salah satu model tersebut.

Random Forest adalah salah satu algoritma pembelajaran mesin yang paling populer. Mereka sangat sukses karena mereka secara umum memberikan kinerja prediksi yang baik, overfitting rendah, dan interpretasi yang mudah. Penafsiran ini diberikan oleh fakta bahwa sangat mudah untuk menurunkan pentingnya setiap variabel pada keputusan pohon. Dengan kata lain, mudah untuk menghitung seberapa besar kontribusi setiap variabel terhadap keputusan. Pemilihan fitur menggunakan Random Forest berada di bawah kategori metode tertanam yang menggabungkan kualitas metode filter dan pembungkus. Mereka diimplementasikan oleh algoritma dengan metode pemilihan fitur bawaannya sendiri.

Pada penelitian ini menggunakan metode eksperimen dengan membuat aplikasi untuk menerapkan model usulan. Kemudian menguji model menggunakan dataset sekunder yang diunduh dari Kaggle dan mengukur kinerjanya.

2 Metodologi

Pada penelitian ini digunakan pendekatan kuantitatif. Pada pendekatan penelitian kuantitatif, dilakukan analisa kuantitatif secara teliti terhadap beberapa generasi informasi yang berbentuk kuantitatif. Umumnya pendekatan kuantitatif memiliki tiga bentuk yang berbeda, yaitu pendekatan inferensial, pendekatan eksperimental, dan pendekatan simulasi. Pada pendekatan inferensial, sampel yang diperoleh digunakan untuk membuat dugaan karakteristik populasi, relasinya, dan lain-lain. Pada pendekatan ini, peneliti tidak memiliki kontrol atas karakteristik, variabel, dan responden yang diteliti. Pendekatan eksperimental ditandai dengan adanya kontrol atas lingkungan penelitian oleh peneliti. Eksperimen adalah suatu proses yang sistematis di mana peneliti memiliki kontrol atas variabel berdasarkan pertimbangan agar sesuai dengan tujuan penelitian. Simulasi berarti operasi model numerik yang mewakili struktur proses dinamis. Pada pendekatan simulasi, lingkungan buatan dibuat di mana informasi yang diperlukan dapat dihasilkan.

Tujuan dari metode kuantitatif adalah untuk dapat memahami bagaimana sesuatu dikonstruksi, bagaimana dibangun, dan bagaimana cara kerjanya. Penelitian kuantitatif umumnya didorong oleh hipotesis, kemudian dibuat rumusan dan pengujian secara ketat untuk menunjukkan bahwa hipotesisnya salah. Sehingga usaha yang dilakukan adalah membuktikan bahwa hipotesis yang dibuat adalah salah, jika hipotesisnya tahan uji, maka hipotesis tersebut dianggap benar. Tetapi jika tidak tahan uji, maka hipotesisnya dianggap salah.

Sudut pandang kuantitatif menekankan bahwa pengukuran merupakan dasar yang dapat digunakan untuk menunjukkan hubungan antara observasi dan formalisasi model, teori, dan hipotesis. Penelitian dan metode kuantitatif akan menghasilkan pengembangan model, teori, dan hipotesis yang berkaitan dengan fenomena alam.

Pendakatan yang akan digunakan pada penelitian ini adalah pendekatan eksperimen. Pada penelitian eksperimen dilakukan dengan cara menginvestigasi hubungan sebab-akibat menggunakan pengujian yang dikontrol oleh peneliti. Pada penelitian semieksperimental sering mendapatkan kendala pada tidak cukupnya akses terhadap sampel, masalah etika, dan sebagainya. Untuk pengembangan, evaluasi, dan pemecahan masalah proyek biasanya dilakukan dengan eksperimen.

Sejak munculnya COVID-19, para peneliti dalam pembelajaran mesin dan radiologi telah bergegas mengembangkan algoritme yang dapat membantu diagnosis, triase, dan pengelolaan penyakit. Akibatnya, ribuan model diagnostik dan prognostik menggunakan radiografi dada dan CT telah dikembangkan. Namun, tanpa pendekatan standar untuk pengembangan atau evaluasi, sulit, bahkan bagi para ahli, untuk menentukan model mana yang paling bermanfaat secara klinis. Di sini, kami berbagi keprihatinan utama kami dan menyajikan beberapa solusi yang mungkin.

Pandemi ini terus menantang sistem medis di seluruh dunia dalam banyak aspek, termasuk peningkatan tajam dalam permintaan tempat tidur rumah sakit dan kekurangan peralatan medis, sementara banyak petugas kesehatan sendiri telah terinfeksi. Dengan demikian, kapasitas untuk keputusan klinis segera dan penggunaan sumber daya kesehatan yang efektif sangat penting. Tes diagnosis COVID-19 yang paling tervalidasi, menggunakan reverse transcriptase polymerase chainreaction (RT-PCR), telah lama kekurangan di negara berkembang. Hal ini berkontribusi pada peningkatan tingkat infeksi dan penundaan tindakan pencegahan kritis. Skrining yang efektif memungkinkan diagnosis COVID-19 yang cepat dan efisien dan dapat mengurangi beban pada sistem perawatan kesehatan. Model prediksi yang menggabungkan beberapa fitur untuk memperkirakan risiko infeksi telah dikembangkan, dengan harapan dapat membantu staf medis di seluruh dunia dalam melakukan triase pasien, terutama dalam konteks sumber daya perawatan kesehatan yang terbatas. Model-model ini menggunakan fitur-fitur seperti pemindaian tomografi komputer (CT), gejala klinis, tes laboratorium, dan integrasi fitur-fitur ini. Namun, sebagian besar model sebelumnya didasarkan pada data dari pasien yang dirawat di rumah sakit, sehingga tidak efektif dalam skrining SARS-CoV-2 dalam populasi umum.

Random Forest adalah model terawasi yang mengimplementasikan pohon keputusan dan metode bagging. Idenya adalah bahwa dataset pelatihan disampel ulang sesuai dengan prosedur yang disebut "bootstrap". Setiap sampel berisi subset acak dari kolom asli dan digunakan agar sesuai dengan pohon keputusan. Jumlah model dan jumlah kolom adalah hyperparameter yang akan dioptimalkan.

Akhirnya, prediksi pohon dicampur bersama-sama menghitung nilai rata-rata (untuk regresi) atau menggunakan pemungutan suara lunak (untuk klasifikasi). Ide bagging adalah bahwa, dengan merata-ratakan output dari pohon keputusan tunggal, kesalahan standar berkurang dan begitu juga varians model menurut tradeoff bias-varians. Itu sebabnya Random Forest menjadi sangat terkenal dalam beberapa tahun terakhir.

Setiap pohon di Random Forest dapat menghitung pentingnya fitur sesuai dengan kemampuannya untuk meningkatkan kemurnian daun. Ini adalah topik yang berkaitan dengan cara kerja Classification And Regression Trees (CART). Semakin tinggi peningkatan kemurnian daun, semakin tinggi pentingnya fitur tersebut. Ini dilakukan untuk setiap pohon, kemudian dirata-ratakan di antara semua pohon dan, akhirnya, dinormalisasi menjadi 1. Jadi, jumlah skor kepentingan yang dihitung dengan Random Forest adalah 1.

Setelah kami mengetahui pentingnya setiap fitur, kami melakukan pemilihan fitur menggunakan prosedur yang disebut Penghapusan Fitur Rekursif. Dalam artikel ini, saya akan berbicara tentang versi yang menggunakan validasi silang k-fold.

Idenya adalah untuk menyesuaikan model, kemudian menghapus fitur yang kurang relevan dan menghitung nilai rata-rata dari beberapa metrik kinerja di CV. Kemudian kami menghapus fitur penting kedua terakhir, menyesuaikan model lagi dan menghitung kinerja rata-rata. Kami terus melakukan pendekatan ini sampai tidak ada fitur yang tersisa. Kumpulan fitur yang memaksimalkan kinerja di CV adalah kumpulan fitur yang harus kami kerjakan. Harap dicatat bahwa seluruh prosedur harus bekerja dengan nilai yang sama untuk hyperparameter.

2.1 Data Pendukung

Dataset yang digunakan pada penelitian ini merupakan data sekunder yang diambil dari website Kaggle dengan link https://www.kaggle.com/code/midouazerty/symptoms-covid-19-using-7-machine-learning-98/data. Spesifikasi dataset yang digunakan ditunjukkan pada Tabel 1.

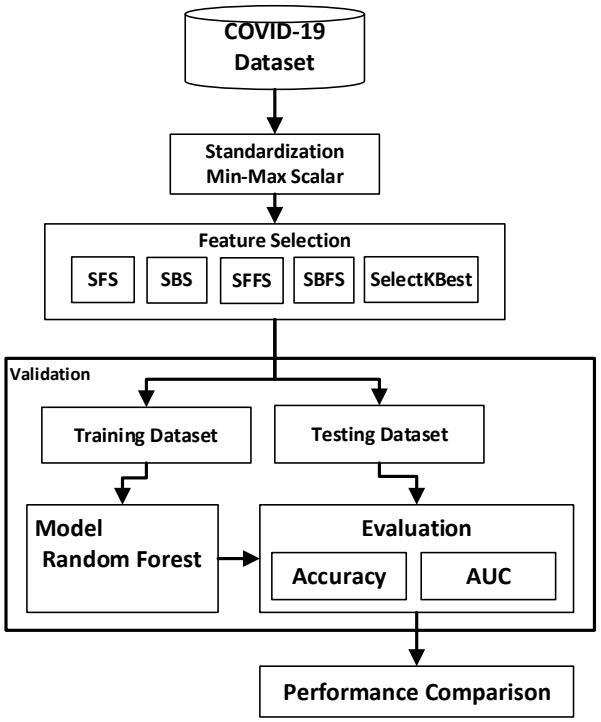
Tabel 1 Spesifikasi Fitur/Atribut Dataset

No.	Atribut	Nilai
1	Breathing Problem	Yes/No
2	Fever	Yes/No
3	Dry Cough	Yes/No
4	Sore throat	Yes/No
5	Running Nose	Yes/No
6	Asthma	Yes/No
7	Chronic Lung Disease	Yes/No
8	Headache	Yes/No
9	Heart Disease	Yes/No
10	Diabetes	Yes/No
11	Hyper Tension	Yes/No
12	Fatigue	Yes/No
13	Gastrointestinal	Yes/No
14	Abroad travel	Yes/No
15	Contact with COVID Patient	Yes/No
16	Attended Large Gathering	Yes/No
17	Visited Public Exposed Places	Yes/No
18	Family working in Public Exposed Places	Yes/No
19	Wearing Masks	No
20	Sanitization from Market	No
21	COVID-19	Yes/No

Berdasarkan Tabel 1, fitur/atribut Wearing Masks dan Sanitization from Market hanya berisi nilai No. Kedua fitur/atribut tersebut dianggap tidak berpengaruh, maka dikeluarkan dalam penerapan model yang diusulkan.

2.2 Model yang Diusulkan

Model yang diusulkan untuk prediksi COVID-19 dalam penelitian ini ditunjukkan pada Gambar 1. Yang pertama adalah melakukan standardisasi data COVID-19. Kemudian fitur dipilih menggunakan algoritma seleksi fitur (feature selection) yang diusulkan. Dataset hasil seleksi fitur tersebut kemudian digunakan untuk melatih dan menguji model menggunakan teknik 10-fold cross validation. Hasil pengujian dimasukkan ke dalam tabel confusion matrix untuk mengukur kinerja model. Pada model yang diusulkan menerapkan algoritma Random Forest.



Gambar 1 Model yang Diusulkan

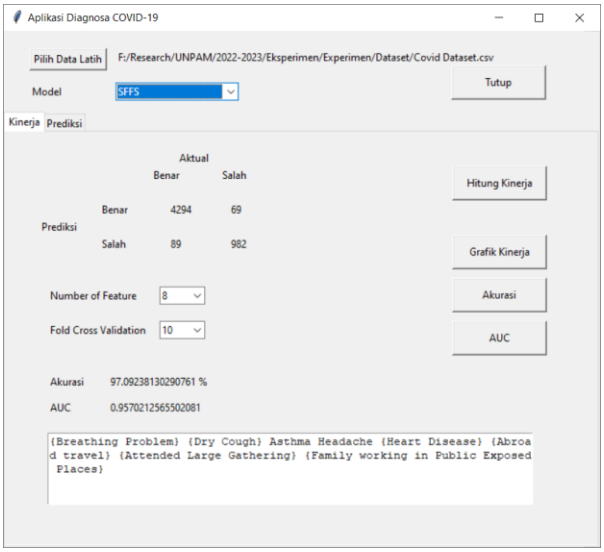
Berdasarkan model yang diusulkan pada Gambar 1, maka akan dibuat 6 model yaitu Random Forest, SFS, SBS, SFFS, SBFS dan SelectKBest. Kinerja kelima model tersebut dibandingkan untuk mendapatkan model terbaik. SFS adalah metode pemilihan fitur deterministik yang menggunakan pencarian hill climbing untuk menambah dan menilai semua kemungkinan

ekstensi atribut tunggal untuk subset saat ini. Sedangkan SBS bekerja berlawanan arah dengan SFS (Liu et al., 2016). SFS dan SBS memilih fitur secara satu arah, sehingga fitur yang telah dievaluasi tidak dapat dipilih lagi, tetapi kelemahan tersebut dihindari pada SFFS dan SBFS(Xue et al., 2016). SelectKBest adalah modul di pustaka scikit learn yang memilih k fitur yang memiliki skor tertinggi. Skor tersebut dihitung berdasarkan analisis statistik univariat(Nair & Bhagat, 2019), yaitu analisis variabel satu per satu.

Hasil pengujian tersebut dimasukkan ke dalam tabel confusion matrix dan dilakukan penghitungan performansi pengklasifikasi dalam bentuk akurasi dan AUC (Area Under the Curve). Confusion matrix adalah alat yang sangat berguna untuk menganalisis kinerja model klasifikasi dan mampu mengenali tupel dan fitur dari kelas yang berbeda (Jiawei et al., 2012). Analisis menggunakan confusion matrix dilakukan dengan menghitung jumlah objek yang diprediksi dengan benar dan tidak tepat untuk mengetahui performansi model (Gorunescu, 2011). Nilai validasi yang telah masuk ke dalam confusion matrix digunakan untuk menghitung nilai Accuracy atau AUC masing-masing model dan mengukur performansi model. Dalam penelitian ini menggunakan bahasa pemrograman python yang telah menyediakan banyak library yang dapat digunakan, di antaranya untuk mengukur performansi model dalam bentuk confusion matrix, akurasi, dan AUC.

3 Hasil dan Pembahasan

Untuk mengimplementasikan model yang diusulkan maka dikembangkan aplikasi menggunakan bahasa pemrograman Python. Tampilan aplikasi Diagnosa COVID-19 yang dikembangkan ditunjukkan pada Gambar 2.



Gambar 2 Tampilan Aplikasi Diagnosa COVID-19

Berdasarkan nilai kinerja model pada Tabel 2 dan Tabel 3 didapatkan model dengan kinerja tertinggi saat menggunakan salah satu fitur, yaitu Heart Problem. Nilai kinerja tertinggi adalah 97,09% untuk akurasi dan 0,9570 untuk nilai AUC. Nilai kinerja model hasil eksperimen kemudian divisualisasikan menggunakan grafik yang ditunjukkan pada Gambar 3 dan Gambar 4.

Tabel 2 Kinerja Model (Akurasi)

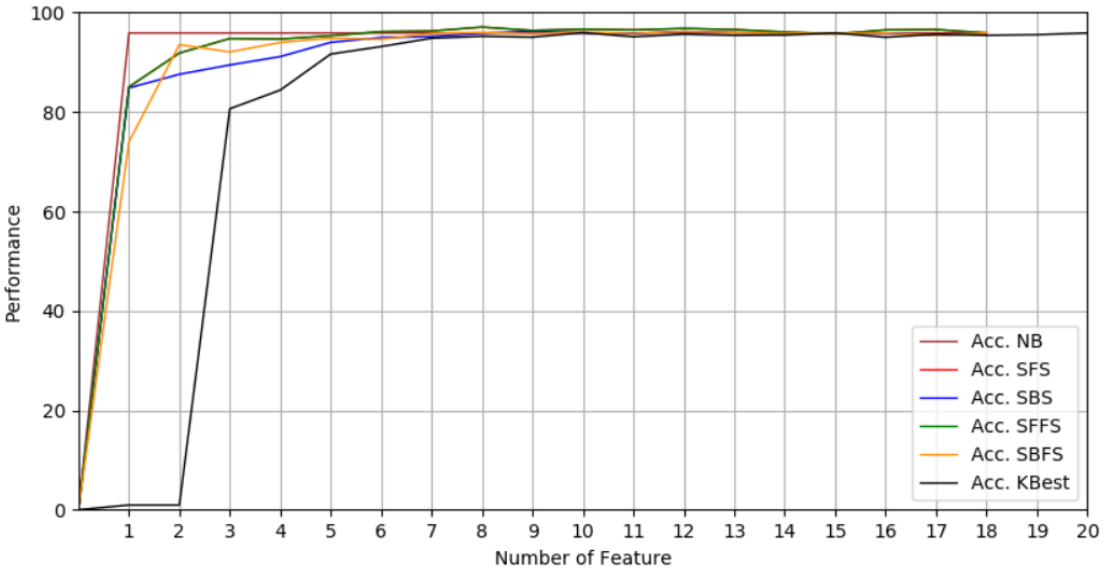
Number of Feature	RF	SFS	SBS	SFFS	SBFS	KBest
1	95,90	85,06	84,85	85,06	74,00	80,66
2	95,90	91,87	87,60	91,87	93,56	84,41
3	95,90	94,79	89,47	94,79	92,09	91,66
4	95,90	94,68	91,15	94,68	94,00	93,17
5	95,90	95,34	94,02	95,34	94,87	94,85
6	95,90	96,17	94,98	96,17	94,64	95,23
7	95,90	96,30	95,18	96,30	95,80	95,07
8	95,90	97,09	95,84	97,09	95,99	95,97
9	95,90	96,41	96,23	96,41	95,77	95,14
10	95,90	96,67	95,91	96,67	96,03	95,66
11	95,90	96,52	95,75	96,52	95,75	95,42
12	95,90	96,82	96,21	96,82	96,21	95,49
13	95,90	96,60	95,95	96,60	95,95	95,90
14	95,90	96,10	95,82	96,10	95,82	95,05
15	95,90	95,68	95,79	95,68	95,79	95,64

ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v7i4.40755

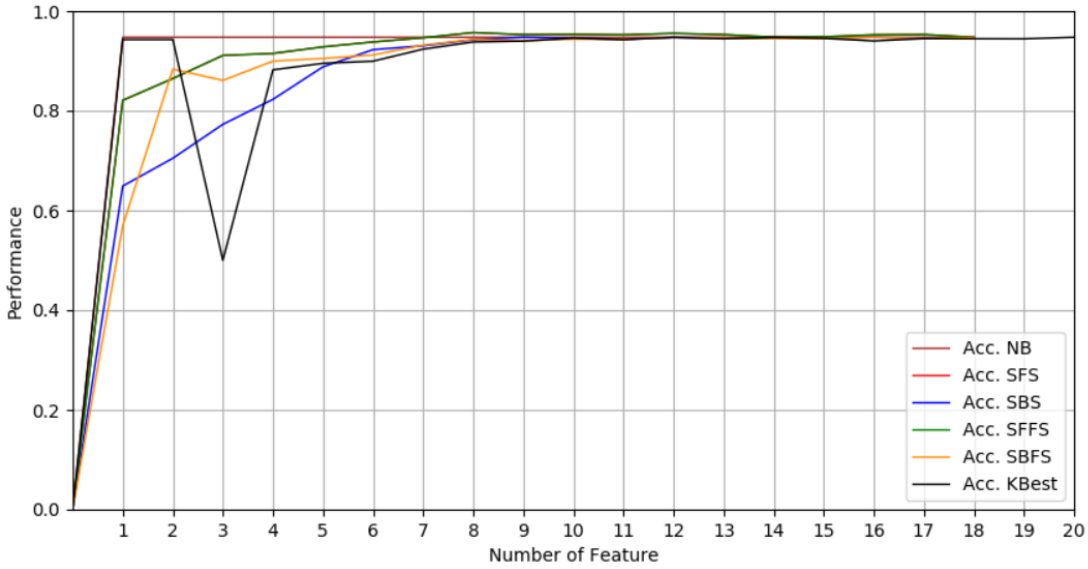
Number of Feature	RF	SFS	SBS	SFFS	SBFS	KBest
16	95,90	96,52	95,88	96,52	95,88	95,42
17	95,90	96,60	95,45	96,60	95,45	95,56
18	95,90	95,90	95,90	95,90	95,90	95,90

Tabel 3 Kinerja Model (AUC)

Number of Feature	RF	SFS	SBS	SFFS	SBFS	KBest
1	0,9474	0,8209	0,6490	0,8209	0,5690	0,5000
2	0,9474	0,8646	0,7047	0,8646	0,8834	0,8820
3	0,9474	0,9109	0,7727	0,9109	0,8609	0,8952
4	0,9474	0,9150	0,8229	0,9150	0,8995	0,8995
5	0,9474	0,9281	0,8881	0,9281	0,9049	0,9236
6	0,9474	0,9379	0,9226	0,9379	0,9118	0,9379
7	0,9474	0,9463	0,9303	0,9463	0,9317	0,9398
8	0,9474	0,9570	0,9424	0,9570	0,9433	0,9457
9	0,9474	0,9528	0,9473	0,9528	0,9401	0,9428
10	0,9474	0,9533	0,9454	0,9533	0,9435	0,9470
11	0,9474	0,9524	0,9433	0,9524	0,9433	0,9445
12	0,9474	0,9553	0,9472	0,9553	0,9472	0,9471
13	0,9474	0,9529	0,9456	0,9529	0,9456	0,9453
14	0,9474	0,9476	0,9448	0,9476	0,9448	0,9400
15	0,9474	0,9482	0,9446	0,9482	0,9446	0,9448
16	0,9474	0,9524	0,9473	0,9524	0,9473	0,9445
17	0,9474	0,9529	0,9447	0,9529	0,9447	0,9443
18	0,9474	0,9474	0,9474	0,9474	0,9474	0,9474



Gambar 3 Grafik Akurasi Model



Gambar 4 Grafik AUC Model

Berdasarkan nilai pada tabel dan grafik kinerja model tersebut dapat dinyatakan bahwa model yang menerapkan fitur seleksi SFS dan SFFS dengan 8 fitur dapat memberikan nilai terbaik. 8 Fitur yang dipilih adalah Breathing Problem, Dry Cough, Asthma, Headache, Heart Disease, Abroad travel, Attended Large Gathering, dan Family working in Public Exposed Places. Dari hasil penerapan pemilihan fitur menunjukkan bahwa tidak semua fitur relevan dan dapat berpengaruh positif terhadap kinerja model.

4 Kesimpulan

Diagnosa untuk mengetahui seseorang terkena COVID-19 sangat penting agar dapat segera ditangani dengan cepat dan mengurangi penyebaran. Diagnosa dapat dilakukan dengan mengembangkan model prediksi/diagnosa menggunakan machine learning. Dataset COVID-19 memiliki atribut yang banyak dapat memperlambat proses diagnosa/prediksi dan dapat menurunkan kinerja jika tidak relevan. Maka pada penelitian ini diusulkan penerapan seleksi fitur. Hasil percobaan menunjukkan bahwa penerapan fitur seleksi dapat membantu memilih fitur yang relevan dan dapat meningkatkan kinerja model. Algoritma pemilihan fitur terbaik untuk memprediksi COVID-19 adalah SFS dan SFFS dengan 8 fitur, yaitu Breathing Problem, Dry Cough, Asthma, Headache, Heart Disease, Abroad travel, Attended Large Gathering, dan Family working in Public Exposed Places.

5 Saran

Hasil eksperimen menunjukkan bahwa model yang diusulkan belum memberikan kinerja yang sangat baik. Untuk penelitian selanjutnya, dapat dicoba dengan algoritma pengklasifikasi yang lain seperti SVM (*Support Vektor Machine*), *Random Forests*, ANN (*Artificial Neural Network*), DT (*Decision Tree*) atau yang lainnya. Atau ditingkatkan dengan teknik ensemble, seperti teknik *boosting*, *bagging*, dan *stacking*.

References

Burhan, E., Susanto, A. D., Nasution, S. A., Ginanjar, E., Pitoyo, W., Susilo, A., & Dkk. (2020). Pedoman Tatalaksana COVID-19. In *Pedoman Tatalaksana COVID-19* (3rd ed.).

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Napoli, R. Di. (2022). Features, Evaluation, and Treatment of Coronavirus (COVID-19). *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–., December 2020*, 1–49.

Cennimo, D. J. (2024). *Coronavirus Disease 2019 (COVID-19)* (pp. 1–4). Medscape.

Gorunescu, F. (2011). Data mining: concepts and techniques. In *Chemistry &* https://doi.org/10.1007/978-3-642-19721-5

Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., Tan, Y.-Y., Chen, S.-D., Jin, H.-J., Tan, K.-S., Wang, D.-Y., & Yan, Y. (2020). The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak – An Update on the

Status. *Military Medical Research*, *7*(11), 1–10. https://doi.org/10.1186/s40779-020-00240-0

Jiawei, H., Kamber, M., Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *San Francisco, CA, itd: Morgan Kaufmann*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Liu, H., Jiang, H., & Zheng, R. (2016). The Hybrid Feature Selection Algorithm Based on Maximum Minimum Backward Selection Search Strategy for Liver Tissue Pathological Image Classification. *Computational and Mathematical Methods in Medicine*, *2016*. https://doi.org/10.1155/2016/7369137

Mendez, C. M. (2020). Solidarity" Clinical Trial for COVID -19 Treatments. *Brazilian Journal of Impleantology and Health Sciences*, 1–6.

Nair, R., & Bhagat, A. (2019). Feature selection method to improve the accuracy of classification algorithm. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6), 124–127.

Oxford University. (2022). *The RECOVERY Trial - two years on* (pp. 1–2). Oxford University.

Sahin, A. R., Erdogan, A., Mutlu Agaoglu, P., Dineri, Y., Cakirci, A. Y., Senel, M. E., Okyay, R. A., & Tasdogan, A. M. (2020). 2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature. *Eurasian Journal of Medicine and Oncology*, *4*(1), 1–7. https://doi.org/10.14744/ejmo.2020.12220

Sindi, S., Ningse, W. R. O., Sihombing, I. A., Ilmi R.H.Zer, F., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia. *Jti*, *4*(1), 166–173.

World Health Organization. (2020). Laboratory Testing for 2019 Novel Coronavirus (2019-nCoV) in Suspected Human Cases. In *WHO - Interim guidance* (pp. 1–7). World Health Organization. https://www.who.int/publications/i/item/WHO-2019-nCoV-lab-testing-2021.1-eng

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation*, *20*(4), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

Yanti, E., Fridalni, N., & Harmawati. (2020). Mencegah Penularan Virus Corona. *Journal Abdimas Saintika*, *2*, 7. https://jurnal.syedzasaintika.ac.id/index.php/abdimas/article/view/553/pdf