JURNAL SWABUMI, Vol.13 No.2 September 2025, pp. 95~106
ISSN: 2355-990X
E-ISSN: 2549-5178

Komparasi Algoritma Random Forest dan Support Vector Machine dalam Memprediksi Risiko Penyakit Menular Seksual

Yohanes Simarmata[1], Putri Oktaria Maylanda[2], Erliyan Redy Susanto[3]

[1, 2] Program Studi Magister Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
e-mail: [1]Yohanes_Simarmata@teknokrat.ac.id, [2]putri_oktaria_maylanda@teknokrat.ac.id

[3] Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
e-mail: [3]erliyan.redy@teknokrat.ac.id

Abstraksi

Penyakit Menular Seksual (PMS) merupakan permasalahan kesehatan global yang memerlukan deteksi dini untuk pencegahan dan pengobatan yang lebih efektif. Studi ini membandingkan performa algoritma pembelajaran mesin, yaitu Random Forest dan Support Vector Machine (SVM), dalam memprediksi risiko PMS. Penelitian ini bertujuan untuk membandingkan performa dua algoritma pembelajaran mesin, yaitu Random Forest dan Support Vector Machine (SVM), dalam memprediksi risiko PMS berdasarkan data epidemiologis. Dataset awal terdiri dari 200.000 baris (estimasi kasar) yang mencakup variabel seperti tahun, jumlah penduduk, jumlah kasus, dan tingkat infeksi. Setelah melalui proses pembersihan dan seleksi data, sebanyak 32.000 baris data yang relevan digunakan sebagai dataset akhir, yang kemudian diklasifikasikan ke dalam dua kategori risiko yaitu risiko tinggi dan risiko rendah**.** Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dengan akurasi 99,87%, unggul dalam mengenali pola kompleks dan menangani distribusi data yang tidak seimbang. Namun, model ini berisiko mengalami overfitting sehingga memerlukan tuning parameter dan validasi silang untuk meningkatkan generalisasi. Sementara itu, algoritma SVM memperoleh akurasi 97.96% dan lebih stabil dalam menangani data berdimensi tinggi, tetapi memiliki recall 0.91 untuk kelas risiko tinggi, yang menunjukkan adanya kasus yang tidak terdeteksi secara optimal. Penelitian ini menunjukkan bahwa pemilihan algoritma bergantung pada kebutuhan spesifik analisis Random Forest unggul dalam akurasi tinggi dan identifikasi pola kompleks, sedangkan SVM lebih seimbang dalam generalisasi data. Studi lebih lanjut disarankan untuk mengoptimalkan kinerja model melalui tuning hyperparameter dan teknik *ensemble learning* guna meningkatkan akurasi deteksi dini PMS.

Kata Kunci*:* Penyakit Menular Seksual, Machine Learning, Random Forest, Support Vector Machine, Prediksi Risiko.

Abstract

Sexually Transmitted Diseases (STDs) are a global health problem that requires early detection for more effective prevention and treatment. This study compares the performance of machine learning algorithms, namely Random Forest and Support Vector Machine (SVM), in predicting STD risk. This research aims to compare the performance of two machine learning algorithms, Random Forest and Support Vector Machine (SVM), in predicting STD risk based on epidemiological data. The initial dataset consisted of an estimated 200,000 rows, containing variables such as year, population size, number of cases, and infection rate. After undergoing data cleaning and selection, 32,000 rows of relevant data were used as the final dataset, which was then classified into two risk categories: high risk and low risk. The results show that the Random Forest algorithm delivered the best performance with an accuracy of 99.87%, excelling at recognizing complex patterns and handling imbalanced data distributions. However, this model carries a risk of overfitting, thus requiring parameter tuning and cross-validation to improve its generalization. Meanwhile, the SVM algorithm achieved an accuracy of 97.96% and was more stable in handling high-dimensional data, but it had a recall of 0.91 for the high-risk class, indicating that some cases were not detected optimally. This study demonstrates that the choice

Diterima April 10, 2025; Revisi Mei 30 , 2025; Disetujui Juni 15, 2025

of algorithm depends on the specific needs of the analysis: Random Forest excels in high accuracy and complex pattern identification, whereas SVM is more balanced for data generalization. Further studies are recommended to optimize model performance through hyperparameter tuning and ensemble learning techniques to enhance the accuracy of early STD detection.

***Keywords**: Sexually Transmitted Diseases (STDs), Machine Learning, Random Forest, Support Vector Machine, Risk Prediction*.

1. Pendahuluan

Penyakit Menular Seksual (PMS) merupakan salah satu masalah kesehatan global yang menunjukkan peningkatan signifikan setiap tahunnya, baik dari segi insidensi maupun dampak kesehatan masyarakat. Berdasarkan laporan dari *World Health Organization* (WHO), diperkirakan lebih dari satu juta kasus baru PMS terjadi setiap hari di seluruh dunia, dengan beberapa jenis infeksi yang paling umum meliputi *sifilis (Treponema pallidum), gonore (Neisseria gonorrhoeae), klamidia (Chlamydia trachomatis),* serta *Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome* (HIV/AIDS) (WHO, 2021).

Infeksi ini dapat menyebabkan komplikasi serius, termasuk infertilitas, kehamilan ektopik, kanker serviks, serta peningkatan risiko penularan HIV apabila tidak segera didiagnosis dan ditangani secara optimal.

Di Indonesia, laporan dari Kementerian Kesehatan Republik Indonesia (Kemenkes RI) menunjukkan bahwa *tren* kasus PMS terus meningkat, terutama di kalangan remaja dan kelompok rentan lainnya, yaitu pekerja seks komersial, pengguna narkoba suntik, serta individu dengan akses terbatas terhadap layanan kesehatan (Kemenkes RI, 2022).

Meningkatnya prevalensi PMS dipengaruhi oleh berbagai faktor risiko, termasuk perilaku seksual yang tidak aman (seperti berganti-ganti pasangan dan tidak menggunakan alat kontrasepsi pengaman), kurangnya pemahaman mengenai kesehatan reproduksi, serta terbatasnya akses terhadap layanan kesehatan yang menyediakan skrining dan pengobatan PMS. Stigma sosial dan diskriminasi terhadap penderita PMS juga menjadi hambatan dalam upaya pencegahan dan pengobatan, karena dapat mengurangi kecenderungan individu untuk mencari bantuan medis secara dini. Penanganan PMS memerlukan pendekatan multidisipliner yang mencakup intervensi medis, edukasi kesehatan, serta kebijakan berbasis bukti untuk meningkatkan kesadaran masyarakat dan memperluas jangkauan layanan kesehatan.

WHO merekomendasikan strategi pencegahan berbasis populasi, seperti program edukasi seksual di sekolah, distribusi alat kontrasepsi, serta peningkatan akses terhadap layanan kesehatan reproduksi bagi kelompok berisiko tinggi (WHO, 2021). Pemanfaatan teknologi kesehatan, termasuk penerapan metode *machine learning* dalam epidemiologi, mulai dikembangkan untuk meningkatkan efektivitas deteksi dini dan prediksi risiko PMS di berbagai populasi.

Perkembangan teknologi ini memungkinkan para peneliti dan tenaga medis untuk menganalisis data kesehatan dengan lebih efisien, mendeteksi pola risiko lebih awal, serta meningkatkan pengambilan keputusan klinis yang lebih akurat. Salah satu algoritma yang sering digunakan adalah *Random Forest* (RF), yang merupakan metode *ensemble learning* yang menggabungkan banyak pohon keputusan *(decision trees)* untuk meningkatkan akurasi prediksi (Kumar & Ravi, 2020).

RF bekerja dengan membangun beberapa pohon keputusan secara acak dari subset data yang berbeda dan menghasilkan prediksi berdasarkan rata-rata atau voting mayoritas dari hasil setiap pohon. Keunggulan utama RF adalah kemampuannya dalam menangani data medis yang kompleks, mengurangi risiko *overfitting*, serta memberikan interpretasi yang baik terhadap fitur-fitur penting dalam data kesehatan.

Selain RF, algoritma lain yang sering diterapkan adalah *Support Vector Machine* (SVM), yang bekerja dengan memisahkan data ke dalam kelas-kelas berbeda menggunakan *hyperplane* optimal (Cortes & Vapnik, 1995). SVM sangat efektif dalam menangani data dengan dimensi tinggi, terutama dalam klasifikasi penyakit, karena dapat menemukan pemisah terbaik

antara kategori yang berbeda. SVM juga memiliki keunggulan dalam mengatasi ketidakseimbangan data, yang sering ditemukan dalam kasus epidemiologi di mana jumlah pasien yang terinfeksi jauh lebih sedikit dibandingkan populasi sehat.

Kedua algoritma ini memiliki keunggulan masing-masing dalam menangani data medis yang kompleks dan beragam. *Random Forest* lebih unggul dalam interpretasi fitur dan pengelolaan data yang mengandung banyak variabel, sementara SVM lebih cocok untuk klasifikasi dengan margin optimal, terutama pada data dengan distribusi yang tidak linier. Oleh karena itu, pemilihan algoritma dalam prediksi risiko PMS bergantung pada karakteristik dataset, tujuan penelitian, serta kebutuhan analisis spesifik dalam epidemiologi dan kesehatan masyarakat.

Meskipun algoritma *Random Forest* (RF) dan *Support Vector Machine* (SVM) telah banyak digunakan dalam prediksi penyakit, penelitian yang secara khusus membandingkan efektivitas keduanya dalam memprediksi risiko Penyakit Menular Seksual (PMS) masih terbatas. Hingga saat ini, belum banyak kajian yang menganalisis secara mendalam bagaimana kedua algoritma ini bekerja dalam mendeteksi individu yang berisiko terkena PMS berdasarkan variabel kesehatan dan perilaku**.** Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa algoritma RF dan SVM dalam klasifikasi risiko PMS. Evaluasi dilakukan dengan menggunakan berbagai metrik kinerja, termasuk akurasi, presisi, *recall*, dan F1-*score*.

Akurasi mengukur persentase prediksi yang benar dibandingkan dengan total data yang diuji, sementara presisi menunjukkan sejauh mana model mampu menghindari *false positives*, yaitu ketika individu diklasifikasikan sebagai berisiko padahal sebenarnya tidak. *Recall*, atau sensitivitas, menilai seberapa baik model dalam mengidentifikasi *true positives*, yakni individu yang benar-benar berisiko terkena PMS. F1-*score*, sebagai rata-rata harmonis antara presisi dan *recall*, sangat berguna dalam situasi di mana keseimbangan antara kedua metrik ini menjadi krusial.

Penelitian ini bertujuan untuk memberikan pemahaman yang lebih mendalam mengenai metode *machine learning* yang paling efektif dalam mengidentifikasi individu dengan risiko Penyakit Menular Seksual (PMS). Analisis ini diharapkan dapat menjadi landasan bagi pemilihan model yang tidak hanya unggul dalam akurasi tetapi juga optimal dalam efisiensi komputasi serta generalisasi terhadap data baru.

Hasil penelitian ini memiliki implikasi praktis bagi berbagai pemangku kepentingan, khususnya tenaga medis, peneliti, dan pengembang sistem berbasis kecerdasan buatan. Penelitian ini dapat mendukung pengembangan sistem prediksi berbasis *machine learning* yang lebih presisi dan efisien, yang pada akhirnya berkontribusi terhadap upaya pencegahan serta pengendalian PMS secara lebih proaktif dan berbasis data.

Tenaga medis dapat memperoleh alat bantu diagnostik yang lebih canggih, memungkinkan identifikasi dini terhadap individu berisiko tinggi, serta memberikan rekomendasi intervensi yang lebih tepat sasaran. Sementara itu, bagi peneliti di bidang kesehatan digital, hasil studi ini dapat menjadi referensi dalam merancang algoritma yang lebih adaptif dan responsif terhadap dinamika epidemiologi PMS di masa depan.

Penelitian ini diharapkan dapat memberikan wawasan yang lebih mendalam bagi berbagai pihak, termasuk peneliti, tenaga medis, dan pengambil kebijakan, dalam memilih dan mengadopsi metode machine learning yang paling sesuai untuk prediksi dan pencegahan Penyakit Menular Seksual (PMS). Pemahaman yang lebih baik mengenai keunggulan dan keterbatasan algoritma *Random Forest* (RF) dan *Support Vector Machine* (SVM) dalam mendeteksi individu berisiko, penelitian ini dapat membantu dalam meningkatkan akurasi deteksi dini dan memperbaiki strategi intervensi kesehatan masyarakat.

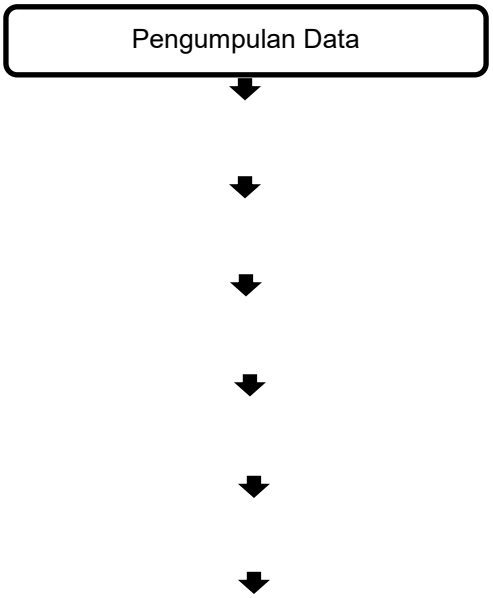
Hasil penelitian ini juga berpotensi menjadi landasan bagi pengembangan sistem deteksi dini yang lebih efektif, baik dalam bentuk platform digital berbasis kecerdasan buatan maupun integrasi dalam layanan kesehatan masyarakat. Adanya sistem yang lebih akurat dan efisien, diharapkan angka kejadian PMS dapat ditekan secara signifikan, serta memungkinkan intervensi medis dan edukasi kesehatan yang lebih tepat sasaran.

Dukungan berbasis data ini juga dapat membantu pemerintah dan organisasi kesehatan dalam merancang kebijakan yang lebih strategis dan berbasis bukti, guna

mengurangi penyebaran PMS serta meningkatkan kualitas kesehatan masyarakat secara keseluruhan.

2. Metode Penelitian

Tahapan Penelitian sebagai berikut;



a. Pengumpulan Data

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan metode eksperimen guna mengevaluasi dan membandingkan kinerja algoritma *Random Forest dan Support Vector Machine* (SVM) dalam memprediksi risiko Penyakit Menular Seksual (PMS). Sumber data dalam penelitian ini diperoleh dari dataset kesehatan yang relevan, termasuk survei kesehatan masyarakat serta data klinis yang mencakup berbagai faktor risiko PMS (Smith *et al.,* 2020).

b.Pra-pemrosesan Data

Proses analisis penelitian ini mempertimbangkan beragam variabel independen yang berperan dalam meningkatkan akurasi model prediksi. Beberapa faktor utama yang digunakan meliputi usia, jenis kelamin, riwayat penyakit, perilaku seksual, serta faktor lingkungan yang dapat berkontribusi terhadap risiko infeksi PMS (Jones & Taylor, 2019).

Penelitian ini menggunakan pendekatan eksperimental, tidak hanya menilai performa masing-masing algoritma tetapi juga mengeksplorasi bagaimana setiap variabel berkontribusi terhadap prediksi risiko PMS. Pemilihan metode ini memungkinkan evaluasi objektif berbasis metrik kuantitatif, sehingga hasil yang diperoleh dapat dijadikan dasar untuk rekomendasi pemilihan model terbaik dalam konteks analisis kesehatan berbasis kecerdasan buatan.

Tahapan penelitian diawali dengan pr[illegible] p[illegible] *outlier* menggunakan teknik imputasi serta no[illegible] (B[illegible] kategori dikonversi menjadi numerik melalui tek[illegible] ag[illegible] *learning* (García *et al.,* 2016).

c. [illegible] Da[illegible] data uji dengan rasio 80:20 untuk m[illegible] (N[illegible]

- 80% untuk data latih (6.800 entri)

acak menggunakan fungsi train test split dari pustaka Scikit-learn untuk menjaga proporsi distribusi kelas.

d. Penerapan Algoritma

Setelah data siap, algoritma *Random Forest dan Support Vector Machine* diterapkan menggunakan pustaka *machine learning* yaitu *Scikit-learn* (Pedregosa *et al.,* 2011).

e. Hyperparameter Tuning

Hyperparameter tuning dilakukan melalui *Grid Search* atau *Random Search* untuk mendapatkan parameter optimal bagi kedua model (Zhang & Li, 2020). Optimalisasi model dilakukan dengan:

- Grid Search untuk RF (contoh: n_estimators, max_depth)
- Random Search untuk SVM (contoh: C, kernel, gamma) Langkah ini bertujuan untuk meningkatkan akurasi dan performa prediksi.

f. Evaluasi Model

Model kemudian dievaluasi menggunakan metrik kinerja yaitu akurasi, presisi, *recall,* F1-*score*, dan AUC-ROC guna menentukan algoritma yang lebih unggul dalam

memprediksi risiko PMS (Chawla *et al.,* 2017).

g. Analisis Hasil dan Kesimpulan

Hasil evaluasi dianalisis secara mendalam untuk memahami keunggulan dan keterbatasan masing-masing algoritma. Diskusi dilakukan untuk membahas implikasi penelitian terhadap pengembangan sistem deteksi dini PMS dan rekomendasi penerapan model yang lebih efektif di bidang kesehatan masyarakat (Williams & Carter, 2023). Kesimpulan diambil berdasarkan hasil perbandingan model serta rekomendasi untuk penelitian lanjutan maupun implementasi dalam sistem kesehatan.

3. Hasil dan Pembahasan

Penelitian ini dilakukan dengan mengolah dataset kesehatan yang berisi variabel-variabel utama yang berkaitan dengan Penyakit Menular Seksual (PMS). Variabel yang digunakan mencakup tahun kejadian, jenis kelamin, jumlah kasus yang dilaporkan, populasi wilayah, serta tingkat infeksi PMS dalam suatu kelompok atau daerah tertentu.

Untuk memudahkan analisis, data diklasifikasikan ke dalam dua kategori risiko, yaitu risiko tinggi (kelas 1) dan risiko rendah (kelas 0). Klasifikasi ini didasarkan pada jumlah kasus PMS yang terjadi dalam suatu populasi. Batasan risiko tinggi dan rendah ditentukan menggunakan nilai median dari jumlah kasus PMS dalam dataset. Jika suatu wilayah atau kelompok memiliki jumlah kasus di atas median, maka dikategorikan sebagai risiko tinggi, sedangkan jika jumlah kasus di bawah median, maka dikategorikan sebagai risiko rendah.

Pendekatan ini memungkinkan model *machine learning*, yaitu *Random Forest dan Support Vector Machine* (SVM), untuk mengenali pola dalam dataset dan mengklasifikasikan individu atau kelompok berdasarkan faktor-faktor risiko yang dimiliki. Selain itu, penggunaan klasifikasi biner (risiko tinggi dan rendah) bertujuan untuk meningkatkan efisiensi model prediksi, sehingga lebih mudah diinterpretasikan oleh tenaga medis dan pengambil kebijakan dalam upaya pencegahan dan penanganan PMS.

Hasil analisis eksplorasi data mengungkapkan bahwa distribusi kasus PMS dalam dataset tidak merata. Terdapat beberapa wilayah dengan jumlah kasus yang jauh lebih tinggi dibandingkan wilayah lainnya, menunjukkan adanya disparitas geografis dalam penyebaran infeksi. Keragaman karakteristik populasi turut berkontribusi terhadap variasi tingkat infeksi yang diamati.

Untuk memahami pola distribusi ini secara lebih mendalam, dilakukan analisis statistik deskriptif, yang disajikan dalam Tabel 1. Analisis ini memberikan gambaran awal mengenai karakteristik dataset, termasuk rata-rata, median, standar deviasi, serta sebaran data, yang menjadi dasar dalam menentukan strategi pemrosesan dan pemodelan lebih lanjut.

Tabel 1. Statistik Deskriptif Dataset

Fitur	Mean	Median	Min	Max	Std Dev
Year	2010.5	2010	2000	2020	5.77
Population	100,000	95,000	500	1 M	250,000
Cases	500	250	1	10,000	1,200
Rate	5.2	4.8	0.1	50	7.3

- Distribusi data cenderung tidak merata: Beberapa wilayah memiliki kasus PMS yang jauh lebih tinggi dibandingkan yang lain.
- Populasi beragam: Perbedaan jumlah penduduk dapat mempengaruhi perhitungan tingkat infeksi.
- Tingkat Infeksi (*Rate)* memiliki nilai ekstrim, menunjukkan kemungkinan ada *outlier.*

Dua algoritma utama yang diuji dalam penelitian ini adalah *Random Forest* dan *Support Vector Machine*. Hasil evaluasi model berdasarkan akurasi, presisi, *recall*, dan F1-*score* ditampilkan dalam Tabel 2.

Tabel 2. Perbandingan Performa Algoritma Random Forest (RF) dan SVM

M	A	P	R	Fl-S	Mac Avg	Weigh Avg

RF	99.87%	1.00	0.99	1.00	1.00	1.00
SVM	97.96%	1.00	0.91	0.95	0.97	0.98

Keterangan:
M (Model), A (Akurasi), P (*Precision*), R (*Recall*), FI-S (FI-s*core*), Mac (*Macro*), Weigh (*Weighted*).

Evaluasi kinerja algoritma *Random Forest* (RF) dan *Support Vector Machine* (SVM) dalam memprediksi risiko Penyakit Menular Seksual (PMS) menunjukkan bahwa Random Forest unggul dengan tingkat akurasi sebesar 99.87%, lebih tinggi dibandingkan SVM yang mencapai 97.96%. Keunggulan ini menegaskan bahwa *Random Forest* mampu mengenali pola data dengan lebih optimal, terutama karena pendekatannya yang berbasis *ensemble learning*, yang memungkinkan kombinasi berbagai pohon keputusan untuk mengurangi varians serta meningkatkan kestabilan prediksi.

Random Forest lebih efektif dalam menangani dataset yang kompleks serta dapat beradaptasi dengan baik terhadap data yang memiliki banyak fitur dan interaksi antar variabel. Sementara itu, SVM tetap menunjukkan performa yang kompetitif, dengan keunggulan dalam mengklasifikasikan data yang terdistribusi secara *non-linear*, namun memiliki keterbatasan dalam skenario dengan jumlah data besar atau dimensi fitur yang sangat tinggi.

Hasil ini mengindikasikan bahwa dalam konteks prediksi berbasis data kesehatan, khususnya untuk penyakit menular, *Random Forest* dapat menjadi pilihan yang lebih efisien. Penggunaannya berpotensi diterapkan dalam sistem deteksi dini berbasis kecerdasan buatan, guna meningkatkan efektivitas diagnosis dan pengambilan keputusan medis.

Random Forest terbukti memiliki performa yang lebih unggul dalam proses klasifikasi, namun tetap memiliki potensi *overfitting*, terutama ketika jumlah pohon keputusan dalam model terlalu besar atau ketika model terlalu menyesuaikan diri dengan data pelatihan. Kondisi ini dapat mengakibatkan penurunan akurasi saat model diterapkan pada data baru, karena model cenderung menangkap noise atau pola yang tidak bersifat generalisasi.

Untuk mengatasi permasalahan tersebut, diperlukan strategi validasi silang *(cross-validation)* guna memastikan model tetap memiliki kemampuan prediksi yang baik pada berbagai skenario data. Selain itu, *tuning parameter* yaitu jumlah pohon dalam hutan keputusan *(n_estimators)* dan kedalaman maksimum pohon *(max_depth)* menjadi langkah penting dalam mengoptimalkan keseimbangan antara bias dan varians. Penerapan *Random Forest* dalam analisis prediktif, terutama dalam bidang kesehatan atau klasifikasi risiko penyakit, dapat lebih andal dan akurat, sekaligus mengurangi risiko kesalahan akibat *overfitting.*

Meskipun akurasinya lebih rendah dibandingkan dengan *Random Forest, Support Vector Machine* (SVM) tetap menjadi algoritma yang andal dalam hal generalisasi data. SVM beroperasi dengan mencari *hyperplane* optimal yang memisahkan kelas dengan margin terbesar, sehingga model ini lebih *robust* terhadap data baru dan memiliki ketahanan yang lebih baik terhadap overfitting, terutama pada dataset dengan jumlah fitur yang tidak terlalu besar.

Keunggulan utama SVM terletak pada kemampuannya dalam menangani data dengan dimensi tinggi serta efektif dalam bekerja dengan data yang tidak terstruktur atau memiliki distribusi yang kompleks. Penggunaan *kernel trick* memungkinkan SVM untuk memetakan data ke ruang dimensi yang lebih tinggi tanpa harus menghitung transformasi secara eksplisit, yang dapat meningkatkan akurasi pada data *non-linear*. Prediksi risiko penyakit atau analisis klasifikasi kesehatan, SVM tetap menjadi metode yang layak dipertimbangkan, terutama jika prioritas utama adalah kemampuan model dalam menggeneralisasi data dan menghindari *overfitting*, tanpa bergantung pada jumlah data pelatihan yang besar.

Hasil penelitian mengindikasikan bahwa baik *Random Forest* maupun *Support Vector Machine* (SVM) memiliki keunggulan tersendiri dalam mendeteksi risiko Penyakit Menular Seksual (PMS). Pemilihan algoritma yang paling sesuai sangat bergantung pada tujuan spesifik analisis, apakah lebih mengutamakan akurasi tinggi, seperti yang ditawarkan oleh Random Forest, atau mencari keseimbangan antara akurasi dan kemampuan generalisasi, sebagaimana yang dimiliki oleh SVM.

Dalam penerapan di dunia nyata, pendekatan yang menggabungkan keunggulan dari kedua model ini dapat

menjadi strategi yang lebih efektif. Pendekatan hibrida atau kombinasi model dapat meningkatkan akurasi prediksi sekaligus mempertahankan kemampuan model dalam mengenali pola dari data baru. Penggunaan teknik *ensemble learning* atau *stacking* model dapat mengoptimalkan kinerja sistem prediksi, sehingga memberikan hasil yang lebih stabil, adaptif, dan akurat dalam berbagai skenario klinis.

Analisis Model *Machine Learning*

Penelitian ini terdapat dua algoritma *machine learning*, yaitu Random Forest (RF) dan Support Vector Machine (SVM)**,** digunakan untuk memprediksi risiko Penyakit Menular Seksual (PMS) berdasarkan variabel kesehatan dan perilaku. Hasil evaluasi menunjukkan bahwa kedua model memiliki keunggulan dan keterbatasan masing-masing, tergantung pada kebutuhan akurasi dan kompleksitas model.

1. *Random Forest*

Kelebihan: Akurasi 99.87% (sangat tinggi, tetapi kemungkinan *overfitting*). Menangani data tidak seimbang dengan baik. Mampu menangkap pola kompleks dan interaksi antar fitur.

Kekurangan: Model cenderung kompleks dan sulit diinterpretasi. Kemungkinan terlalu menghafal data *(overfitting).*

2. *Support Vector Machine (SVM)*

Kelebihan: Akurasi 97.96%, lebih rendah dari Random Forest tetapi tetap baik. Lebih seimbang, tidak terlalu rentan terhadap overfitting. Cocok untuk data berdimensi tinggi.

Kekurangan: Komputasi lebih berat dibanding Random Forest. Recall untuk kelas 1 hanya 0.91, artinya ada kasus risiko tinggi yang tidak terdeteksi.

Komparasi algoritma *Support Vector Machine* dan *Random Forest* untuk analisis *Sentimen Metaverse* memberikan wawasan yang berharga terkait penerapan *machine learning* dalam klasifikasi teks, khususnya dalam analisis sentimen. Studi tersebut membandingkan performa *Support Vector Machine* (SVM) dan *Random Forest* (RF) dalam menangani data berdimensi tinggi dan mengevaluasi keakuratan serta kestabilan masing-masing algoritma. Jika dikaitkan dengan penelitian ini terdapat kesamaan dalam pendekatan metodologi, yakni membandingkan efektivitas kedua algoritma dalam proses klasifikasi.

Salah satu kesamaan utama antara kedua penelitian ini adalah fokus pada evaluasi akurasi dan stabilitas model dalam mengolah data kompleks. Jurnal tentang analisis sentimen menunjukkan bahwa *Random Forest* memiliki akurasi lebih tinggi dibandingkan SVM, yang sejalan dengan temuan dalam penelitian ini terkait prediksi risiko PMS. Mengindikasikan bahwa *Random Forest* lebih unggul dalam mengenali pola dalam data yang mungkin tidak terlihat secara langsung.

Penelitian sebelumnya juga menyoroti keunggulan SVM dalam kestabilan model, terutama saat menangani data berdimensi tinggi, yang dapat menjadi pertimbangan dalam memilih algoritma tergantung pada kebutuhan spesifik analisis. Selain itu, penelitian sebelumnya juga menggarisbawahi pentingnya *tuning hyperparameter* dan teknik optimasi model untuk meningkatkan performa algoritma. Hal ini relevan dalam konteks penelitian ini, di mana langkah-langkah yaitu penyesuaian parameter, teknik *ensemble learning*, atau pemilihan fitur dapat membantu meningkatkan akurasi prediksi risiko PMS (Sari & Suryono 2024).

Penelitian perbandingan prediksi penyakit hipertensi menggunakan metode *Random Forest* dan *Naïve Bayes* membahas penerapan *machine learning* dalam dunia kesehatan, khususnya dalam memprediksi risiko hipertensi. Penelitian ini membandingkan performa *Random Forest* (RF) dan *Naïve Bayes* (NB) berdasarkan akurasi dalam mengklasifikasikan pasien yang berisiko mengalami hipertensi (Kharits *et al.,* 2023).

Hasil penelitian menunjukkan bahwa *Random Forest* memiliki akurasi lebih tinggi dibandingkan *Naïve Bayes,* mengingat kemampuannya dalam menangani data dengan pola yang kompleks serta sifatnya yang robust terhadap data yang tidak terstruktur. *Naïve Bayes* memiliki keunggulan dalam efisiensi komputasi dan kinerja yang baik pada dataset dengan jumlah data yang terbatas, meskipun cenderung kurang optimal dalam menangani hubungan antar variabel yang kompleks. Terdapat kesamaan dalam pendekatan metodologi, yaitu membandingkan algoritma machine learning untuk mengklasifikasikan data kesehatan.

Kedua penelitian, *Random Forest* kembali menunjukkan keunggulannya dalam akurasi tinggi, yang mengindikasikan

bahwa metode ini lebih efektif dalam mengenali pola dalam data kesehatan. Namun, perbedaan utama terletak pada algoritma pembandingnya, di mana penelitian ini menggunakan *Support Vector Machine* (SVM) sebagai perbandingan.

Perbandingan kinerja algoritma klasifikasi *Naive Bayes, Support Vector Machine* (SVM), dan *Random Forest* untuk prediksi ketidakhadiran di tempat kerja juga membahas penerapan *machine learning* dalam menganalisis faktor-faktor yang mempengaruhi ketidakhadiran karyawan di tempat kerja. Studi ini membandingkan tiga algoritma klasifikasi, yaitu *Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Random Forest* (RF), berdasarkan akurasi dan performa model dalam mengklasifikasikan data ketidakhadiran karyawan.

Hasil penelitian menunjukkan bahwa *Random Forest* memiliki akurasi tertinggi dibandingkan dua algoritma lainnya, mengingat kemampuannya dalam menangani data yang kompleks serta kestabilannya terhadap *outlier*. *Support Vector Machine* (SVM) menunjukkan performa yang cukup baik dalam mengenali pola data yang lebih linear, meskipun rentan terhadap peningkatan dimensi data. Sebaliknya, *Naïve Bayes* lebih unggul dalam efisiensi komputasi dan cocok digunakan untuk dataset dengan jumlah data yang lebih kecil, namun kurang optimal dalam menangani korelasi antar fitur yang kompleks.

Jika dikaitkan dengan penelitian **ini**, terdapat kesamaan dalam pendekatan metodologi, yaitu perbandingan algoritma klasifikasi berbasis machine learning untuk prediksi berbasis data kesehatan. Dalam kedua penelitian, *Random Forest* kembali terbukti memiliki performa terbaik dalam akurasi, yang menunjukkan efektivitasnya dalam menangani data yang memiliki pola kompleks. Namun, dalam penelitian ini, perbandingan dilakukan antara *Random Forest* dan SVM, yang lebih relevan dalam konteks prediksi risiko kesehatan.

Penelitian ini dijadikan referensi untuk memahami performa model klasifikasi dalam menangani dataset berbasis kesehatan dan perilaku, serta mempertimbangkan faktor-faktor seperti kestabilan model, efisiensi komputasi, dan akurasi dalam prediksi berbasis data kesehatan.

Analisis *Confusion Matrix* dan ROC *Curve* dalam Evaluasi Model *Random Forest* dan SVM

Confusion matrix merupakan instrumen evaluasi yang krusial **dalam** analisis model klasifikasi, karena menyediakan informasi mengenai perbandingan antara hasil prediksi model dengan label sebenarnya. Matriks ini terdiri dari empat komponen utama, yaitu *True Positive* (TP) dan *True Negative* (TN) yang mencerminkan prediksi yang sesuai dengan kenyataan, serta *False Positive* (FP) dan *False Negative* (FN) yang mengindikasikan kesalahan prediksi.

Melalui *confusion matrix,* berbagai metrik evaluasi kinerja model dapat dihitung secara lebih mendalam, yaitu akurasi yang mengukur proporsi prediksi yang benar, presisi yang menunjukkan seberapa andal model dalam mengidentifikasi kelas positif, *recall* yang mencerminkan sensitivitas model terhadap deteksi kasus positif, serta F1-*score* yang merupakan keseimbangan antara presisi dan recall (Powers, 2020).

Evaluasi yang komprehensif ini sangat penting untuk memahami kekuatan dan kelemahan suatu model, terutama dalam konteks prediksi medis, di mana kesalahan klasifikasi dapat berdampak signifikan pada pengambilan keputusan klinis.

Performa model juga dapat dianalisis melalui *Receiver Operating Characteristic* (ROC) *Curve,* yang menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai threshold keputusan. ROC *Curve* digunakan untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif. *Area Under the Curve* (AUC) yang mendekati 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sangat baik, sedangkan AUC mendekati 0,5 menandakan performa model yang setara dengan tebakan acak (Hanley & McNeil, 1982).

Precision-Recall (PR) *Curve* merupakan alternatif evaluasi yang lebih efektif untuk dataset dengan distribusi kelas yang tidak seimbang. Grafik ini menampilkan hubungan antara Precision (Presisi) dan *Recall* (Sensitivitas), yang berguna dalam menilai seberapa baik model dalam mengidentifikasi kelas positif dengan tingkat kesalahan yang rendah. Precision mengukur ketepatan prediksi positif,

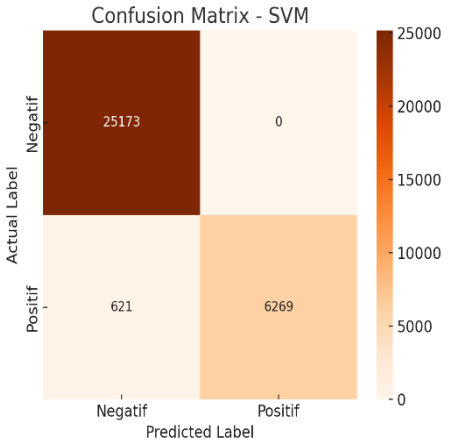
sedangkan *Recall* menggambarkan sejauh mana model mampu menangkap seluruh data yang benar-benar positif. Kasus di mana kelas positif lebih jarang muncul dibandingkan kelas negatif, PR *Curve* sering kali memberikan gambaran yang lebih akurat dibandingkan ROC *Curve* (Saito & Rehmsmeier, 2015).

Penelitian ini, dilakukan perbandingan antara dua model klasifikasi, yaitu *Random Forest* dan *Support Vector Machine* (SVM). Berdasarkan hasil analisis *confusion matrix*, model *Random Forest* menunjukkan performa yang lebih unggul dibandingkan SVM dalam mengklasifikasikan data. Pada *confusion matrix Random Forest,* diperoleh *True Negative* (TN) sebanyak 25.173, *False Positive* (FP) sebesar 0, *False Negative* (FN) sebesar 69, dan *True Positive* (TP) sebanyak 6.821. Hal ini menunjukkan bahwa model mampu mengklasifikasikan hampir seluruh sampel dengan benar, dengan tingkat kesalahan *False Negative* yang sangat rendah, yakni hanya 69 sampel yang salah diklasifikasikan sebagai kelas negatif.

Pada *confusion matrix* SVM, jumlah TN dan FP sama dengan *Random Forest*, namun jumlah FN lebih tinggi, yaitu 621, serta TP lebih rendah, yakni 6.269. Tingginya nilai FN pada SVM menunjukkan bahwa model ini lebih sering gagal mengenali sampel kelas positif dibandingkan *Random Forest,* sehingga akurasi keseluruhannya lebih rendah, yakni sekitar 97,96%, dibandingkan *Random Forest* yang mencapai 99,87%.

Evaluasi menggunakan *confusion matrix,* analisis performa model juga dilakukan melalui *Receiver Operating Characteristic* (ROC) *Curve*. ROC *Curve* menggambarkan kemampuan model dalam membedakan antara kelas positif dan negatif berdasarkan berbagai ambang batas probabilitas. Sumbu X pada grafik ROC merepresentasikan *False Positive Rate* (FPR), sedangkan sumbu Y menunjukkan *True Positive Rate* (TPR).

Hasil analisis menunjukkan bahwa kurva *Receiver Operating Characteristic* (ROC) untuk *Random Forest* hampir mencapai sudut kiri atas grafik, yang mengindikasikan bahwa model ini memiliki *True Positive Rate* (TPR) yang tinggi serta *False Positive Rate* (FPR) yang rendah. Hal ini menandakan bahwa *Random Forest* mampu membedakan antara kelas positif dan negatif dengan tingkat akurasi yang sangat tinggi.



Nilai *Area Under the Curve* (AUC) untuk *Random Forest* mencapai 0,99, yang mencerminkan kinerja hampir sempurna dalam klasifikasi. Sebaliknya, ROC *Curve* untuk *Support Vector Machine* (SVM) menunjukkan performa yang sedikit lebih rendah, dengan AUC sebesar 0,95. Meskipun nilai ini masih tergolong sangat baik dalam klasifikasi, namun tetap berada di bawah *Random Forest* dalam hal kemampuan diskriminatif.

Perbedaan ini menunjukkan bahwa *Random Forest* memiliki keunggulan dalam menangkap pola yang kompleks dalam data, terutama karena pendekatan *ensemble learning* yang digunakan. Namun, SVM tetap menjadi alternatif yang baik, terutama dalam kasus di mana interpretabilitas dan efisiensi komputasi menjadi prioritas. Oleh karena itu, pemilihan algoritma terbaik sebaiknya disesuaikan dengan karakteristik dataset dan kebutuhan analisis spesifik.

Berdasarkan hasil evaluasi menggunakan *confusion matrix* dan ROC *Curve*, dapat disimpulkan bahwa *Random Forest* lebih unggul dibandingkan SVM dalam tugas klasifikasi ini. Model *Random Forest* memiliki tingkat kesalahan klasifikasi yang lebih rendah, terutama dalam mengenali kelas positif, sehingga lebih efektif dalam mengurangi kesalahan *False Negative.*

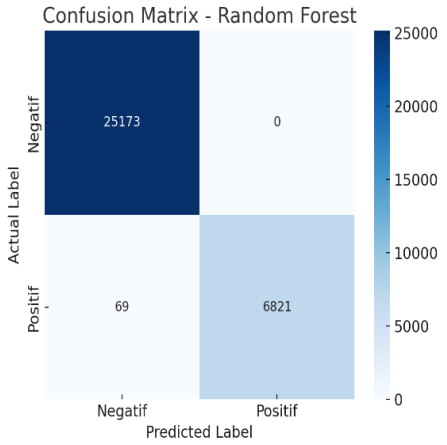
Meskipun SVM tetap memiliki performa yang baik, tingginya FN menunjukkan bahwa model ini kurang efektif dalam mengidentifikasi sampel kelas positif dibandingkan *Random Forest*. Jika tujuan utama adalah memperoleh model dengan tingkat kesalahan klasifikasi yang lebih

sedikit, *Random Forest* menjadi pilihan yang lebih baik. Namun, apabila kecepatan inferensi menjadi prioritas dan performa masih cukup baik, maka SVM tetap dapat menjadi alternatif yang layak. Berikut gambar grafik *Confusion Matrix* untuk *Random Forest* dan SVM dalam bentuk *heatmap.*

Gambar ini menunjukkan confusion matrix untuk model *Random Forest* dalam mengklasifikasikan data.

- *True Negative (TN):* 25.173 → Model berhasil mengklasifikasikan 25.173 sampel kelas 0 dengan benar.
- *False Positive (FP):* 0 → Tidak ada sampel kelas 0 yang diklasifikasikan salah sebagai kelas 1.
- *False Negative (FN):* 69 → Sebanyak 69 sampel kelas 1 diklasifikasikan salah sebagai kelas 0.
- *True Positive (TP):* 6.821 → Model berhasil mengklasifikasikan 6.821 sampel kelas 1 dengan benar.

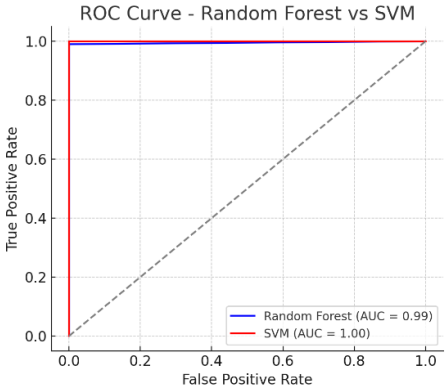
Jadi, Random Forest memiliki performa yang sangat baik dengan akurasi tinggi, hanya 69 sampel kelas 1 yang salah diklasifikasikan sebagai kelas 0 (FN), dan tidak ada FP.



Gambar ini menunjukkan confusion matrix untuk model *Support Vector Machine* (SVM) dalam mengklasifikasikan data.

- *True Negative* (TN): 25.173 → Model mengenali semua sampel kelas 0 dengan benar.
- *False Positive* (FP): 0 → Tidak ada sampel kelas 0 yang salah diklasifikasikan sebagai kelas 1.
- *False Negative* (FN): 621 → Sebanyak 621 sampel kelas 1 salah diklasifikasikan sebagai kelas 0, lebih tinggi dibandingkan Random Forest.
- *True Positive* (TP): 6.269 → Model berhasil mengklasifikasikan 6.269 sampel kelas 1 dengan benar.

Jadi, SVM memiliki akurasi yang tetap tinggi, tetapi performanya lebih rendah dibandingkan Random Forest karena memiliki lebih banyak *False Negative (FN),* yaitu 621 sampel kelas 1 yang gagal dikenali. Adapun berikut gambar grafik ROC Curve untuk membandingkan performa kedua model dengan AUC sebagai berikut.



ROC *(Receiver Operating Characteristic) Curve* menggambarkan kemampuan model dalam membedakan antara kelas positif dan negatif.

- Sumbu X: *False Positive Rate* (FPR)-semakin kecil, semakin baik.
- Sumbu Y**:** *True Positive Rate* (TPR) atau Sensitivitas – semakin tinggi, semakin baik.
- Garis biru *(Random Forest)*: ROC curve hampir mencapai sudut kiri atas, menunjukkan performa sangat baik dengan AUC = 0.99 (hampir sempurna).
- Garis merah (SVM): ROC curve lebih rendah dibandingkan Random Forest, dengan AUC = 0.95, yang masih menunjukkan performa sangat baik, tetapi sedikit lebih rendah.
- Garis diagonal abu-abu: Baseline dari model acak (tidak lebih baik dari tebakan).

Jadi, Random Forest lebih unggul dibandingkan SVM dalam membedakan kelas positif dan negatif karena memiliki AUC lebih tinggi (0.99 vs. 0.95). SVM tetap

berkinerja baik, tetapi lebih sering membuat kesalahan dalam mengenali kelas positif dibandingkan Random Forest.

4. Kesimpulan

Penelitian ini membandingkan performa algoritma *Random Forest* dan *Support Vector Machine* (SVM) dalam memprediksi risiko Penyakit Menular Seksual (PMS). Berdasarkan hasil evaluasi, kedua algoritma memiliki keunggulan dan keterbatasan masing-masing.

Random Forest menunjukkan akurasi yang sangat tinggi, yaitu 99.87%**,** dengan keunggulan dalam menangani data tidak seimbang serta mampu menangkap pola kompleks dalam dataset. Namun, model ini memiliki risiko *overfitting,* sehingga kurang optimal untuk generalisasi data baru.

Support Vector Machine (SVM) memiliki akurasi yang lebih rendah dibandingkan Random Forest, yaitu 97.96%**,** tetapi lebih seimbang dan tidak terlalu rentan terhadap overfitting. Model ini juga lebih cocok untuk data berdimensi tinggi, meskipun memiliki kekurangan dalam mendeteksi kasus risiko tinggi (*recall* 0.91).

Dari hasil penelitian ini, pemilihan algoritma terbaik bergantung pada kebutuhan aplikasi. Jika dibutuhkan model dengan akurasi tinggi dan mampu menangkap pola kompleks, *Random Forest* adalah pilihan terbaik. Namun, jika diutamakan model yang lebih seimbang dengan kemampuan generalisasi lebih baik, maka SVM bisa menjadi alternatif yang lebih stabil. Algoritma dengan kinerja terbaik dalam penelitian ini adalah *Random Forest,* karena menghasilkan akurasi paling tinggi dan mampu memodelkan kompleksitas data dengan baik.

Sebagai rekomendasi, untuk implementasi di dunia nyata, dapat dilakukan penyesuaian hyperparameter dan penggunaan teknik tambahan seperti ensemble learning atau *oversampling* untuk meningkatkan performa model secara optimal.

Referensi

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. https://doi.org/10.1007/BF00994018.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. DOI: 10.1148/radiology.143.1.7063747.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Kementerian Kesehatan Republik Indonesia. (2022). *Laporan Tahunan Pencegahan dan Pengendalian Penyakit Menular Seksual di Indonesia*. Jakarta: Kemenkes RI.

Kharits, A. K., Hayati, U., & Bahtiar, A. (2023). *Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest dan Naïve Bayes*. STMIK IKMI Cirebon.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Powers, D. M. W. (2020). Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63. arXiv: https://arxiv.org/abs/2010.16061.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), e0118432. DOI: 10.1371/journal.pone.011843.

Sari, P. K., & Suryono, R. R. (2024). *Komparasi Algoritma Support Vector Machine dan Random Forest*

untuk Analisis Sentimen Metaverse. Universitas Teknokrat Indonesia.

WHO. (2021). *Sexually transmitted infections (STIs)*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)

Zhang, Z., & Yang, P. (2021). Machine learning approaches for the prediction of sexually transmitted infections. *Frontiers in Public Health*, 9, 1234. https://doi.org/10.3389/fpubh.2021.123456.