JTRISTE, Vol.11, No.1, Maret 2024, pp. 60~66
ISSN: **2355-3677 | e-ISSN: 2460-8548**

PERBANDINGAN METODE KLASIFIKASI RANDOM FOREST DAN SUPPORT VECTOR MACHINE TERHADAP DATASET RESIKO KANKER SERVIKS

Jesly Putri Kristiani B[1], Louisa Leokadja[2], Iwan Binanto[3*]
[1,2,3] Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma
e-mail: [1] jeslybudiman@gmail.com, [2] louisaleokadja18@gmail.com, [3]iwan@usd.ac.id

Abstrak

Kanker serviks, merupakan salah satu masalah kesehatan global yang signifikan, kanker ini merupakan jenis kanker yang berkembang dari sel-sel leher rahim. Penelitian ini memfokuskan pada perbandingan efektivitas dua metode klasifikasi, yaitu Random Forest (RF) dan Support Vector Machine (SVM), dalam menilai risiko kanker serviks. Dengan menggunakan dataset yang relevan, penelitian ini bertujuan untuk mengidentifikasi keunggulan dan kelemahan masing-masing metode serta mengevaluasi kemampuan kedua algoritma tersebut dalam memberikan prediksi resiko kanker serviks. Melalui analisis perbandingan, diharapkan penelitian ini dapat memberikan wawasan yang berharga untuk pengembangan metode penilaian resiko kanker serviks yang lebih efisien. Hasil penelitian ini diharapkan dapat memberikan kontribusi pada pemahaman lebih lanjut tentang perbandingan performa antara Random Forest (RF) dan Support Vector Machine (SVM) dalam konteks penilaian resiko kanker serviks, membuka peluang untuk penerapan metode klasifikasi yang lebih optimal dalam upaya pencegahan dan deteksi dini penyakit ini.

Kata kunci: *Kanker serviks, klasifikasi, Random Forest, SVM*

Abstract

Cervical cancer is a significant global health issue, representing a type of cancer that develops from the cells of the cervix. This research focuses on comparing the effectiveness of two classification methods, namely Random Forest (RF) and Support Vector Machine (SVM), in assessing the risk of cervical cancer. Utilizing relevant dataset, the study aims to identify the strengths and weaknesses of each method and evaluate their ability to provide predictions of cervical cancer risk. Through comparative analysis, it is anticipated that this research will offer valuable insights for the development of more efficient methods for assessing the risk of cervical cancer. The results of this study are expected to contribute to a deeper understanding of the performance comparison between Random Forest (RF) and Support Vector Machine (SVM) in the context of assessing the risk of cervical cancer, opening opportunities for the optimal application of classification methods in efforts for the prevention and early detection of this disease.

Keywords*: Cervical cancer, classification, Random Forest, SVM*

1. Pendahuluan

Kanker serviks adalah tumor ganas pada wanita yang menempati urutan kedua di seluruh dunia dan serius mengancam kesehatan wanita. Menurut data dari Profil Kesehatan Indonesia tahun 2021, kanker serviks menempati peringkat kedua setelah kanker payudara, yaitu sebanyak

[*] Corresponding author : Iwan Binanto (iwan@usd.ac.id)

Diterima *: April, 2024* ***Disetujui*** *: April, 2024* ***Dipublikasikan*** *: April, 2024*

36.633 kasus atau 17,2% dari seluruh kanker pada wanita. Infeksi berisiko tinggi berkelanjutan dengan virus papilloma manusia (*Human Papilloma Virus*) telah ditentukan sebagai penyebab yang diperlukan dari kanker serviks [1]. Sebagai bentuk pencegahan dan deteksi dini, penelitian terus dilakukan untuk mengembangkan metode yang efektif dalam memprediksi risiko kanker serviks. Data kanker cerviks yang digunakan dalam penelitian ini di *download* dari www.kaggle.com [2].

Memprediksi data untuk kanker serviks sudah dilakukan sebelumnya, tentunya juga dengan berbagai algoritma, seperti naïve bayes, KNN, dan juga SVM [3], [4]. Penelitian ini akan melakukan perbandingan dua algoritma, yakni Random Forest (RF) dan Support Vector Machine (SVM) [5], [6]. Random Forest (RF) dan Support Vector Machine (SVM) merupakan algoritma dalam *machine learning*, dimana keduanya digunakan untuk tugas klasifikasi dan regresi [6]. Penelitian ini difokuskan pada perbandingan dua metode klasifikasi utama, yaitu Random Forest (RF) dan Support Vector Machine (SVM), dalam konteks deteksi dan evaluasi risiko kanker serviks. Keduanya telah menjadi fokus kajian mendalam dalam literatur machine learning untuk aplikasi kesehatan, dengan potensi memberikan wawasan yang berharga terkait efektivitas serta kelebihan dan kelemahan masing-masing metode [7].

Penelitian ini bertujuan untuk menyelidiki efektivitas dan perbandingan antara metode Random Forest (RF) dan Support Vector Machine (SVM) dalam menilai resiko kanker serviks. Penerapan metode klasifikasi seperti Random Forest (RF) dan Support Vector Machine (SVM) dalam penilaian risiko kanker serviks menjadi langkah krusial dalam mengoptimalkan upaya pencegahan dan deteksi dini penyakit yang memiliki dampak kesehatan masyarakat yang signifikan.

2. Tinjauan Pustaka

Penelitian ini bertujuan untuk menginvestigasi efektivitas dan perbandingan antara metode Random Forest (RF) dan Support Vector Machine (SVM) dalam penilaian risiko kanker serviks. Dengan menyatukan algoritma keduanya, penulis berharap dapat menghasilkan model yang memberikan prediksi yang lebih akurat, memungkinkan pencegahan dan deteksi dini yang lebih efisien. Dengan merangkum temuan-temuan tersebut, maka dalam penelitian ini diharapkan memberikan kontribusi pada pengembangan metode penilaian risiko kanker serviks yang lebih canggih dan akurat. Hasilnya diharapkan tidak hanya meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi risiko kanker serviks tetapi juga menyediakan dasar untuk tindakan pencegahan yang lebih terarah, potensial mengurangi angka kesakitan dan kematian akibat penyakit ini [8].

Berdasarkan penelusuran artikel, ditemukan bahwa metode Decision Tree mencapai jumlah tertinggi pada tahun 2019 dengan 15 artikel sedangkan metode Random Forest (RF) mendominasi dengan 23 artikel dari total 67 artikel yang dianalisis. Metode Naïve Bayes mendominasi penggunaannya pada tahun 2020 dengan 12 artikel. Sedangkan metode Support

Vector Machine (SVM) mencapai jumlah tertinggi, yaitu 18 artikel dari total 62 artikel yang dipublikasikan. Pada tahun 2021, metode Neural Network menjadi yang paling banyak digunakan dengan 18 artikel dari total 36 artikel yang diteliti [9], [10].

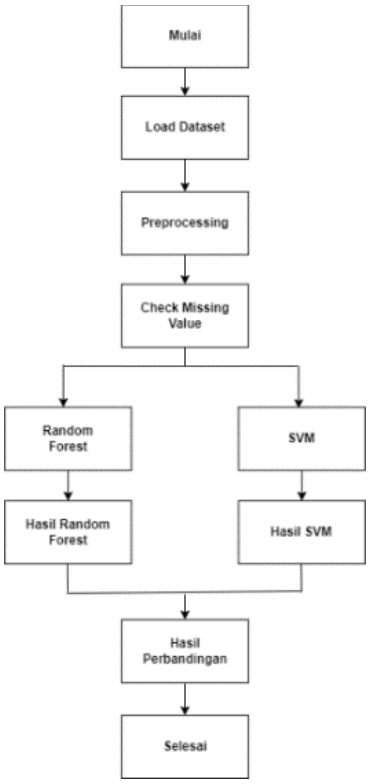
Random Forest (RF) adalah algoritma yang termasuk dalam metode penggabungan pembelajaran, khususnya *bagging ensemble learning*. Dalam proses ini, set pelatihan diambil sampel sebanyak jumlah pohon yang diinginkan menggunakan metode *Simple Random Sampling With Replacement* (SRS WR), yang dikenal sebagai *bagging*. Setiap sampel dari set pelatihan membentuk satu pohon keputusan. Saat menguji objek dalam set pengujian, proses dilakukan oleh semua pohon yang terbentuk, dan keputusan klasifikasi akhir diambil berdasarkan mayoritas suara atau keputusan yang paling banyak dipilih oleh pohon-pohon tersebut. Terdapat juga perhitungan kesalahan untuk objek-objek yang tidak digunakan selama pembuatan pohon keputusan, yang dikenal sebagai estimasi kesalahan *Out-of-bag* (OOB) [11]. Temuan dari penelitian menunjukkan bahwa Random Forest (RF) sebagai model klasifikasi berhasil mencapai akurasi sekitar 97,16% melalui penggunaan validasi silang, dengan nilai *Area Under the Curve* (AUC) sebesar 0,996. Selanjutnya, tingkat akurasi dari model klasifikasi Support Vector Machine (SVM) mencapai sekitar 96,01%, dan nilai AUC-nya adalah 0,543 [12].

Dalam dekade terakhir, Support Vector Machine (SVM) telah terbukti menjadi metode yang sangat efektif dalam pola klasifikasi, mencapai tingkat keberhasilan tinggi dalam berbagai bidang aplikasi. Kinerja yang luar biasa dari Support Vector Machine (SVM) dalam menyelesaikan berbagai masalah pembelajaran telah menarik minat banyak komunitas machine learning untuk memahami dan mengembangkannya. Support Vector Machine (SVM) bertujuan untuk menemukan *hyperplane* terbaik yang dapat memisahkan dua kelas dalam ruang input. Dengan menggunakan data pelatihan, algoritma klasifikasi Support Vector Machine (SVM) membentuk model yang dapat digunakan untuk memprediksi kelas data baru, yang disebut sebagai data pengujian [13].

Dalam hal ini, penelusuran literatur mendukung pandangan bahwa metode klasifikasi seperti Random Forest (RF) dan Support Vector Machine (SVM) contoh algoritma yang populer. Dengan demikian, penelitian ini memberikan kontribusi yang berharga dalam konteks penelitian kanker serviks dan dapat membuka jalan bagi penelitian lebih lanjut dalam bidang ini.

3. Metode Penelitian

Dataset kanker serviks yang sudah tersedia di Kaggle akan digunakan untuk menganalisa kanker serviks dalam penelitian ini, dengan dua algoritma yakni Random Forest (RF) dan Support Vector Machine (SVM) sebagai bahan perbandingan. Adapun metode penelitiannya seperti terlihat pada gambar 1.



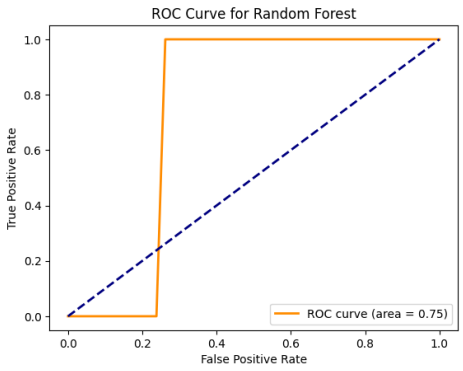
Gambar 1. Metode Penelitian

Age	0
Number of sexual partners	0
First sexual intercourse	0
Num of pregnancies	0
Smokes	0
Smokes (years)	0
Smokes (packs/year)	0
Hormonal Contraceptives	0
Hormonal Contraceptives (years)	0
IUD	0
IUD (years)	0
STDs	0
STDs (number)	0
STDs:condylomatosis	0
STDs:cervical condylomatosis	0
STDs:vaginal condylomatosis	0
STDs:vulvo-perineal condylomatosis	0
STDs:syphilis	0
STDs:pelvic inflammatory disease	0
STDs:genital herpes	0
STDs:molluscum contagiosum	0
STDs:AIDS	0
STDs:HIV	0
STDs:Hepatitis B	0
STDs:HPV	0
STDs: Number of diagnosis	0
STDs: Time since first diagnosis	0
STDs: Time since last diagnosis	0
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0
dtype: int64	

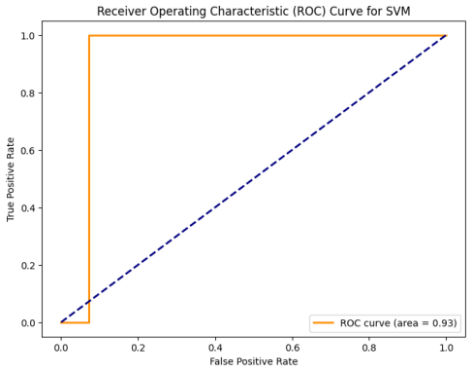
Gambar 2. Pengecekan *missing value*

Pada dataset ini semua kolom dapat digunakan untuk klasifikasi. Setelah dilakukan preprocessing data, dilakukan pengecekan *missing value*. Pada saat pengecekan ini tidak terdapat *missing value*, seperti terlihat pada Gambar 2.

Setelah *preprocessing* dan pengecekan *missing value* selesai, dilakukan *split* data berdasarkan data *testing* dan *training*. Setelah itu melakukan klasifikasi dengan algoritma Random Forest (RF) yaitu mencari nilai yang paling optimal dari jumlah yang ditetapkan pada "n_estimators". Untuk klasifikasi Support Vector Machine (SVM) memasukkan jumlah iterasi untuk dicoba ke data *training*.



Gambar 3. ROC Curve Random Forest



Gambar 4. ROC Curve SVM

4. Hasil dan Pembahasan

Kedua algoritma memiliki hasil yang berbeda dan kurva *Receiver Operator Characteristic* (ROC) yang berbeda juga, hal seperti akurasi, presisi, recall, dan F1-score pada dataset juga mempengaruhi interpretasi kinerja model secara keseluruhan.

Kurva *Receiver Operator Characteristic* (ROC) berikut melibatkan evaluasi model Random Forest (RF) dengan *cross-validation*, pelatihan pada data *training*, evaluasi pada *holdout* dataset kemudian di visualisasikan dengan kurva *Receiver Operator Characteristic* (ROC). Hal ini terlihat pada Gambar 3.

Sama dengan visualisasi kurva di Random Forest (RF) untuk Support Vector Machine (SVM) ini menggambarkan kinerja model dalam membedakan antara kelas positif dan negatif, kemudian memberikan ukuran kinerja secara keseluruhan. Hal ini terlihat pada Gambar 4.

Hasil eksperimen dirangkum pada Tabel 1.

Tabel 1. Hasil Eksperimen

	Random Forest	**Support Vector Machine**
Precision	0.953	0.977
Recall	0.953	0.977
F1-Score	0.953	0.977
Accuracy	0.953	0.977
Time	0.81367 detik	0.09291 detik

5. Kesimpulan

Dalam penelitian ini, perbandingan efektivitas antara dua metode klasifikasi, yaitu Random Forest (RF) dan Support Vector Machine (SVM), terhadap dataset risiko kanker serviks telah dilakukan. Penelitian sebelumnya dalam domain ini telah mencoba berbagai algoritma, seperti Naïve Bayes, KNN, dan Support Vector Machine (SVM) [3], [4], [6]. Namun, fokus utama penelitian ini adalah pada dua algoritma Random Forest (RF) dan Support Vector Machine (SVM), dimana kedua algoritma tersebut termasuk metode yang sering digunakan dan efektif dalam pola klasifikasi [14]. Hasil eksperimen menunjukkan bahwa keduanya memberikan hasil yang berbeda, dengan kurva *Receiver Operator Characteristic* (ROC) yang berbeda pula. Melalui evaluasi kinerja, terlihat bahwa Support Vector Machine (SVM) memiliki nilai precision, recall, F1-score, dan akurasi yang tinggi. Sementara Random Forest (RF) juga menunjukkan kinerja yang baik, namun dengan waktu komputasi yang lebih lama dibandingkan dengan Support Vector Machine (SVM) [15].

Penelitian ini memberikan kontribusi pada pemahaman perbandingan performa antara Random Forest (RF) dan Support Vector Machine (SVM) dalam konteks penilaian risiko kanker serviks. Melalui hasil algoritma keduanya, dapat disimpulkan bahwa Support Vector Machine (SVM) dapat menjadi pilihan yang baik untuk tugas klasifikasi ini, memberikan prediksi yang akurat dan efisien.

Saran untuk penelitian selanjutnya adalah melakukan percobaan ulang dengan dataset yang lebih besar atau berbeda, serta mempertimbangkan penggunaan algoritma lain untuk memberikan hasil yang lebih komprehensif. Dengan demikian, penelitian ini dapat membuka peluang untuk pengembangan metode penilaian risiko kanker serviks yang lebih canggih dan optimal.

Daftar Pustaka

[1] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, “Cervical cancer: Epidemiology, risk factors and screening,” Chinese J. Cancer Res., vol. 32, no. 6, pp. 720–728, 2020, doi: 10.21147/j.issn.1000-9604.2020.06.05.

[2] Gokagglers, “Cervical Cancer Risk Classification,” www.kaggle.com. Accessed: Nov. 21, 2023. [Online]. Available: https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification/

[3] T. Praningki and I. Budi, “Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN,” Creat. Inf. Technol. J., vol. 4, no. 2, p. 83, 2018, doi: 10.24076/citec.2017v4i2.100.

[4] S. S. Arifin, A. M. Siregar, A. Ratna, and T. Al Mudzakir, “Klasifikasi Penyakit Kanker Serviks Menggunakan Algoritma Support Vector Machine (SVM),” 4th Conf. Innov. Appl. Sci. Technol. (CIASTECH 2021), no. 4, pp. 521–528, 2021.

[5] Z. Rustam, E. Sudarsono, and D. Sarwinda, “Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD),” in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052066.

[6] H. Syahputra and A. Wibowo, “Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites,” J. Ilm. Tek. Elektro Komput. dan Inform., vol. 9, no. 1, pp. 165–173, 2023, doi: 10.26555/jiteki.v9i1.25861.

[7] B. Falini, G. Martino, and S. Lazzi, “A comparison of the International Consensus and 5th World Health Organization classifications of mature B-cell lymphomas,” Leukemia, vol. 37, no. 1, pp. 18–34, 2023, doi: 10.1038/s41375-022-01764-1.

[8] Maryati, “Faktor-Faktor Yang Mempengaruhi Skrining Kanker Serviks Di Indonesia: Scoping Review,” J. Persat. Perawat Nas. Indones., vol. 8, no. 1, p. 12, 2023, doi: 10.32419/jppni.v8i1.404.

[9] P. Pangestu and R. Novita, “Systematic Literature Review : Perbandingan Algoritma Klasifikasi,” pp. 431–440, 2023.

[10] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, “Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja,” J. Inform. Univ. Pamulang, vol. 5, no. 4, p. 578, 2021, doi: 10.32493/informatika.v5i4.7575.

[11] R. I. Arumnisaa and A. W. Wijayanto, “SISTEMASI: Jurnal Sistem Informasi Perbandingan Metode Ensemble Learning: Random Forest, Support Vector Machine, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM) Comparison of Ensemble Learning

Method: Random Forest, Support Vector Machine, AdaB," Januari, vol. 12, no. 1, pp. 2540–9719, 2023, [Online]. vailable: http://sistemasi.ftik.unisi.ac.id

[12] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," J. Transform., vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.

[13] E. Suryati, A. Ari Aldino, N. Penulis Korespondensi, and E. Suryati Submitted, "Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," J. Teknol. dan Sist. Inf., vol. 4, no. 1, pp. 96–106, 2023, [Online]. Available: https://doi.org/10.33365/jtsi.v4i1.2445

[14] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," J. Media Inform. Budidarma, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[15] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," J. Inform. Upgris, vol. 7, no. 1, pp. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099