JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 1, JANUARI, 2026, Pp. 114 – 126
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Model Prediksi Risiko Kanker Serviks dengan Pendekatan Support Vector Machine

Juwita Stefany Hutapea [(1)*], Nisa Hanum Harani [(2)], Cahyo Prianto [(3)]
Departemen Teknik Informatika, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia
e-mail : juwitastefany13@gmail.com, {nisa,cahyo}@ulbi.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 5 Agustus 2025, direvisi 15 Januari 2026, diterima 15 Januari 2026, dan dipublikasikan 25 Januari 2026.

Abstract

Cervical cancer is one of the leading causes of death in women, especially in developing countries due to delays in early diagnosis. Developing a risk prediction model based on the Support Vector Machine (SVM) algorithm is one way to support a more accurate and efficient early detection process. The research object is medical records of female patients obtained from hospitals in Medan City, with a total of 164 patient data. The development process was carried out through the CRISP-DM stages, which include data cleaning, feature transformation, class balancing with SMOTE, and dimensionality reduction using PCA. The evaluation results showed that the best model was obtained with a PCA configuration with 9 principal components (90% variance) and a test size of 80:20, resulting in an accuracy of 88%, a precision of 88%, a recall of 84%, and an F1-score of 86%. Cross-validation evaluation with 5 folds provided the best average performance and the smallest standard deviation, indicating model stability. The final model was implemented in a web-based system to facilitate digital early detection. This study shows that SVM with the SMOTE and PCA approaches is effective in predicting cervical cancer risk accurately and efficiently.

Keywords: Cervical Cancer, Prediction, Support Vector Machine, SMOTE, PCA

Abstrak

Kanker serviks merupakan salah satu penyebab kematian tertinggi pada wanita terutama di negara berkembang akibat keterlambatan dalam diagnosis dini. Pengembangan model prediksi risiko berbasis algoritma *Support Vector Machine* (SVM) menjadi salah satu cara dalam mendukung proses deteksi dini yang lebih akurat dan efisien. Objek penelitian berupa data rekam medis pasien wanita yang diperoleh dari rumah sakit di Kota Medan, dengan total 164 data pasien. Proses pengembangan dilakukan melalui tahapan CRISP-DM yang mencakup pembersihan data, transformasi fitur, penyeimbangan kelas dengan SMOTE dan reduksi dimensi menggunakan PCA. Hasil evaluasi menunjukkan bahwa model terbaik diperoleh pada konfigurasi PCA dengan 9 komponen utama (90% variansi) dan test size 80:20 yang menghasilkan akurasi sebesar 88%, precision 88%, recall 84%, dan F1-score 86%. Evaluasi *cross validation* dengan fold 5 memberikan performa rata-rata terbaik dan standar deviasi terkecil, menandakan kestabilan model. Model akhir diimplementasikan dalam sistem berbasis web untuk mempermudah deteksi dini secara digital. Penelitian ini menunjukkan bahwa SVM dengan pendekatan SMOTE dan PCA efektif untuk memprediksi risiko kanker serviks secara akurat dan efisien.

Kata Kunci: Kanker Serviks, Prediksi, Support Vector Machine, SMOTE, PCA

1. PENDAHULUAN

Kanker serviks merupakan salah satu jenis kanker yang memiliki dampak signifikan secara global. Penyakit ini berkembang dari sel-sel abnormal pada leher rahim (serviks). Berdasarkan data dari *Global Cancer Observatory* (Globocan) yang dikeluarkan oleh *World Health Organization* (WHO) pada tahun 2020, terdapat 604.127 kasus dan 341.831 orang meninggal karena kanker serviks di seluruh dunia (Singh et al., 2023; Sung et al., 2021). Hampir 90% dari kasus kematian tersebut terjadi di negara-negara berkembang. Di Indonesia, kanker serviks

menjadi penyakit kedua yang paling sering terjadi pada wanita, dengan perkiraan 36.633 kasus baru dan 21.003 kematian setiap tahun (Indarti, 2023). Hal ini menunjukkan bahwa kanker serviks masih menjadi masalah kesehatan yang serius dan membutuhkan solusi tepat dalam medeteksi secara dini.

Salah satu hambatan dalam penanganan kanker serviks adalah keterlambatan dalam diagnosis yang menyebabkan angka kematian tinggi. Metode pemeriksaan konvensional seperti *Pap smear* dan inspeksi visual dengan asam asetat (IVA) sering dibatasi karena interpretasi hasil yang bersifat subjektif, ketersediaan tenaga medis yang terbatas, serta akses yang kurang merata di berbagai wilayah (Ekawati et al., 2024). Oleh karena itu, diperlukan pendekatan berbasis teknologi yang dapat mendukung sistem bantuan keputusan untuk membantu memprediksi kanker serviks lebih objektif, efisien, dan akurat.

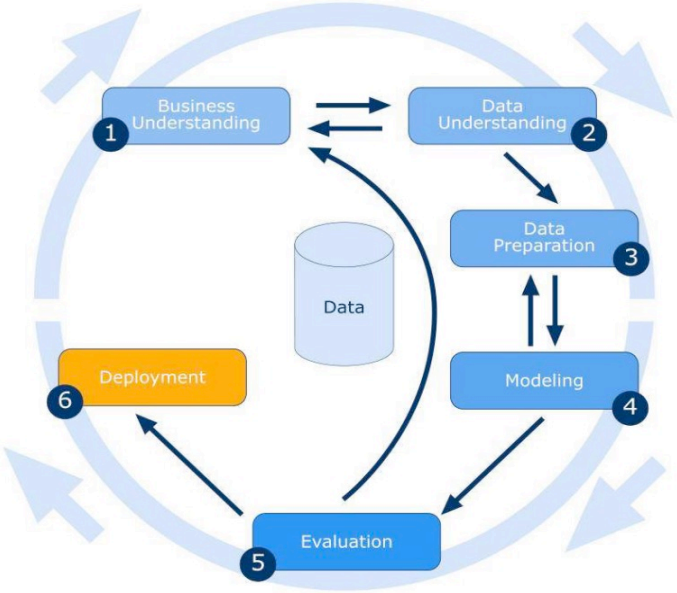
Perkembangan teknologi kecerdasan buatan terutama *machine learning* menjadi peluang baru dalam analisis data medis untuk memprediksi penyakit. Algoritma *Support Vector Machine* (SVM) dikenal sebagai metode klasifikasi yang baik dalam menangani data berdimensi tinggi dan pola yang tidak linear, sehingga cocok untuk analisis data rekam medis yang kompleks. Sejumlah penelitian terdahulu telah menunjukkan potensi besar machine learning dalam prediksi kanker serviks. Studi yang dilakukan oleh (Binanto et al., 2024) berfokus pada perbandingan efektivitas algoritma *Random Forest* (RF) dan *Support Vector Machine* (SVM) dalam menilai risiko kanker serviks. Penelitian ini menggunakan dataset dari Kaggle dan menyoroti tahapan preprocessing, pembagian data, serta evaluasi performa model menggunakan ROC curve dan metrik seperti akurasi, precision, recall, dan F1-score. Hasilnya menunjukkan bahwa RF mencapai akurasi sebesar 97,16% dengan AUC 0,996, sementara SVM hanya mencapai akurasi 96,01% dan AUC 0,543, menjadikan RF lebih unggul dalam konteks ini.

Penelitian oleh (Alsmariy et al., 2020) mengusulkan model prediksi kanker serviks dengan mengombinasikan metode voting classifier. Dataset UCI digunakan bersama teknik SMOTE untuk penyeimbangan kelas dan PCA untuk reduksi dimensi. Model ini menunjukkan peningkatan signifikan dalam akurasi dan sensitivitas (hingga lebih dari 5%) dibandingkan model-model sebelumnya. Penelitian ini menegaskan bahwa kombinasi SMOTE dan PCA efektif dalam meningkatkan kinerja klasifikasi dan efisiensi komputasi untuk prediksi kanker serviks. Berdasarkan hasil penelitian-penelitian sebelumnya menunjukkan bahwa SVM mampu memberikan hasil yang baik dalam klasifikasi berbagai penyakit, termasuk kanker serviks terutama ketiga digunakan dengan teknik pra-pemrosesan seperti SMOTE untuk menyeimbangkan data dan PCA untuk mengurangi dimensi.

Penelitian ini bertujuan untuk mengembangkan model prediksi risiko kanker serviks dengan menggunakan data rekam medis pasien kanker serviks di salah satu rumah sakit di Kota Medan dengan menggunakan algoritma *Support Vector Machine* dan penerapan teknik SMOTE serta PCA. Proses pengembangan model dijalani melalui tahapan CRISP-DM, yaitu pembersihan data, transformasi fitur, penyeimbangan kelas, pemodelan, dan evaluasi. Diharapkan model yang dihasilkan dapat menjadi dasar sistem bantuan keputusan untuk memprediksi risiko kanker serviks secara dini dan membantu meningkatkan kualitas layanan kesehatan wanita di Indonesia.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan CRISP-DM (*Cross Industry Standard Process for Data Mining*) sebagai kerangka kerja utama dalam pengembangan model prediksi risiko kanker serviks berbasis algoritma *Support Vector Machine* (SVM) (Pinto et al., 2020). Metodologi ini dipilih karena merupakan standar dalam data mining dan memiliki struktur yang sistematis dalam proses pengolahan data (Hasanah et al., 2021). Alur proses penelitian dapat dilihat pada Gambar 1. Secara umum, tahapan penelitian ini meliputi pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment.



Gambar 1 Diagram Alur CRISP-DM

2.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah salah satu algoritma machine learning berbasis *supervised learning* yang dirancang untuk menyelesaikan permasalahan klasifikasi dan regresi. Teknik ini bertujuan untuk mencari sebuah bidang pemisah (*hyperlane*) yang paling optimal dalam memisahkan dua kelas data. Prinsip utama dari SVM adalah memilih bidang pemisah yang memiliki margin terbesar antara dua kelompok data, di mana margin tersebut merupakan jarak antara *hyperlane* dengan titik-titik data terdekat dari masing-masing kelas yang dikenal sebagai support vectors (Jalil et al., 2024).

2.2 Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) adalah metode oversampling yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam kumpulan data (Fatmawati et al., 2025). Kumpulan data yang tidak seimbang, di mana satu kelas (kelas mayoritas) memiliki jumlah data yang jauh lebih banyak daripada kelas lainnya (kelas minoritas) dapat menghasilkan model klasifikasi yang bias dan berkinerja buruk dalam mengidentifikasi kelas minoritas (Sulistiyono et al., 2021). SMOTE bekerja dengan cara menghasilkan data sintetis baru dari kelas minoritas melalui interpolasi antar titik data yang saling berdekatan (nearest neighbors), bukan hanya menggandakan data yang sudah ada (Parman et al., 2024).

2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik statistik yang digunakan untuk mengurangi jumlah fitur atau variabel dalam sebuah dataset tanpa kehilangan informasi penting. Metode ini sangat bermanfaat ketika data memiliki banyak atribut yang saling berkorelasi sehingga dapat menyederhanakan kompleksitas model dan mempercepat proses pelatihan algoritma seperti *Support Vector Machine* (SVM) (Oktafiani & Itje Sela, 2024). PCA memiliki tujuan untuk menemukan pola dan struktur tersembunyi dalam data berdimensi tinggi dengan mentransformasikannya ke dalam ruang berdimensi lebih rendah, namun tetap mempertahankan informasi yang paling penting. Proses ini dilakukan dengan menghilangkan korelasi antar variabel independent melalui transformasi ke dalam variabel-variabel baru yang saling tidak berkorelasi yang dikenal sebagai principal components. Hasilnya adalah representasi data dalam dimensi yang lebih rendah namun tetap mempertahankan informasi paling penting dari data asli (Mulla et al., 2021).

2.4 Pemahaman Bisnis (*Business Understanding*)

Tahap pertama dalam penelitian ini adalah business understanding untuk mengetahui sebuah masalah (Cahyaningtyas et al., 2022). Tingginya angka kematian akibat kanker serviks disebabkan oleh keterlambatan dalam proses diagnosis dini. Metode skrining konvensional seperti *Pap smear* dan IVA memiliki keterbatasan dalam hal akurasi dan aksesibilitas. Penelitian ini bertujuan mengembangkan model prediktif berbasis *Support Vector Machine* (SVM) untuk mengidentifikasi pasien dengan risiko tinggi menggunakan data rekam medis. Model ini diharapkan dapat mendukung proses skrining yang lebih cepat, objektif, dan akurat dalam konteks deteksi dini kanker serviks.

2.5 Pemahaman Data (*Data Understanding*)

Setelah tujuan penelitian ditetapkan, tahap selanjutnya adalah memahami data yang akan digunakan. Tahap ini diawali dengan pengumpulan data yang menjadi dasar untuk analisis dan pengambilan kesimpulan (Hasanah et al., 2021). Dataset yang digunakan dalam penelitian ini berasal dari data rekam medis pasien di salah satu rumah sakit di Kota Medan dengan total 164 data yang berisi informasi klinis seperti usia, gejala utama, riwayat *pap smear*, dan lainnya yang disajikan pada Tabel 1.

Tabel 1 Dataset

No.	Atribut	Keterangan
1	Nomor_Rekam_Medis	Nomor identitas unik pasien di rumah sakit
2	Umur	Usia pasien
3	Status_Perkawinan	Status pernikahan pasien (contoh: Belum menikah, Menikah)
4	Gejala_Utama	Gejala utama yang dirasakan pasien saat pertama kali memeriksakan diri
5	Durasi_Gejala_Minggu	Lama gejala berlangsung dalam minggu
6	Riwayat_Pap_Smear	Riwayat pemeriksaan Pap Smear sebelumnya (Pernah/Tidak Pernah)
7	Riwayat_Kontrasepsi	Riwayat penggunaan kontrasepsi (Pernah/Tidak Pernah)
8	Jumlah_Kehamilan	Total jumlah kehamilan yang pernah dialami pasien
9	Tes_HPV	Riwayat melakukan tes HPV (Ya/Tidak)
10	Jumlah_Pasangan_Seksual	Jumlah pasangan seksual yang pernah dimiliki pasien
11	Risiko_Jatuh_1th	Risiko pasien mengalami jatuh dalam 1 tahun terakhir (Ya/Tidak)
12	Psikologis	Kondisi psikologis pasien (Cemas/Tenang)
13	Berat_Badan	Berat badan pasien dalam satuan kilogram (kg)
14	Tinggi_Badan	Tinggi badan pasien dalam satuan sentimeter (cm)
15	BMI	Indeks massa tubuh pasien (Body Mass Index)
16	Tekanan_Sistole	Tekanan darah sistolik pasien (mmHg)
17	Tekanan_Diastole	Tekanan darah diastolik pasien (mmHg)
18	Frekuensi_Nadi	Denyut nadi per menit (bpm - beats per minute)
19	Frekuensi_Nafas	Jumlah nafas per menit
20	Suhu_Tubuh	Suhu tubuh pasien dalam derajat Celcius
21	Skor_GCS	Skor Glasgow Coma Scale untuk menilai kesadaran pasien
22	Skor_Karnofsky	Skor untuk menilai kemampuan fungsional pasien
23	Tipe_Perawatan	Jenis perawatan yang diterima (Rawat Inap/Rawat Jalan)
24	Tahun	Tahun dilakukannya pemeriksaan atau pengambilan data
25	Hasil_Biopsi	Hasil pemeriksaan biopsi (Positif/Negatif untuk kanker serviks)

Tahap eksplorasi awal dilakukan untuk melihat struktur dan isi dataset, termasuk beberapa baris awal, tipe data, dan statistik deskriptif. Salah satunya dengan menggunakan fungsi *head* yang

menampilkan baris pertama data. Pada Gambar 2, menampilkan data rekam medis 5 pasien yang berisi informasi demografis (usia, status perkawinan), keluhan (gejala utama, durasi), riwayat kesehatan (*Pap Smear*, kontrasepsi, jumlah kehamilan), dan hasil tes HPV untuk setiap pasien.

	Nomor_Rekam_Medis	Umur	Status_Perkawinan	Gejala_Utama	Durasi_Gejala_Minggu	Riwayat_Pap_Smear	Riwayat_Kontrasepsi	Jumlah_Kehamilan	Tes_HPV
0	01.22.90.03	50	Menikah	Pendarahan	4	Tidak Pernah	Pernah	4	Ya
1	03.77.34.73	63	Menikah	Nyeri Perut	6	Pernah	Tidak Pernah	2	Tidak
2	01.22.85.54	48	Menikah	Pendarahan	2	Tidak Pernah	Tidak Pernah	1	Tidak
3	01.23.39.32	42	Menikah	Pendarahan	16	Tidak Pernah	Pernah	3	Ya
4	01.22.81.55	45	Menikah	Pendarahan	4	Tidak Pernah	Tidak Pernah	4	Ya
5	01.23.36.99	69	Menikah	Nyeri Perut	4	Tidak Pernah	Tidak Pernah	6	Ya

Gambar 2 Data Baris Pertama

Sebelum memilih fitur yang akan dipakai dalam pelatihan, dilakukan uji korelasi untuk memahami hubungan antar fitur dengan target (Hasil_Biopsi). Korelasi dihitung menggunakan korelasi *pearson* yang mengukur kekuatan dan arah hubungan linear antara dua variabel dengan nilai antara -1 hingga 1. Korelasi *pearson* memberikan gambaran tentang hubungan linear antar variabel yang berguna dalam memilih fitur yang relevan untuk model. Diagram batang yang memvisualisasikan korelasi antar berbagai fitur (variabel independent) terhadap variabel target yaitu Hasil_Biopsi ditunjukkan pada Gambar 3. mengurutkan berbagai faktor (fitur) berdasarkan kekuatan hubungannya (korelasi) dengan Hasil Biopsi. Batang yang mengarah ke kanan menunjukkan korelasi positif, sedangkan batang ke kiri menunjukkan korelasi negatif. Panjang batang merepresentasikan kekuatan hubungan, semakin panjang batangnya semakin kuat korelasinya.

Gambar 3 Korelasi Antar Fitur

Berdasarkan hasil uji korelasi terhadap target dalam Gambar 3, variabel Umur, Gejala_Utama, Durasi_Gejala_Utama, Riwayat_Pap_Smear, Riwayat_Kontrasepsi, Jumlah_Kehamilan, Tes_HPV, Jumlah_Pasangan_Seksual, Psikologis, Risiko_Jatuh_1th, dan Tahun dipilih karena menunjukkan hubungan yang signifikan, baik secara positif maupun negatif terhadap target. Misalnya, Tes_HPV dan Riwayat_Pap_Smear memiliki korelasi positif tinggi yang menunjukkan bahwa variabel ini berhubungan erat dengan Hasil_Biopsi yang positif. Jumlah_Kehamilan dan Gejala_Utama juga menunjukkan korelasi positif yang sedang sehingga relevan untuk dimasukkan dalam model prediksi. Sementara, Riwayat_Kontrasepsi memiliki korelasi negatif tinggi yang artinya kemungkinan untuk mendapatkan Hasil_Biopsi positif cenderung lebih rendah. Variabel lain seperti Umur, Tahun, dan Risiko_Jatuh_1th memiliki nilai korelasi yang lebih rendah, tetapi tetap dipertimbangkan karena masih memiliki sedikit pengaruh terhadap hasil akhir prediksi dan relevan secara klinis. Pemilihan variabel ini menunjukkan keseimbangan antara kekuatan korelasi dan relevansi kontekstual dalam prediksi risiko kanker serviks.

2.6 Persiapan Data (*Data Preparation*)

Tahap ini bertujuan untuk mengubah data mentah menjadi dataset yang bersih dan siap untuk diolah sebelum diterapkan ke dalam model pembelajaran mesin (Studer et al., 2021). Proses ini mencakup pembersihan, transformasi, dan penyusunan data agar siap digunakan dalam proses modeling. Persiapan data dilakukan untuk menghilangkan noise, mengisi nilai yang hilang, mengubah data kategorikal menjadi numerik, melakukan normalisasi, dan mengatasi ketidakseimbangan kelas agar model prediksi dapat dilatih dengan optimal.

Dataset yang digunakan terdiri atas data bertipe kategorikal dan numerik. Untuk memungkinkan pemrosesan oleh algoritma machine learning, data kategorikal perlu dikonversi terlebih dahulu ke dalam bentuk numerik melalui proses encoding, sehingga seluruh fitur dapat dikenali dan diolah secara optimal oleh model. Agar interpretasi data tetap jelas, dilakukan pencetakan hasil mapping dari nilai numerik hasil encoding ke label aslinya yang ditunjukkan dalam Gambar 4. Pada Gambar 4 menunjukkan proses mapping data, di mana setiap kategori data yang berupa teks diubah menjadi format angka untuk keperluan analisis oleh komputer. Sebagai contoh, pada kolom 'Hasil_Biopsi', nilai 'negatif' diubah menjadi 0 dan 'positif' diubah menjadi 1.

```
Mapping untuk kolom 'Status_Perkawinan':
  0 → Belum Menikah
  1 → Menikah

Mapping untuk kolom 'Gejala_Utama':
  0 → Keputihan
  1 → Nyeri Perut
  2 → Pendarahan

Mapping untuk kolom 'Riwayat_Pap_Smear':
  0 → Pernah
  1 → Tidak Pernah

Mapping untuk kolom 'Hasil_Biopsi':
  0 → negatif
  1 → positif

Mapping untuk kolom 'Riwayat_Kontrasepsi':
  0 → Pernah
  1 → Tidak Pernah
```

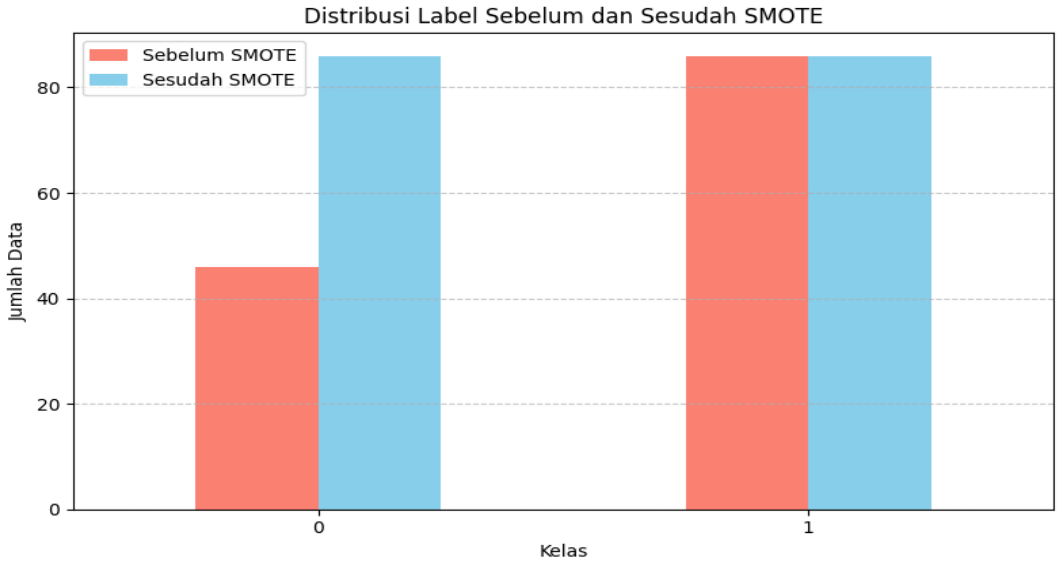
Gambar 4 Korelasi Fitur Terhadap Target

Setelah data dibersihkan dan dikonversi, langkah selanjutnya adalah memisahkan fitur (X) dan label (y) untuk keperluan pelatihan model pada tahap selanjutnya. Fitur yang digunakan meliputi variabel umur, gejala utama, durasi gejala dalam minggu, riwayat pap smear, riwayat penggunaan kontrasepsi, jumlah kehamilan, hasil tes HPV, jumlah pasangan seksual, kondisi psikologis, risiko jatuh dalam satu tahun terakhir, dan tahun pemeriksaan. Sementara itu, label yang digunakan sebagai target prediksi adalah hasil biopsi.

Data yang telah melalui tahap praproses dan dipisahkan antara fitur dan target selanjutnya dibagi menjadi dua bagian, yaitu 131 data latih (*training*) dan 33 data (*testing*). Pembagian ini bertujuan untuk memastikan bahwa model dilatih menggunakan sebagian data, sementara performanya diuji menggunakan data yang belum pernah dilihat sebelumnya. Dalam penelitian ini, digunakan beberapa variasi proporsi data pengujian (test size), yaitu 0.2, 0.3, dan 0.5 dari total 164 dataset untuk mengevaluasi konsistensi dan stabilitas performa model pada berbagai skenario pembagian data. Proporsi test size sebesar 0,2 menghasilkan 33 data uji dan 131 data latih, test size 0,3 menghasilkan 49 data uji dan 115 data latih, sedangkan test size 0,5 menghasilkan 82 data uji dan 82 data latih.

Setelah data dibagi menjadi data pelatihan dan pengujian, tahap selanjutnya adalah membangun pipeline untuk mendukung proses pelatihan model secara terstruktur dan sistematis. Penggunaan pipeline memungkinkan integrasi beberapa tahapan seperti normalisasi, penyeimbangan kelas, reduksi dimensi, dan pelatihan model secara berurutan dan konsisten setiap kali model dijalankan. Pendekatan ini tidak hanya meningkatkan efisiensi pemrosesan, tetapi juga meminimalkan potensi kesalahan akibat ketidakkonsistenan dalam penanganan data. Selain itu, pipeline memastikan bahwa seluruh transformasi hanya diterapkan pada data pelatihan dalam setiap lipatan validasi silang, sehingga mencegah terjadinya kebocoran data (*data leakage*) ke dalam proses pelatihan.

Pada tahap oversampling, dilakukan penanganan terhadap ketidakseimbangan kelas pada data target. SMOTE digunakan untuk mengatasi masalah ini dengan cara membuat sampel sintetik baru pada kelas minoritas sehingga distribusi antar kelas menjadi lebih seimbang. Hasil dari proses penyeimbangan data ditampilkan pada Gambar 5. Awalnya, dataset tidak seimbang dengan 86 data untuk kelas '1' dan hanya 46 data untuk kelas '0'. Setelah diterapkan SMOTE, jumlah data pada kelas minoritas '0' ditingkatkan sehingga kini kedua kelas menjadi seimbang dengan masing-masing memiliki 86 data.



Gambar 5 Penerapan SMOTE

Penerapan PCA dilakukan untuk mereduksi jumlah fitur menjadi sejumlah komponen utama yang mampu menjelaskan total varians data. Teknik ini digunakan untuk menyaring informasi paling relevan serta mengurangi noise atau redundansi yang tidak signifikan terhadap klasifikasi. Untuk menentukan jumlah komponen utama yang optimal tanpa mengorbankan terlalu banyak informasi penting, dilakukan pengujian terhadap 2 variasi nilai n_components, sehingga diperoleh konfigurasi terbaik dalam mereduksi dimensi sekaligus mempertahankan kinerja model secara maksimal.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 1, JANUARI, 2026: 114 – 126

2.7 Pemodelan (*Modeling*)

Pada tahap ini merupakan inti dari proses data mining, di mana dilakukan penerapan algoritma yang sesuai untuk membangun model prediktif berdasarkan data yang telah dipersiapkan (Schröer et al., 2021). Model yang digunakan dalam penelitian ini adalah *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF). SVM dipilih karena kemampuannya yang andal dalam menangani klasifikasi dengan dimensi tinggi dan data yang tidak terpisah secara linear. Selain itu, pendekatan ini dilengkapi dengan teknik *cross-validation* dan penyesuaian bobot kelas (*class weighting*) untuk meningkatkan generalisasi dan mengurangi bias akibat ketidakseimbangan data.

2.8 Evaluasi Model (*Evaluation*)

Setelah model dibangun, kinerjanya dievaluasi secara objektif menggunakan data uji yang belum pernah digunakan selama proses pelatihan. Evaluasi ini bertujuan untuk mengukur seberapa baik model dapat menggeneralisasi pengetahuannya pada data baru (Dzulhijjah et al., 2024). Kinerja model dievaluasi menggunakan beberapa metrik klasifikasi yaitu akurasi, precision, recall, F1-score, dan confusion matrix. Metrik recall menjadi perhatian utama untuk memastikan model mampu mendeteksi sebanyak mungkin kasus risiko tinggi kanker serviks.

Penilaian kinerja model selanjutnya dilakukan dengan beberapa metrik. Pertama, akurasi, yaitu metrik sederhana namun penting yang menunjukkan persentase prediksi model yang benar (baik positif maupun negatif) dari total keseluruhan data (Tangkelayuk & Mailoa, 2022). Akurasi sangat bermanfaat jika distribusi data antar kelas seimbang. Rumus akurasi dapat dihitung menggunakan Pers. (1). Selanjutnya, precision, yaitu metrik yang mengukur seberapa banyak dari prediksi untuk suatu kelas yang benar. Precision yang tinggi menunjukkan bahwa model menghasilkan sedikit kesalahan (*false positive*) (Admojo & Ahsanawati, 2020). Di mana rumus precision dapat dilihat pada Pers. (2). Metrik lainnya adalah recall, metrik ini mengukur kemampuan model untuk mendeteksi semua data yang sebenarnya termasuk dalam suatu kelas. Recall yang tinggi berarti model jarang melewatkan data yang penting (*false negative*) (Prasetyo et al., 2023). Persamaan untuk menghitung recall dapat dihitung menggunakan Pers. (3). Terakhir, F1-score, yaitu metrik kombinasi antara precision dan recall yang memberikan keseimbangan antara keduanya (Sylvia et al., 2024). F1-score berguna pada saat mempertimbangkan baik *false positives* maupun *false negatives*. Di mana rumus F1-score dapat dilihat pada Pers. (4).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$Presisi = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

2.9 Implementasi (*Deployment*)

Sebagai tahap terakhir dalam kerangka CRISP-DM adalah deployment. Dalam penelitian ini, tahap ini berfokus pada implementasi sistem (Triginandri & Subhiyakto, 2024). Dilakukan pembangunan sistem prototipe berbasis web menggunakan framework *Flask* dan *database* MySQL. Sistem ini memungkinkan pengguna menginputkan data dan memperoleh hasil prediksi risiko secara langsung, serta visualisasi hasil prediksi melalui antarmuka pengguna yang sederhana.

3. HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan model prediksi menggunakan algoritma *Support Vector Machine* (SVM) yang dibangun untuk memprediksi risiko kanker serviks berdasarkan data rekam medis. Model diuji pada berbagai scenario pengolahan data dan evaluasi untuk mendapatkan konfigurasi terbaik yang menghasilkan kinerja optimal yang ditunjukkan pada Tabel 2. Berdasarkan evaluasi awal terhadap variasi proporsi pembagian data pada Tabel 2, diperoleh hasil akurasi sebesar 0.88 untuk test_size 0.2, 0.82 untuk test_size 0.3, dan 0.85 untuk test_size 0.5. Proporsi 80:20 dipilih sebagai konfigurasi terbaik karena memberikan akurasi tertinggi sekaligus mempertahankan keseimbangan antara data pelatihan dan data pengujian.

Tabel 2 Akurasi Berdasarkan Test Size

Test_size	Hasil Akurasi
0.2	0.88
0.3	0.82
0.5	0.85

Pengujian terhadap nilai n_components pada PCA dilakukan dengan dua variasi nilai yang diuji untuk menentukan jumlah komponen utama yang dapat dipertahankan. Hasil pengujian menunjukkan bahwa pemilihan komponen yang sesuai mampu meningkatkan efisiensi tanpa menurunkan akurasi karena fitur-fitur yang kurang relevan telah dieliminasi. PCA berhasil membantu menyederhanakan kompleksitas data sekaligus mempertahankan informasi penting untuk prediksi.

Tabel 3 Evaluasi Model pada Berbagai Nilai PCA

PCA	Principal Component	Evaluasi						
		Accuracy	Precision		Recall		F1-Score	
			0	1	0	1	0	1
0.90	9	0.88	0.89	0.88	0.73	0.95	0.80	0.91
0.75	7	0.85	0.80	0.87	0.73	0.91	0.76	0.89

Berdasarkan hasil pengujian pada Tabel 3, n_components 0.90 dipilih karena mempertahankan 9 komponen dan memberikan hasil evaluasi terbaik secara keseluruhan. Akurasi model pada komponen ini mencapai 0.88 tertinggi dibandingkan komponen lain, selain itu nilai precision, recall dan F1-score untuk masing-masing kelas juga menunjukkan performa yang sangat baik dan seimbang. Sementara n_components 0.75 hasil evaluasinya lebih rendah dibandingkan dengan 0.90.

Tabel 4 Perbandingan Cross Validation

Test_size	Fold (cv)	Akurasi	Rata-Rata	Standar Deviasi
0.2	2	0.88	0.8106	0.0379
	3		0.7879	0.0214
	5		0.8023	0.0665
0.3	2	0.82	0.7564	0.0195
	3		0.8345	0.0264
	5		0.8348	0.0696
0.5	2	0.85	0.7561	0.0000
	3		0.7685	0.0330
	5		0.7691	0.0404

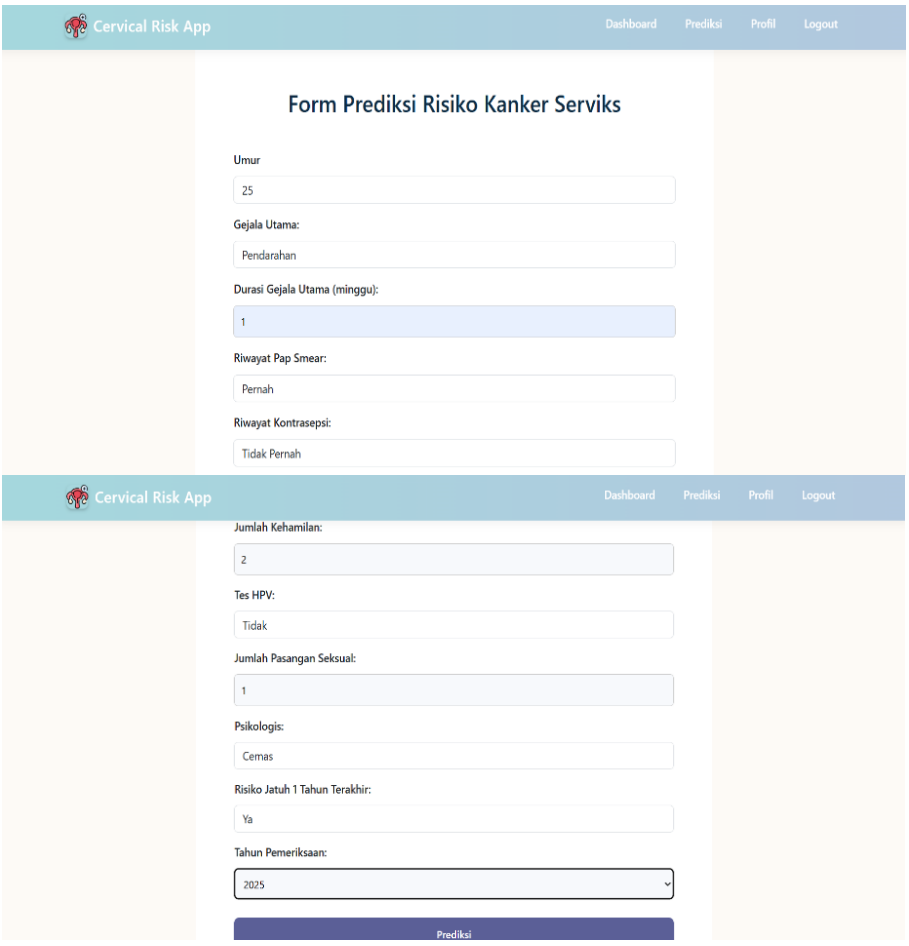
Untuk menguji konsistensi model, dilakukan evaluasi menggunakan teknik *cross validation* dengan nilai fold sebanyak 2, 3, dan 5 yang dikombinasikan dengan berbagai test size. Hasil evaluasi ditunjukkan dalam Tabel 4. Berdasarkan hasil evaluasi di Tabel 4, dapat dilihat bahwa penggunaan cv = 5 dengan test_size 0.2 menghasilkan performa yang paling baik dengan nilai

rata-rata yang diperoleh cukup baik yaitu sebesar 0.8023 dan standar deviasi yaitu 0.0665 dibandingkan dengan cv = 2 atau cv = 3 pada data test_size yang sama. Penggunaan 5 fold ini menunjukkan bahwa model cukup konsisten dalam melakukan prediksi pada data yang berbeda-beda.

Model akhir yang telah dibentuk melalui pipeline kemudian dievaluasi menggunakan *confusion matrix* untuk melihat kinerja klasifikasi secara rinci. Model menghasilkan akurasi sebesar 0.88, precision sebesar 0.88, recall sebesar 0.84, dan F1-score sebesar 0.86. Hasil evaluasi lengkap disajikan dalam Tabel 5. Model yang telah dibangun dan dianalisis hingga mencapai tingkat akurasi sebesar 88%, kemudian diimplementasikan dalam bentuk sistem sederhana berbasis web. Dashboard ini memungkinkan pengguna untuk menginputkan data dan memperoleh hasil prediksi risiko kanker serviks secara *real-time*, serta melihat visualisasi data. Implementasi ini bertujuan untuk mempermudah proses skrining awal secara digital, mempercepat pengambilan keputusan, dan meningkatkan efisiensi layanan deteksi dini. Tampilan dari dashboard dapat dilihat pada Gambar 6.

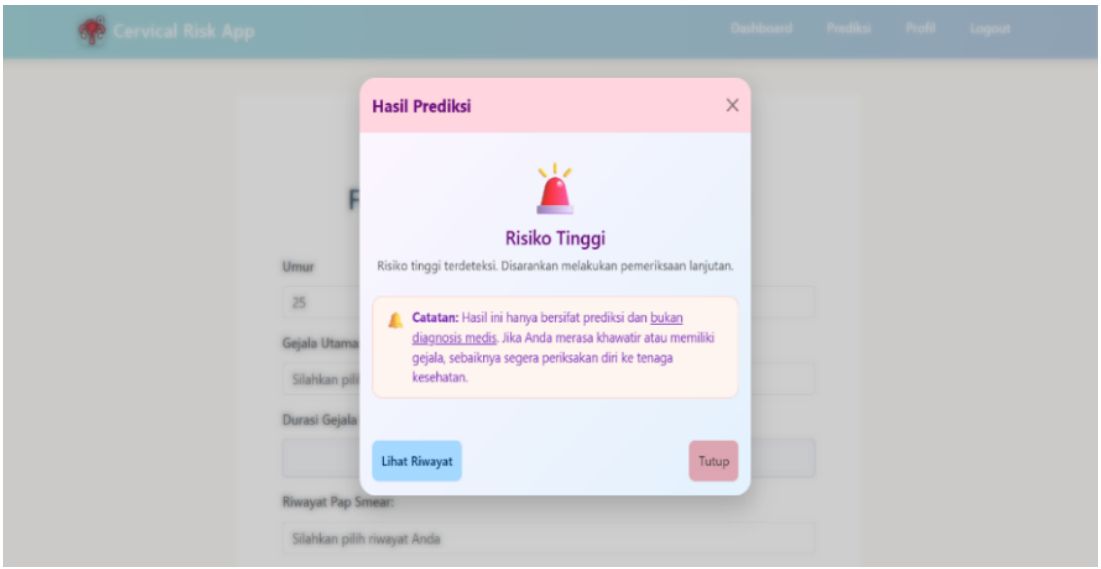
Tabel 5 Performa Kinerja Model

Kelas	Precision	Recall	F1-score	Support
0	0.89	0.73	0.80	11
1	0.88	0.95	0.91	22
Accuracy			**0.88**	**33**
Macro Avg	**0.88**	**0.84**	**0.86**	**33**
Weighted Avg	**0.88**	**0.88**	**0.88**	**33**



Gambar 6 Form Prediksi

Dapat dilihat Gambar 6 menampilkan antarmuka sistem yang dirancang untuk memprediksi risiko kanker serviks berdasarkan data yang diinputkan. Pengguna dapat memasukkan informasi seperti usia, gejala utama, durasi gejala, riwayat pap smear, penggunaan kontrasepsi, jumlah kehamilan, hasil tes HPV, jumlah pasangan seksual, kondisi psikologis, serta faktor risiko lainnya. Setelah data lengkap diinputkan, pengguna dapat menekan tombol Prediksi untuk memperoleh hasil prediksi risiko. Sistem akan menghasilkan output berupa prediksi risiko kanker serviks yang memetakan data pasien ke dalam kategori tinggi atau rendah, sehingga dapat membantu proses skrining awal secara efisien. Hasil visualisasi dari prediksi tersebut dapat dilihat pada Gambar 7 berikut ini.



Gambar 7 Hasil Prediksi

Model yang digunakan dalam penelitian ini adalah algoritma *Support Vector Machine* (SVM) yang telah dioptimalkan melalui tahapan preprocessing, balancing data, dan reduksi dimensi. Sebagaimana ditampilkan pada Gambar 7, sistem prediksi berbasis web yang dikembangkan menghasilkan visualisasi hasil prediksi risiko dalam bentuk modal pop-up yang interaktif. Dalam contoh ini, sistem mendeteksi bahwa pengguna memiliki risiko tinggi terhadap kanker serviks dan memberikan rekomendasi untuk melakukan pemeriksaan lanjutan. Visualisasi ini juga dilengkapi dengan catatan penting yang menekankan bahwa hasil prediksi bukan merupakan diagnosis medis, melainkan alat bantu skrining awal. Penyajian hasil prediksi dalam format visual ini tidak hanya mempermudah interpretasi bagi pengguna, tetapi juga mempercepat pengambilan keputusan terhadap langkah selanjutnya yang harus dilakukan. Implementasi antarmuka ini bertujuan untuk mendukung proses prediksi dini kanker serviks secara digital dan efisien.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan algoritma *Support Vector Machine* (SVM) dalam prediksi risiko kanker serviks mampu menghasilkan model yang akurat dan andal. Dengan bantuan teknik SMOTE untuk menyeimbangkan jumlah data antar kelas dan PCA untuk menyederhanakan jumlah fitur, model mampu mencapai akurasi sebesar 88%, dengan precision 0.88, recall 0.84, dan F1-score 0.86. Nilai tersebut menunjukkan bahwa model bekerja dengan baik dalam mengenali pasien yang berisiko tinggi. Uji validasi silang (*cross-validation*) dengan 5 fold dan pembagian data 80:20 juga membuktikan bahwa model cukup stabil dan konsisten. Hasil penelitian ini menunjukkan bahwa penggabungan metode pembelajaran mesin dengan teknik praproses yang tepat dapat meningkatkan akurasi prediksi pada data kesehatan. Sebagai rekomendasi, model ini dapat dikembangkan lebih lanjut dengan jumlah data yang lebih besar dan diintegrasikan dalam sistem digital di fasilitas kesehatan untuk mendukung proses skrining awal kanker serviks secara lebih luas.

DAFTAR PUSTAKA

Admojo, F. T., & Ahsanawati. (2020). Klasifikasi Aroma Alkohol Menggunakan Metode KNN. *Indonesian Journal of Data and Science*, *1*(2), 34–38. https://doi.org/10.33096/ijodas.v1i2.12

Alsmariy, R., Healy, G., & Abdelhafez, H. (2020). Predicting Cervical Cancer Using Machine Learning Methods. *International Journal of Advanced Computer Science and Applications*, *11*(7), 173–184. https://doi.org/10.14569/IJACSA.2020.0110723

Binanto, I., B, J. P. K., & Leokadja, L. (2024). Perbandingan Metode Klasifikasi Random Forest dan Support Vector Machine Terhadap Dataset Resiko Kanker Serviks. *JTRISTE*, *11*(1), 60–66. https://doi.org/10.55645/jtriste.v11i1.507

Cahyaningtyas, C., Manongga, D., & Sembiring, I. (2022). Algorithm Comparison and Feature Selection for Classification of Broiler Chicken Harvest. *Jurnal Teknik Informatika (Jutif)*, *3*(6), 1717–1727. https://doi.org/10.20884/1.jutif.2022.3.6.493

Dzulhijjah, D. A., Herlambang, M. B., & Haifan, M. (2024). Implementasi Framework CRISP-DM untuk Proses Data Mining Aplikasi Credit Scoring PT. XYZ. *Seminar Nasional Sains dan Teknologi "SainTek" Seri I*, *1*(1), 238–251. https://conference.ut.ac.id/index.php/saintek/article/view/2337

Ekawati, F. M., Listiani, P., Idaiani, S., Thobari, J. A., & Hafidz, F. (2024). Cervical Cancer Screening Program in Indonesia: Is It Time for HPV-DNA Tests? Results of a Qualitative Study Exploring the Stakeholders' Perspectives. *BMC Women's Health*, *24*(1), Article ID: 125. https://doi.org/10.1186/s12905-024-02946-y

Fatmawati, A., Latifah, U. B., S, A. S., & Zuhriya, T. K. (2025). Klasifikasi Emosi Teks Pengguna Twitter Menggunakan Metode SVM. *SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, *9*, 1999–2007. https://doi.org/10.29407/hq9jyy12

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Indarti, J. (2023). The Role of Social Obstetrics and Gynecology in the Coverage of Cervical Cancer Screening in the Era of Health Transformation in Indonesia. *Indonesian Journal of Obstetrics and Gynecology*, 198–200. https://doi.org/10.32771/inajog.v11i4.2181

Jalil, A., Homaidi, A., & Fatah, Z. (2024). Implementasi Algoritma Support Vector Machine untuk Klasifikasi Status Stunting pada Balita. *G-Tech: Jurnal Teknologi Terapan*, *8*(3), 2070–2079. https://doi.org/10.33379/gtech.v8i3.4811

Mulla, G. A. A., Demir, Y., & Hassan, M. (2021). Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *10*(3), 858–869. https://doi.org/10.17798/bitlisfen.939733

Oktafiani, R., & Itje Sela, E. (2024). Breast Cancer Classification with Principal Component Analysis and SMOTE Using Random Forest Method and Support Vector Machine. *Article in International Journal of Computer Applications*, *186*(16), 1–8. https://doi.org/10.5120/ijca2024923537

Parman, N. H., Hassan, R., & Zakaria, N. H. (2024). Breast Cancer Prediction Using Support Vector Machine Ensemble with PCA Feature Selection Method. *International Journal of Innovative Computing*, *14*(1), 15–19. https://doi.org/10.11113/ijic.v14n1.461

Pinto, A., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2020). Data Mining to Predict Early Stage Chronic Kidney Disease. *Procedia Computer Science*, *177*, 562–567. https://doi.org/10.1016/j.procs.2020.10.079

Prasetyo, S. D., Hilabi, S. S., & Nurapriani, F. (2023). Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN. *Jurnal KomtekInfo*, *10*, 1–7. https://doi.org/10.35134/komtekinfo.v10i1.330

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Singh, D., Vignat, J., Lorenzoni, V., Eslahi, M., Ginsburg, O., Lauby-Secretan, B., Arbyn, M., Basu, P., Bray, F., & Vaccarella, S. (2023). Global Estimates of Incidence and Mortality of

Cervical Cancer in 2020: A Baseline Analysis of the WHO Global Cervical Cancer Elimination Initiative. *The Lancet Global Health*, *11*(2), e197–e206. https://doi.org/10.1016/S2214-109X(22)00501-0

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392–413. https://doi.org/10.3390/make3020020

Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *SISTEMASI*, *10*(2), 445–459. https://doi.org/10.32520/stmsi.v10i2.1303

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Sylvia, P. H., Arifin, O., Arpan, A., Permata, R., Handoko, D., & Fitriyah. (2024). Evaluasi Kinerja Algoritma Naïve Bayes, KNN, dan SVM dalam Analisis Sentimen Media Sosial. *Jusikom: Jurnal Sistem Komputer Musi Rawas*, *9*(2), 157–166.

Tangkelayuk, A., & Mailoa, E. (2022). Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *9*(2), 1109–1119. https://doi.org/10.35957/jatisi.v9i2.2048

Triginandri, R., & Subhiyakto, E. R. (2024). Deteksi Dini Cacar Monyet menggunakan Convolutional Neural Network (CNN) dalam Aplikasi Mobile. *Edumatic: Jurnal Pendidikan Informatika*, *8*(2), 516–525. https://doi.org/10.29408/edumatic.v8i2.27625