Cervical Cancer Prediction by Classification using Support Vector Machines Algorithm

FINAL PROJECT

Student Identity:
Fathurrahman Nur Aziz (20.11.3694)

Course : Big Data & Data Mining
Lecturer : Anna Baita, M.Kom

Program Studi Informatika
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta

January, 2023

Abstraksi

Adanya peningkatan jumlah kasus kanker serviks di Indonesia berdasarkan data pada tahun 2020 menunjukkan perlunya upaya untuk menekan kenaikan melalui berbagai upaya pencegahan primer dan sekunder. Upaya primer yang dapat dilakukan di antaranya adalah seperti menanamkan pola hidup sehat serta melakukan vaksinasi HPV. Langkah ini tentunya perlu didukung dengan upaya pencegahan sekunder, yakni dengan melakukan skrining atau deteksi dini guna memastikan kesehatan leher rahim penduduk wanita Indonesia, sehingga pengembangan teknologi skrining perlu terus dilakukan demi menghasilkan teknologi skrining yang semutakhir mungkin. Pada penelitian ini, penulis berupaya meningkatkan *akurasi* dan mempercepat klasifikasi data kanker serviks dengan memisahkan dua kelas antara dominan negatif dan positif menggunakan algoritma *Support Vector Machine*. Hasil penelitian, implementasi, dan pengujian algoritma *Support Vector Machine* dalam melakukan klasifikasi untuk memprediksi kemungkinan terkena kanker serviks berdasarkan parameter diagnosis kanker serviks mengunjukkan prediksi hasil yang akurat dengan kesalahan minimal. Dari hasil pengujian algoritma *Support Vector Machine* menggunakan *confusion matrix* dan *classification report* didapatkan hasil klasifikasi dengan performa yang sangat bagus dengan nilai akurasi sebesar 99.00%, nilai presisi sebesar 99.00%, dan nilai *recall* sebesar 99.00%.

Kata Kunci : *kanker serviks, data mining, supervised learning, prediksi, klasifikasi, support vector machines*

Abstract

An elevation in Indonesia cervical cancer cases based on 2020 data shows the need to suppress its increase through various primary and secondary prevention efforts. Primary efforts that can be carried out include instilling a healthy lifestyle and taking the HPV vaccination. Of course, this step needs to be supported by secondary prevention efforts, like conducting screening or early detection to ensure the cervical health of the Indonesian female population, therefore the development of screening technology needs to be carried out continuously to make it as advanced as possible. In this study, the authors attempted to boost the accuration and accelerate the classification of cervical cancer data by separating the two classes between dominant negative and positive using the Support Vector Machine algorithm. The research, implementation, and testing of the Support Vector Machine algorithm in performing classification to predict the likelihood of developing cervical cancer based on cervical cancer diagnosis parameters show accurate prediction results with minimal error. From the testing of the Support Vector Machine algorithm using the confusion matrix and classification report, a very good result is obtained with an accurate value of 99.00%, a precision value of 99.00%, and a recall value of 99.00%.

Keywords: *cervical cancer, data mining, supervised learning, prediction, classification, support vector machines*

Bab I
Pendahuluan

1.1. Latar Belakang

Kasus kanker serviks atau kanker leher rahim masih menempati posisi teratas sebagai faktor utama penyebab kematian kanker pada penduduk usia produktif negara berkembang, tak terkecuali Indonesia[11]. Di Indonesia sendiri, kemunculan kasus kanker serviks mengalami peningkatan dari total 32.469 kasus pada tahun 2019 menjadi 36.633 kasus pada 2020[13,14].

Peningkatan ini mengindikasikan perlunya upaya lebih untuk menekan angka tersebut, baik melalui upaya primer dengan menjaga kesehatan dan melakukan vaksinasi HPV serta melakukan skrining sejak dini sebagai upaya pencegahan sekunder, terutama pada kelompok Wanita Usia Subur. Tahapan skrining ini sangat penting dilakukan agar dapat mendeteksi sedini mungkin risiko keberadaan kanker serviks guna memaksimalkan tingkat kesembuhan pasien. Demikian pula sebaliknya, deteksi kanker serviks yang terlampau terlambat dapat menurunkan tingkat kesembuhan secara signifikan.

Fakta tersebut kemudian mendorong penulis untuk mendukung pengembangan instrumen deteksi dan prediksi kanker serviks dengan mencoba mempercepat proses dan meningkatkan akurasi diagnosis kanker serviks dengan teknik *data mining*, tepatnya dengan mempercepat klasifikasi data dalam proses prediksi tersebut. Adapun dalam penelitian ini, klasifikasi dilakukan dengan memisahkan dua kelas antara dominan negatif dan positif menggunakan algoritma *Support Vector Machine*. Dataset yang diambil kemudian dilakukan *preprocessing* dengan cara *cleaning data* untuk kemudian memilih fitur yang relevan dan target yang menjadi label klasifikasi. Setelah itu, dilakukan proses *training*, *testing*, dan terakhir evaluasi hasil.

1.2. Tujuan Penelitian

Penelitian ini bertujuan untuk membantu meningkatkan akurasi dan mempercepat proses diagnosis kanker serviks dengan melakukan prediksi kemungkinan terkena kanker serviks berdasarkan parameter diagnosis kanker serviks.

1.3. Metode Penelitian

Penelitian ini dilakukan menggunakan metode *data mining* untuk mencari pengetahuan dari suatu kumpulan data (*Knowledge Discovery)*. Metode *data mining* yang diterapkan pada penelitian ini adalah *Supervised Learning* berupa klasifikasi menggunakan algoritma *Support Vector Machines* (SVM). Dengan algoritma tersebut, klasifikasi diharapkan memiliki nilai akurasi dan presisi yang tinggi sehingga tidak terjadi kesalahan diagnosis pada kasus kanker serviks.

Bab II
Tinjauan Pustaka

2.1. Kanker Serviks

Merujuk pada pemikiran Evriarti & Yasmon[12], kanker serviks terjadi akibat adanya pertumbuhan tak wajar dari jaringan epitel serviks atau leher rahim sehingga berdampak pada terjadinya infeksi yang persisten *human papillomavirus* (HPV) tipe *high risk* (HRHPV) *onkogenik* dan menyebabkan terbentuknya tumor ganas. Kanker serviks di Indonesia menjadi kasus kanker dengan jumlah penderita terbesar kedua, diperkirakan hingga 207 kasus per 100.000 populasi, sehingga pengembangan sistem diagnosis kanker serviks harus terus dilakukan.

2.2. *Supervised Learning*

Supervised Learning adalah algoritma dalam *machine learning* yang secara khusus digunakan untuk klasifikasi[1]. Alur kerja dari *supervised learning* adalah melakukan *pelabelan* pada suatu kumpulan data yang menjadi masukan (input) untuk dipetakan ke suatu kelas atau luaran (output) yang diinginkan[2]. Metode ini didasarkan pada suatu kumpulan data yang memiliki label yang akan dijadikan sebuah karakteristik distribusi perilaku suatu benda sehingga akan membentuk model perilaku suatu benda tersebut[3].

2.3. Klasifikasi

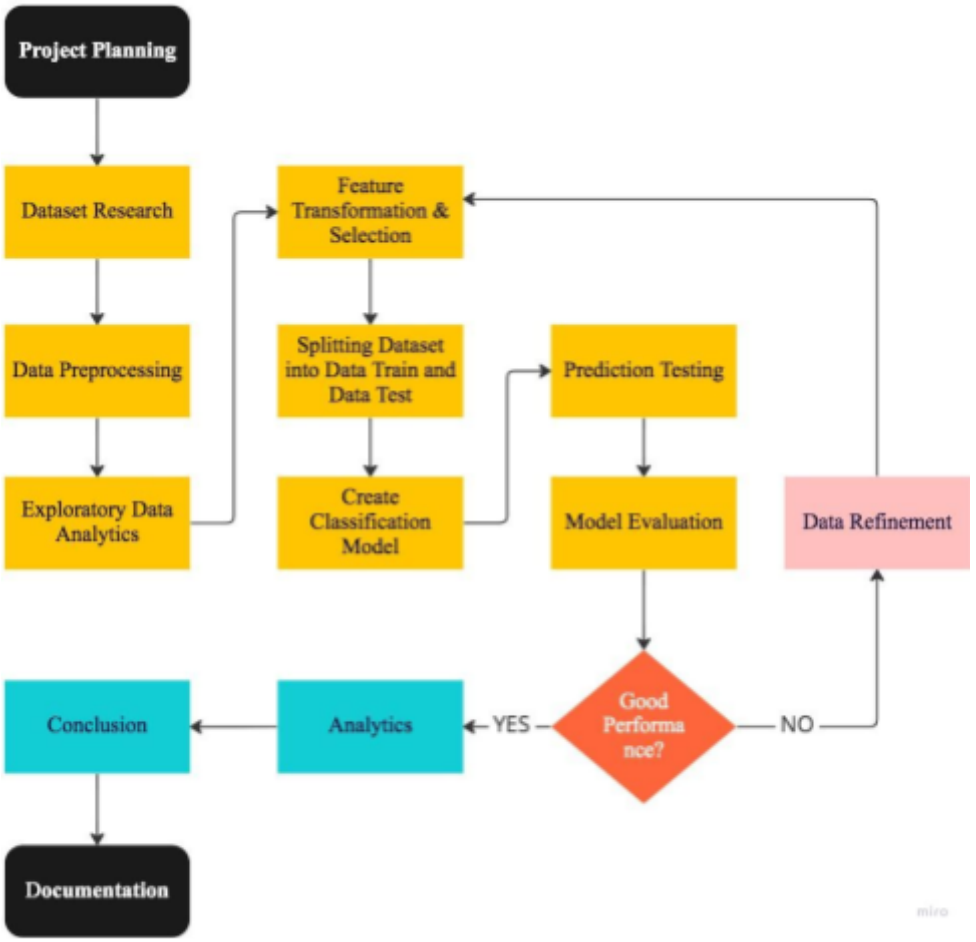
Klasifikasi merupakan cabang ilmu dari *machine learning*. Klasifikasi adalah rentetan proses menemukan fungsi dan perilaku suatu model yang dapat membedakan atau menjelaskan kelas data dengan tujuan memprediksi kelas yang tidak dikenal[4].

2.4. Support Vector Machines

SVM (*Support Vector Machines)* adalah teknik dalam data mining untuk melakukan klasifikasi dengan memisahkan dua sampel data dari dua kelas yang berbeda[5]. Support Vector Machines sangat dikenal karena mampu menghasilkan klasifikasi dengan baik meskipun *data train* yang digunakan lebih sedikit[6],[7]. Meskipun demikian, akurasi dari algoritma SVM sangat bergantung dari kernel dan parameter yang digunakan[8]. Konsep dari algoritma ini adalah mencari *hyperplane* terbaik dengan batas atau margin yang maksimal[9]. Dari margin tersebut, terbentuklah pola-pola terdekat terhadap masing-masing kelas yang disebut *support vector*[10].

Bab III
Metodologi Penelitian

3.1. Alur Penelitian



Gambar 1. Flowchart Penelitian

3.2. Spesifikasi Perangkat

- *Hardware* : Acer E5-475G, Intel i5-7200u, 16GB RAM, 2TB SSD
- *Software* : Google Colaboratory
- *Library* : *pandas, numpy, seaborn, matplotlib.pyplot, sklearn.metrics.*, sklearn.svm, sklearn.model_selection.train_test_split, sklearn.preprocessing.StandardScaler*

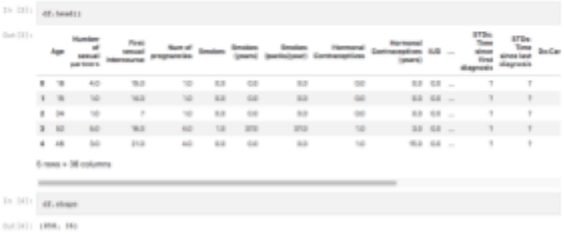
3.3. Alat dan Bahan

- *Dataset:* *https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification*
- *Notebook:* *https://github.com/azizfath/bddm3/blob/main/uas/20_11_3694_UAS.ipynb*

Bab IV
Hasil dan Pembahasan

Dataset

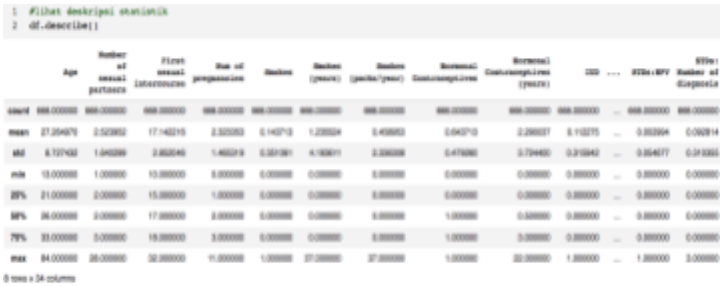
Dataset yang diambil dari kaggle memiliki jumlah kasus sebanyak 858 baris dan 36 fitur seperti yang terlihat pada gambar 2 di bawah ini:



Gambar 2. Dataset dari Kaggle

Data Preprocessing

Dari dataset tersebut, masih terdapat data yang bernilai "?" sehingga akan mengacaukan proses klasifikasi. Untuk mengatasinya, penulis mengubah data tersebut menjadi *null values*. Terdapat pula dua fitur yang memiliki terlalu banyak *null values*, yaitu *"STDs: Time since first diagnosis"* dan *"STDs: Time since last diagnosis"* sehingga perlu di-*drop*. Deskripsi statistik dari dataset tersebut terdapat pada gambar 3:



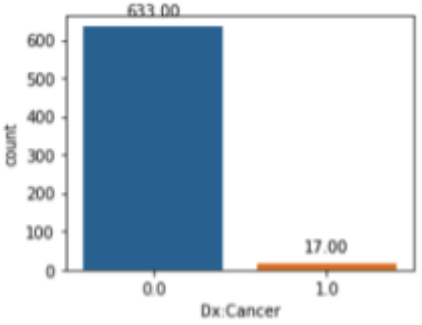
Gambar 3. Deskripsi statistik dari dataset

Dari deskripsi statistik tersebut, terdapat beberapa poin yang dapat diambil, yakni:

- Nilai maksimum dari *"Age"* adalah 84, dimana nilai ini terlalu besar jika dibandingkan dengan nilai maksimum pada kolom lain. Sehingga perlu dilakukan standarisasi skala seluruh kolom agar tidak mengganggu proses klasifikasi.
- Nilai maksimal dari "*Num of pregnancies"* adalah 11, dimana rata-ratanya hanya 2. Maka nilai 11 tersebut termasuk *outlier* yang dapat mengganggu performa klasifikasi.
- Kolom *"STDs:cervical condylomatosis"* dan *"STDs:AIDS"* hanya berisi nilai 0 yang membuat fitur tersebut tidak memiliki arti sehingga perlu dihilangkan. Kolom *"Dx:CIN"*, *"Dx:HPV"* tidak berhubungan dengan diagnosis kanker serviks dan kolom *"Dx"* hanya merupakan jumlah diagnosis sehingga tidak berkaitan dan perlu dihilangkan.

Exploratory Data Analysis

Exploratory Data Analysis yang dilakukan adalah melihat jumlah kasus negatif dan positif dari kolom diagnosis seperti pada gambar 4 di samping.



Gambar 4. Count dari Dx:Cancer

Feature Selection

Pada tahap ini, penulis menentukan variabel dependen (*features)* adalah seluruh fitur selain kolom *"Dx:Cancer"* dan variabel independen (target) adalah kolom *"Dx:Cancer"*.

Classification

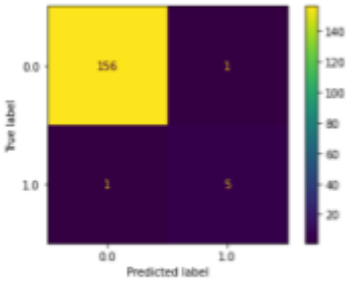
Klasifikasi dilakukan menggunakan algoritma *Support Vector Machines*. Sebelum melakukan klasifikasi penulis melakukan *split data* menjadi data *training* dan data *test* dengan persentase perbandingan 75%:25% dari total keseluruhan data. Selanjutnya penulis membuat model klasifikasi SVM dengan kernel *linear* dan melakukan *fitting* data. Kemudian, didapatkan hasil percobaan prediksi seperti pada gambar 5 berikut ini:

index	cancer
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0
7	0.0
8	0.0
9	0.0

Gambar 5. Hasil prediksi 15 kolom pertama data test

Model Evaluation

Confusion Matrix pada hasil klasifikasi terdapat pada gambar 6:



Gambar 6. Confusion Matrix

Capaian kinerja model pada hasil klasifikasi terdapat pada gambar 7:

Classification Report for SVM Classification

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	157
1.0	0.83	0.83	0.83	6
accuracy			0.99	163
macro avg	0.91	0.91	0.91	163
weighted avg	0.99	0.99	0.99	163

Gambar 7. Classification Report

Analytics

Berdasarkan ***Confusion Matrix***, maka klasifikasi tersebut menghasilkan ***True Negative*: 156 ; *True Positive*: 5; *False Positive*: 1; *False Negative*: 1**. Dengan demikian, dapat dikatakan bahwa klasifikasi tersebut dapat benar-benar melakukan prediksi positif ataupun negatif karena nilai *true*-nya tinggi dan nilai *false*-nya sangat rendah.

Berdasarkan ***Classification Report,*** maka klasifikasi tersebut menghasilkan:

- Akurasi : 99%
- Nilai Presisi Negatif : 99% ; Nilai Presisi Positif : 83%
- Nilai Recall Negatif : 99% ; Nilai Recall Positif : 83%

Berdasarkan nilai presisi di atas, dapat disimpulkan Algoritma SVM telah bagus dalam melakukan klasifikasi dataset tersebut karena banyak *True Positif* dari prediksi yang dilakukan (99% dan 83%). Berdasarkan nilai *Recall*, maka Algoritma SVM telah bagus dalam melakukan klasifikasi dataset tersebut karena banyak True Positif yang sesuai dengan *class* sebenarnya (99% dan 83%) Di samping itu, nilai akurasi yang cukup tinggi (99%) mengindikasikan bahwa klasifikasi dataset dengan SVM tersebut telah menghasilkan prediksi *class* yang baik.

Performa dan kinerja model dari klasifikasi yang dibuat sudah sangat baik sehingga tidak perlu dilakukan perbaikan data untuk meningkatkan performa.

Bab V
Kesimpulan

Berdasarkan penelitian, implementasi, dan pengujian algoritma *Support Vector Machine* dalam melakukan klasifikasi untuk memprediksi kemungkinan terkena kanker serviks berdasarkan parameter diagnosis kanker serviks, dapat diambil kesimpulan sebagai berikut: Penerapan algoritma *Support Vector Machine* dapat menghasilkan hasil prediksi yang akurat serta kesalahan yang minimal untuk mengetahui potensi terkena kanker serviks. Dari hasil pengujian algoritma *Support Vector Machine* menggunakan *confusion matrix* dan *classification report* didapatkan hasil klasifikasi dengan performa yang sangat bagus dengan nilai **akurasi** sebesar **99.00%**, nilai **presisi** sebesar **99.00%** dan nilai ***recall*** sebesar 99.00%.

Referensi

[1] Kotsiantis SB. 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica.

[2] Ayodele, TO. 2010. New Advances in Machine Learning, Yagang Zhang (Ed). London: IntechOpen Limited.

[3] Amei, W., Huailin, D., Qingfeng, W., & Ling, L. (2011). A survey of application-level protocol identification based on machine learning. 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, 3, 201–204.

[4] Nugroho, A., dan Subanar. Klasifikasi Naïve Bayes untuk Prediksi Kelahiran pada Data Ibu Hamil. Berkala MIPA. Vol. 23, No. 3, halaman 297-308, September 2013.

[5] Vapnik, V dan Cortes, C. 1995. Support Vector Networks. Machine Learning, 20, 273-297.

[6] C. Huang, L.S. Davis, dan J.R.G. Townshend, "An Assessment of Support Vector Machines for Land Cover Classification," Int. J. Remote Sens., Vol. 23, No. 4, hal. 725-749, 2002.

[7] A.T. Azar dan S.A. El-Said, "Performance Analysis of Support Vector Machines Classifiers in Breast Cancer Mammography Recognition," Neural Comput. Appl., Vol. 24, No. 5, hal. 1163-1177, 2014.

[8] Santosa, Budi. (2007). Tutorial Support Vector Machine. ITS, Surabaya

[9] C. Pitoy, "Metode Support Vector Machines pada Klasterisasi K-Means Data Nonlinear Separable," Frontiers: Jurnal Sains Dan Teknologi., Vol. 2, No. 1, hal. 71–77, 2010.

[10] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, dan V. Vapnik, "Feature Selection for SVMs," Proc. Advances in Neural Information Processing Systems, 2001, 668-674.

[11] Pratiwi Kurniasari and Fitriana Yuni, "Early Marriage Increase The Risk of Cervical Cancer Events," J. Ilmu Kebidanan, vol. 9, pp. 69–78, 2021.

[12] P. R. Evriarti and A. Yasmon, "Patogenesis Human Papillomavirus (HPV) pada Kanker Serviks," J. Biotek Medisiana Indones., vol. 8, no. 1, pp. 23–32, 2019, doi: 10.22435/jbmi.v8i1.2580.

[13] H. Widowati, "Kasus Kanker Payudara Paling Banyak Terjadi di Indonesia," Katadata, 2019. https://databoks.katadata.co.id/datapublish/2019/06/03/kasus-kanker-payudara-paling-banyak-terjadi-di-indonesia.

[14] M. A. Rizaty, "Ini Jenis Kanker yang Banyak Menyerang Perempuan Indonesia," Katadata, 2022. https://databoks.katadata.co.id/datapublish/2022/04/21/ini-jenis-kanker-yang-banyak-menyerang-perempuan-indonesia.