Urecol Journal. Part E: Engineering
Vol. 1 No. 1 (2021) pp. 1-6
eISSN: 2797-0418

Published by:
KONSORSIUM LPPM PTMA
KOORDINATOR WILAYAH JATENG & DIY

Application of the C4.5 Algorithm for Early Cervical Cancer Classification

Taftazani Ghazi Pratama[1], **Achmad Ridwan**[2], **Agung Prihandono**[3]
[1] Department of Computer Science, Universitas Muhammadiyah Kudus, Indonesia
[2] Department of Information System, Universitas Muhammadiyah Kudus, Indonesia
[3] Department of Computer Science, Universitas Muhammadiyah Kudus, Indonesia
taftazanighazi@umkudus.ac.id

https://doi.org/10.53017/uje.4

Received: 10/02/2021 Revised: 25/02/2021 Accepted: 28/02/2021

Abstract

Cervical cancer is one of the cancers that is of global concern because of the high mortality rate in women. Preventive measures such as early detection are needed so that patients can get treatment more quickly. Fortunately, this disease can be prevented with the role of technology to help doctors in early detection of various types of cancer. The technology developed by the researchers is using machine learning algorithms. Therefore, in this study using the C4.5 algorithm to classify cervical cancer. This algorithm aims to classify 2 classes: people who have cervical cancer, people who are healthy. The results of the experiment obtained from the C4.5 algorithm are getting an accuracy of 98.61%, precision of 98.08%, and recall of 95.24% ROC curve shows 0.982%.

Keywords: Cervical cancer; Algorithm C4.5; Machine learning

Penerapan Algoritma C4.5 untuk Klasifikasi Kanker Serviks Tingkat Awal

Abstrak

Kanker serviks merupakan salah satu kanker yang menjadi perhatian dunia karena tingginya angka kematian pada para perempuan. Langkah pencegahan seperti pendeteksian dini sangat dibutuhkan agar pasien dapat mendapatkan penanganan lebih cepat. Untungnya penyakit ini dapat dicegah dengan adanya peran dari teknologi sehingga membantu para dokter dalam melakukan pendeteksian dini berbagai macam penyakit kanker. Teknologi yang dikembangkan para peneliti yaitu menggunakan algoritma machine learning. Oleh karena itu, pada penelitian ini menggunakan algoritma C4.5 untuk melakukan klasifikasi kanker serviks. Algoritma ini bertujuan untuk mengklasifikasikan 2 kelas: orang yang menderita kanker serviks, orang yang sehat. Hasil dari eksperimen yang diperoleh dari algoritma C4.5 yaitu mendapatkan akurasi sebesar 98,61%, Precision sebesar 98,08%, dan Recall sebesar 95,24% kurva ROC menunjukkan angka 0,982%.

Kata-kata kunci: Kanker serviks; Algoritma C4.5; Machine learning

1. Pendahuluan

Salah satu jenis kanker ganas yang menyerang kaum wanita adalah kanker serviks. Berdasarkan data yang diperoleh dari Kemenkes pada tahun 2019, terdapat kasus kanker serviks sebesar 23,4 per 100.000 penduduk dengan rata-rata kematian 13,9 per 100.000 penduduk. Dengan tingginya kanker serviks di Indonesia, WHO menempatkan Indonesia

sebagai negara yang memiliki jumlah penderita kanker serviks terbanyak di dunia. Ironisnya, 80% penderita kanker serviks sudah masuk dalam stadium lanjut dan 94% pasien dari kasus tersebut meninggal dalam 2 tahun [1].

Jumlah kematian yang seringkali meningkat disebabkan oleh adanya keterlambatan diagnosis dan pemeriksaan penyakit tersebut. Diagnosis kanker serviks terlambat karena gejala yang dialami tidak kasat mata dan baru terasa saat sudah masuk stadium akhir [2]. Oleh karena itu dibutuhkan pendeteksian dini untuk memantau gejala yang dirasa dan sebagai langkah pemeriksaan awal. Pendeteksian kanker serviks umumnya dapat dilakukan melalui pemeriksaan laboratorium menggunakan metode Inspeksi Visual Asam Asetat (IVA) [2]. Namun dengan metode tersebut membutuhkan tenaga medis yang berkompeten dan beberapa pertimbangan fitur untuk mendapatkan hasil diagnosis yang akurat [2]. Oleh karena itu perlu adanya metode pendeteksian dini yang secara cepat dan lebih akurat. Dewasa ini perkembangan teknologi sudah semakin berkembang khususnya di bidang kesehatan. Salah satu teknologi yang digunakan untuk keperluan pendeteksian dini penyakit adalah membuat model pendeteksian dini berbasis komputer. Model tersebut dibuat dengan cara menerapkan algoritme machine learning untuk mempelajari kasus-kasus yang sudah ada melalui training dan kemudian diuji pada kasus-kasus baru. Beberapa penelitian yang mengklasifikasikan atau deteksi dini beberapa penyakit kanker dapat memperoleh akurasi mencapai rata-rata 90% [2]–[4]. Oleh karena itu, penelitian ini bertujuan untuk mengkasifikasikan atau mendeteksi awal individu apakah mengidap kanker serviks atau sehat menggunakan algoritma machine learning yaitu: C4.5.

2. Metode

Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Tujuan dari penelitian ini adalah melakukan klasifikasi dan evaluasi model algoritma C4.5 untuk mengetahui akurasi algoritma C4.5 dalam mengklasifikasikan penyakit kanker serviks.

2.1. Sumber data

Sumber data yang digunakan pada penelitian ini diperoleh dari UCI Machine Learning Repository. Dataset yang digunakan adalah Cervical Cancer Behavior Risk Data Set tahun 2019. Variabel yang digunakan pada penelitian ini adalah sebanyak 20 variabel dengan jumlah data sebanyak 72. Dataset ini memuat data screening kanker serviks berdasarkan perilaku dan determinannya. Metode pengumpulan data ini dibuat dari Sampel dan Kuisioner. Kuisioner kemudian disebarkan kepada 72 responden, terdiri dari: 22 perempuan penderita kanker serviks (30,56%) dan 50 perempuan bukan penderita kanker serviks (69,44%) [5].

2.2. Algoritma C4.5

Algoritma C4.5 adalah salah satu metode untuk membuat decision tree berdasarkan training data yang telah disediakan. Algoritma C4.5 merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada algoritma C4.5 adalah dapat mengatasi missing value, dapat menangani continues data, dan pruning. Algoritma C4.5 merupakan salah satu algoritma top 10 yang sering digunakan untuk klasifikasi [6] tetapi algoritma C4.5 belum bisa menangani ketidak seimbangan kelas [7]. Rumus untuk menghitung entropy dan gain disajikan pada persamaan (1) dan (2) secara berurutan.

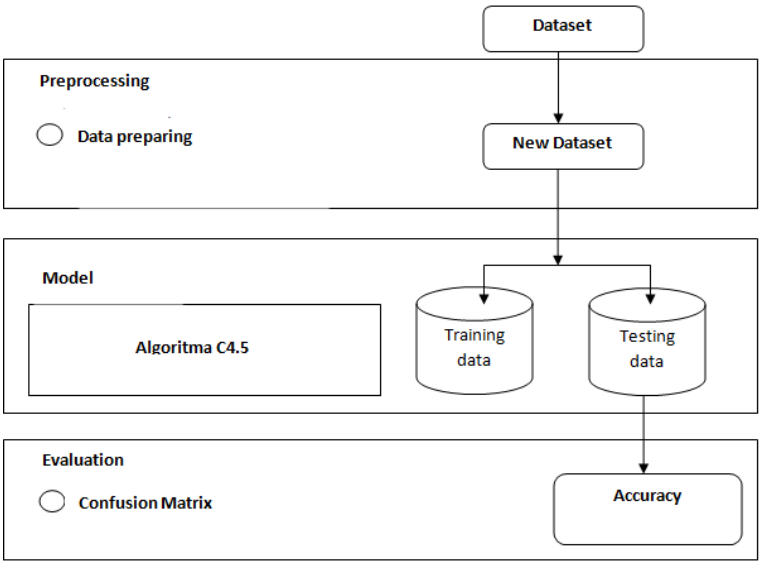
$$Entropy\ (S) = \sum_{i=1}^{n} - \text{ph} * \log 2\ \text{ph} \quad (1)$$

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

Dimana:
S = Himpunan Kasus
A = Atribut
n = jumlah partisi atribut A
|Si| = jumlah kasus pada partisi ke-i
|S| = jumlah kasus dalam S
Ph = Proporsi dari Si terhadap S

2.3. Alur penelitian

Alur penelitian yang dilakukan pada penelitian ini dapat lihat pada Gambar 1 di bawah ini.



Gambar 1. Alur penelitian

Penjelasan Gambar 1:

1. Preprocessing

Ruang lingkup awal ini dilakukan dengan cara menangani nilai yang hilang mengikuti teknik mengabaikan tupel dengan nilai yang telah dipraproses.

2. Model

Pada bagian kedua ini, Tool RapidMiner diterapkan pada Dataset yang baru untuk dibagi menjadi 2 bagian yaitu Training dan Testing pada Algoritma C4.5.

3. Evaluasi

Pada bagian ketiga ini, Dataset penentu perilaku diuji dengan Confusion Matrix serta diukur tingkat akurasinya. Confusion Matrix berisi informasi tentang aktual (actual) dan prediksi (predicted) pada sistem klasifikasi. Kinerja sistem seperti ini biasanya dievaluasi dengan menggunakan data pada matriks. Perhitungan dimasukan kedalam tabel Confusion Matrix [8]. Akurasi adalah perbandingan jumlah prediksi yang benar [7]. Semua ditentukan dengan mengimplementasikan formula sebagaimana diberikan pada Persamaan (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Sensitivity disebut juga dengan recall. Jika sensitivity 100% sama artinya dengan pengklasifikasian menganggap kasus yang diamati positif, yang dihitung dengan Persamaan (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Precision adalah tingkat positif salah adalah perbandingan nilai positif yang salah diklasifikasikan pada kasus negatif, yang perhitungannya menggunakan Persamaan (5).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Kurva ROC digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk menggambarkan pengklasifikasian berdasarkan kinerja algoritma [9], hasil AUC dapat dibagi menjadi beberapa kelompok [7]:

1. 0,90 - 1,00 = Klasifikasi Sangat Baik
2. 0,80 - 0,90 = Klasifikasi Baik
3. 0,70 - 0,80 = Klasifikasi Sedang
4. 0,60 - 0,70 = Klasifikasi Buruk
5. 0,50 - 0,60 = Kegagalan

3. Hasil dan Pembahasan

Eksperimen ini dilakukan menggunakan platform Komputasi: Intel Core i3-6006U @ 2.0 GHz 2.0 GHz CPU, 4 GB RAM, Sistem Operasi Microsoft Windows 10 64-bit, Rapidminer Studio Community versi 7.1.0013 sebagai analisis data. Variabel dataset Cervical Cancer Behavior Risk. Dataset yang digunakan pada penelitian ini disajikan pada Tabel 1.

Tabel 1 Deskripsi variabel dataset

No.	Atribut	Tipe
1	behavior_eating	Integer
2	behavior_personalHygine	Integer
3	intention_aggregation	Integer
4	intention_commitment	Integer
5	attitude_consistency	Integer
6	attitude_spontaneity	Integer
7	norm_significantPerson	Integer
8	norm_fulfillment	Integer
9	perception_vulnerability	Integer
10	perception_severity	Integer
11	motivation_strength	Integer
12	motivation_willingness	Integer
13	socialSupport_emotionality	Integer
14	socialSupport_appreciation	Integer
15	socialSupport_instrumental	Integer
16	empowerment_knowledge	Integer
17	empowerment_abilities	Integer
18	empowerment_desires	Integer
19	behaviour_sexualRisk	Integer
20	ca_cervix	Positive (1), Negative (2).

Tabel 1 menunjukkan 19 atribut dataset penentu perilaku dan 1 Atribut class penentu klasifikasi.

3.1. Preprocessing

Dari dataset Cervical Cancer Behavior Risk Data Set sebagai Data Training dan Data Testing yang ada di klasifikasikan oleh algoritma C4.5.

3.2. Model

Pada bagian model ini Data Training dan Data Testing Cervical Cancer Behavior Risk Dataset akan diproses klasifikasinya menggunakan aplikasi Rapidminer. Adapun hasil dari Confusion Matrix disajikan pada Tabel 2.

Tabel 2 Hasil Class Recall dan Precision

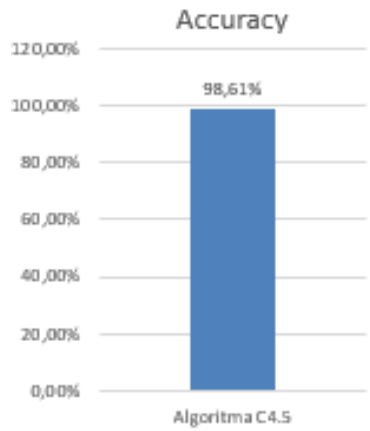
	true range1 [-∞ - 0.500]	true range2 [0.500 - ∞]	class precision
pred.range1 [-∞ - 0.500]	51	1	98.08%
pred.range2 [0.500 - ∞]	0	20	100.00%
class recall	100.00%	95.24%	

Dari Tabel 2 didapatkan Class Precision = 98.08%, dan Class Recall: 95.24%. Kinerja Class Recall lebih rendah dari precision dikarenakan adanya kelas yang tidak seimbang antara class positif dengan kelas negative.

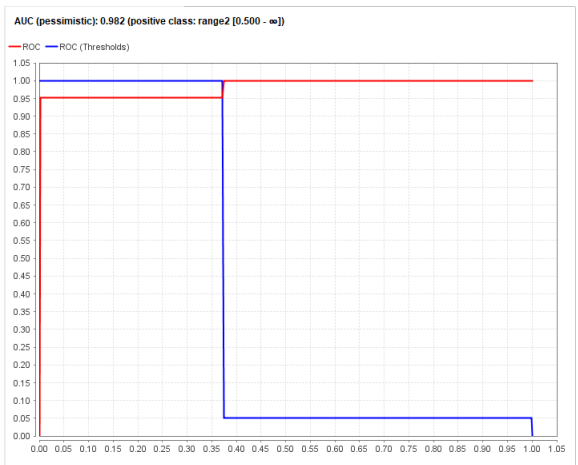
3.3. Evaluasi

Pada bagian Evaluasi ini dari Evaluasi pengklasifikasian dengan Algoritma C4.5 pada Data Training dan Data Testing Cervical Cancer Behavior Risk Dataset menghasilkan seperti Gambar 2. Dari Gambar 2 didapatkan akurasi untuk mengklasifisikan Data Training dan Data Testing Cervical Cancer Behavior Risk Data Set sebesar 98,61%. Adapun hasil dari pengujian ini juga menghasilkan Gambar 3.

Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma C4.5 memiliki nilai AUC sebesar 0,982 yang artinya Excellent Classification (Sangat Bagus) Dalam hasil penelitian, menunjukkan algoritma C4. 5 memberikan akurasi yang bagus untuk Klasifikasi Kanker Serviks pada UCI dataset Machine Learning Repository.



Gambar 2. Akurasi Algoritma C4.5



Gambar 3. Kurva ROC Model C4.5

4. Kesimpulan

Pada Penelitian ini menggunakan Algoritma C4.5 untuk klasifikasi Cervical Cancer Behavior Risk Dataset.Hasil eksperimen menunjukkan akurasi yang dihasilkan untuk mengklasifisikan Dataset Cervical Cancer Behavior Risk Data Set sebesar 98,61%, Class Precision sebesar 98.08%, dan Class Recall sebesar 95.24%. Kurva ROC (Receiver Operating

Characteristic) menunjukkan algoritma C4.5 dengan nilai AUC sebesar 0,982 yang artinya Excellent Classification (sangat bagus). Dengan hasil tersebut diharapkan pada penelitian selanjutnya melakukan kinerja khususnya kinerja recall dengan menerapkan algoritma yang lain sehingga dapat membantu dokter dan tenaga medis dalam pendeteksian awal kanker serviks.

Referensi

[1] "Bagaimana HPV di DIY?" .

[2] U. R. Hidayah, I. Cholissodin, and P. P. Adikara, "Klasifikasi Penyakit Kanker Serviks dengan Extreme Learning Machine," vol. 3, no. 7, pp. 6575–6582, 2019.

[3] D. A. Dharmawan, "Deteksi Kanker Serviks Otomatis Berbasis Jaringan Saraf Tiruan LVQ dan DCT," vol. 03, no. 04, pp. 3–6.

[4] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," Jurnal Infortech, vol. 2, no. 1, pp. 96–101, 2020, doi: 10.31294/infortech.v2i1.8079.

[5] Sobar, R. Machmud, and A. Wijaya, "Behavior determinant based cervical cancer early detection with machine learning algorithm," Advanced Science Letters, vol. 22, no. 10, 2016, doi: 10.1166/asl.2016.7980.

[6] B. Boukenze et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, 2012, doi: 10.1017/S0269888910000032.

[7] A. Ridwan, P. N. Andono, and C. Supriyanto, "Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma Naive," Teknologi Informasi, 2018.

[8] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, 2019, doi: 10.1016/j.patcog.2019.02.023.

[9] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?," Emergency Medicine Journal, 2017, doi: 10.1136/emermed-2017-206735.