Plagiarism Checker X Originality Report

Similarity Found: 8%

Date: Friday, April 14, 2023
Statistics: 138 words Plagiarized / 1752 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

IMPLEMENTASI ALGORITMA C.45 DALAM DIAGNOSA PENYAKIT KANKER PAYUDARA Yulia Eka Praptiningsih1, Winda Widya Ariestya* 2, Ida Astuti 3, Sylvia Nurulita 4 Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma1,2,3,4 Jl. Margonda Raya No. 100 Depok Yulia_eka@staff.gunadarma.ac.id1, winda_widya@staff.gunadarma.ac.id2, astuti@staff.gunadarma.ac.id3 , sylvia.nurulita@gmail.com4 (received: dd-mm-yy, revised: dd-mm-yy, accepted: dd-mm-yy (diisi oleh editor)) Abstract Breast cancer is a condition that affects a large number of Indonesian women.

Because this is a fatal disease, predicting whether someone has malignant or benign breast cancer is critical in order to prevent mortality. This work made use of 570 clinical data records from the UCI Machine Learning Repository. Algorithm C4.5 is used to differentiate between malignant and benign breast cancer patients. Eight determinants were obtained from the classification results on 32 characteristics. The performance evaluation findings demonstrate that the C4.5 algorithm can be utilized to classify breast cancer because the accuracy values obtained are quite high, namely 93.04%, 80.00% precision, and 92.31% recall. Keyword: Data Mining, Classification, C4.5, Breast Cancer Abstrak Penyakit kanker payudara merupakan jenis penyakit yang banyak diderita wanita Indonesia.

Penyakit ini termasuk penyakit yang mematikan, oleh karena itu prediksi seseorag menderita penyakit kanker payudara ganas atau jinak sangat diperlukan, guna menurunkan angka kematian. Dataset yang digunakan pada penelitian sebanyak 570 record data klinis yang berasal dari UCI Machine Learning Repository. Algoritma C4.5 diimplementasikan sebagai metode klasifikasi penderita kanker payudara ganas dan jinak. Dari hasil klasifikasi pada 32 atribut diperoleh 8 atribut sebagai penentu. Hasil

evaluasi performance menunjukan algoritma C4.5

dapat digunakan sebagai algoritma pada klasifikasi penyakit kanker payudara karena nilai akurasi yang diperoleh cukup besar yaitu 93,04%, presisi 80,00% dan recall 92,31%.
Kata Kunci: Data Mining, Klasifikasi, C4.5, Kanker Payudara

Pendahuluan WHO (World Health Organization) mengungkapkan bahwa kanker merupakan sekelompok besar penyakit yang dapat dimulai di hampir semua organ atau jaringan tubuh ketika sel-sel abnormal tumbuh tidak terkendali. Menurut data WHO di tahun 2018, kanker menjadi penyebab utama kematian kedua secara global yaitu sekitar 9,6 juta kematian atau dengan kata lain sebanyak satu dari enam kematian [1].

Kanker paru-paru, prostat, kolorektal, perut, dan hati adalah jenis kanker yang paling umum pada pria, sementara payudara, payudara, kanker kolorektal, paru-paru, serviks, dan tiroid adalah yang paling umum di antara Wanita [1]. Menurut laporan Global Burden of Cancer Study (Globocan) dari World Health Organization (WHO) bahwa pada tahun 2020, terdapat 396.914 penyakit kanker yang menyerang penduduk Indonesia [2]. / Gambar 1. Data Penyakit Kanker di Indonesia Tahun 2020 Sumber: https://gco.iarc.fr/ Berdasarkan jenis penyakitnya, kanker payudara paling banyak dialami di Indonesia yaitu sebanyak 65.858 kasus.

Jumlah ini setara 16,6% dari total kasus penyakit kanker di Indonesia [2]. Guna menyikapi masalah tersebut, diperlukan adanya pendeteksian sejak dini penyakit kanker payudara. Deteksi dini kanker payudara berkontribusi terhadap penurunan kematian akibat kanker payudara dan dapat mengurangi efek biaya yang besar. [3]. Seiring perkembangan teknologi khususnya teknologi informasi, terdapat berbagai teknologi yang dapat dilakukan untuk mengatasi berbagai kesulitan dalam prediksi penyakit kanker.

Salah satu teknologi yang dapat digunakan untuk prediksi dan klasifikasi penyakit adalah Data mining. Data Mining atau penambangan data yang dikenal dengan KDD (Knowledge Discovery in Database) adalah proses penggalian suatu pengetahuan berharga dari data [4]. Dengan kata lain Data merupakan adalah metodologi yang digunakan dalam penemuan yang efektif dan efisien pengetahuan yang ada dalam kumpulan data berskala besar [5].

Salah satu metode Data mining yang dapat digunakan untuk klasifikasi yaitu metode Decision tree dengan algoritma C4.5. Decision tree merupakan salah satu metode klasifikasi yang ebrsifat praktis sehingga menjadi algoritma yang paling populer dan banyak digunakan [6]. Algoritma C4.5 merupakan algoritma yang memiliki kemampuan meniru cara berpikir manusia saat pengambilan keputusan, hal ini berdasarkan hasil penelitian sebelumnya yang membandingkan algoritma decision tree yaitu C&RT, CHAID, QUEST, C4.5

dan ID3 menyatakan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih baik dari metode decision tree lainnya [7]. Beberapa penelitian yang telah dilakukan sebelumnya

digunakan sebagai rujukan pada penelitian ini. Pada suatu penelitian, metode Naïve Bayes berhasil melakukan klasifikasi jenis kanker berdasarkan struktur protein dengan akurasi yang diperoleh sebesar 79,17% [8]. Pada penelitian lain, metode Support Vector Machine (SVM) diimplementasikan dalam deteksi kanker paru dengan hasil akurasi sebesar 87,10% [9]. Pada penelitian lain, metode Decision Tree dengan algoritma C4.5

berhasil mencari rule pada diagnosa stadium kanker serviks dengan hasil akurasi sebesar 85.5% [10]. Penelitian lain metode Decision Tree dengan algoritma C4.5 juga diterapkan pada deteksi penyakit kanker serviks dengan akurasi yang dihasilkan sebesar 98.61% [11]. Berdasarkan uraian penelitian tersebut, metode Decision Tree dengan algoritma C4.5 cukup baik dalam melakukan klasifikasi prediksi penyakit kanker dengan menghasilkan akurasi yang baik, maka pada penelitian ini dilakukan implementasi metode Decision tree dengan algoritma C4.5

untuk melakukan prediksi pada penyakit kanker payudara dan mengetahui atribut penentu dari penderita kanker payudara jinak atau ganas. Metodologi Penelitian Penelitian ini menggunakan metode eksperimen dalam penyelesaian permasalahan. Gambar 2 menjelaskan tahapan dari penelitian yang terdiri dari tiga tahapan dalam penlitian ini, yaitu preprocessing, modeling dan evaluasi. / Gambar 2 Tahapan Penelitian Sumber: Data Diolah (2023) Tahap Preprocessing Sebelum data diolah, terlebih dahulu data dipersiapkan, proses ini dikenal dengan preprocessing [12]. Tahap ini meliputi proses penyiapan dataset dan pembagian data (split data).

Dataset penelitian ini menggunakan data dari database UCI Machine Learning Repository [13]. Dataset berisi 570 record data klinis. / / / Gambar 3. Dataset Penelitian Sumber: UCI Machine Learning Repository (2022) Gambar 3 menggambarkan dataset yang digunakan pada penelitian ini, sebanyak 32 atribut digunakan dan setiap atribut terdapat 3 jenis pengukuran yaitu rata-rata (mean), galat standar (standard error) dan terburuk (worst). Atribut diagnosis dijadikan sebagai atribut penentu apakah seseorang menderita kanker payudara ganas atau jinak. Split data dilakukan terhadap dataset untuk membagi menjadi data training dan data testing.

Prosentase yang digunakan dalam pembagian dataset sebanyak 80% data training yang bertindak sebagai pembentuj pola atau model dan 20% data testing sebagai penguji dari model. Tahap Modelling Penelitian ini menggunakan metode klasifikasi data mining dengan algoritma C4.5. Tahap untuk membuat pohon keputusan dengan algoritma C4.5 secara garis besar [14] yaitu: Tahap awal membentuk pohon keputusan adalah dengan membentuk akar (berada paling atas). Selanjtunya pembagian data sesuai dengan atribut-atribut yang serasi untuk dibentuk daun.

Tahap selanjutnya adalah pemangkasan pohon yang sudah terbentuk atau disebut juga tree pruning, pengenalan dan memangkas cabang yang tidak digunakan pada pohon yang sudah terbentuk. Selain untuk mengurangi ukuran pohon, pemangkasan pohon juga dilakukan dengan tujuan untuk memengkas tingkat eror pada prediksi kasus baru dari hasil pemecahan yang dilakukan dengan divide and conquer. Dua pendekatan dalam Pruning yaitu Pre-pruning dan Post-pruning. Tahap berikutnya dalah membentuk aturan keputusan dari pohon yang telah terbentuk. Aturan diturunkan dari pohon keputusan dengan cara menelusuri mulai dari akar hingga ke daun.

Rapidminer dimanfaatkan sebagai tools dalam mengimplementasikan algoritma C4.5. Rapidminer merupakan aplikasi yang menyediakan prosedur preprocessing, modelling sampai dengan visualisasi pada data mining dan machine learning [15]. Tahap Evaluasi Metode confusion matrix digunakan sebagai metode evaluasi untuk menghitung nilai akurasi penelitian ini. Confusion Matrix dapat mengevaluasi performance algoritma dari Machine Learning (ML), Confusion Matrix merepresentasikan prediksi dan kondisi sebenarnya (aktual) dari data yang dihasilkan oleh algoritma ML [16].

Terdapat tiga performance metrics popular yang digunakan dalam mengukur performance pada Confusion Matrix, yaitu: akurasi, precission, dan recall [17] Akurasi, merupakan rasio prediksi dengan benar (positif dan negatif) dengan keseluruhan data. Persamaan yang digunakan dalam menentukan nilai akurasi adalah: ??????????????= tp+tn tp+tn+fp+fn (1) Precission,merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Persamaan yang digunakan dalam menentukan nilai presisi adalah: ??????????????= tp tp+fp (2) Recall (Sensitifitas), merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif dan salah negatif.

Persamaan yang digunakan dalam menentukan nilai recall adalah: ????????????= tp fn+tp (3) Hasil dan Pembahasan Dari dataset sebanyak 570 record, dibagi menjadi dua bagian. Data training sebanyak 80% yaitu sebanyak 455 record dan data testing sebanyak 20% yaitu sebanyak 105 record. Proses klasifikasi menggunakan algoritma C4.5 dilakukan dengan bantuan Rapidminer seperti pada Gambar 4. / Gambar 4. Proses Klasifikasi Menggunakan Rapidminer Sumber: Hasil Pengolahan (2023) Selanjutnya setelah proses klasifikasi akan menghasilkan pohon keputusan. Pohon keputusan ini digunakan sebagai rule dalam klasifikasi prediksi penyakit kanker payudara.

/ Gambar 5. Pohon Keputusan Sumber: Hasil Pengolahan (2023) Berdasarkan gambar 5, terlihat bahwa dari 32 atribut pada dataset, ada 8 atribut yang memiliki penentu seseorang masuk kriteria penderita penyakit kanker payudara ganas atau jinak, yaitu Id (nomor identitas pasien), Perimeter_worst (keliling), Concavity_worst (tingkat keparahan

dari kontur), Concave points_worst (jumlah konkav), Smoothness_worst (variasi lokal dari nilai radius), Radius_worst (jarak rata-rata dari titik pusat ke tepi), M (diagnosis kanker payudara ganas) dan B (diagnosis kanker payudara jinak). Atribut diagnosis (M/B) dijadikan sebagai label penentu yang menentukan hasil akhir penderita kanker payudara gabas atau jinak.

M mewakili prediksi seseorang menderita kanker payudara ganas, sedangkan B mewakili prediksi seseorang menderita kanker payudara jinak. / Gambar 6. Rules yang Dihasilkan Sumber: Hasil Pengolahan (2023) Terdapat dua rules yang dihasilkan pada gambar 6 yaitu nilai perimeter_worst >114.450 dan nilai perimeter_worst <=114.450 Nilai perimeter_worst >114.450 Jika nilai perimeter_worst >114.450 dan concavity_worst > 0.191. Terdeteksi sebanyak 147 penderita penyakit kanker payudara ganas dan 1 penderita kanker payudara jinak. Sedangkan jika nilai concavity_worst <= 0.191 pada id > 4799740 terdapat 2 penderita kanker payudara jinak.

Untuk id <4799740 terdapat 2 penderita kanker payudara. Nilai parameter_worst <=114.450 Jika nilai perimeter_worst <=114.450 dan concavity_worst >0.176. Terdeteksi sebanyak 12 penderita penyakit kanker payudara ganas. Jika nilai perimeter_worst <=114.450, Concavity_worst <=0.176, smoothness_worst >0.173, terdeteksi 5 penderita penyakit kanker payudara ganas. Jika nilai perimeter_worst <=114.450 Concavity_worst <=0.176, smoothness_worst <=0.173, radius_worstnya > 17.615, terdapat 2 penderita penyakit kanker payudara ganas. Sedangkan jika nilai perimeter_worst <=114.450 Concavity_worst <=0.176, smoothness_worst <=0.173, radius_worstnya <= 17.615, terdapat 18 penderita penyakit kanker payudara ganas dan 265 penyakit kanker payudara jinak. / Gambar 7. Hasil Evaluasi Sumber: Hasil Pengolahan (2023) Tahap terakhir yaitu dilakukan evaluasi dengan melihat nilai performance yang dihasilkan.

Gambar 7 menggambarkan hasil akurasi, class precision dan class recall dari model yang dibentuk. Nilai akurasi yang dihasilkan sebesar 93,04%, Precision sebesar 80,00% dan recall sebesar 92,31%. Kesimpulan Penelitian ini mengimplementasikan algoritma C4.5 pada klasifikasi penderita penyakit kanker payudara. Dataset yang digunakan sebanyak 570 record dibagi menjadi 455 record data training dan data testing 105 record. Dari 32 atribut yang digunakan pada data set diperoleh 8 atribut yang memiliki penentu seseorang masuk kriteria penderita penyakit kanker payudara ganas atau jinak, yaitu atribut nomor identitas pasien, keliling, tingkat keparahan dari kontur, jumlah konkav, variasi lokal dari nilai radius, jarak rata-rata dari titik pusat ke tepi, diagnosis kanker payudara ganas dan diagnosis kanker payudara jinak.

Dari hasil evaluasi dari model klasifikasi yang diterapkan, diperoleh nilai akurasi yang cukup besar yaitu 93,04%, Precision sebesar 80,00% dan recall sebesar 92,31%, hal ini

menunjukan algoritma C4.5 dapat digunakan sebagai algoritma dalam klasifikasi penyakit kanker payudara.

INTERNET SOURCES:

<1% - https://www.researchgate.net/publication/367199663_IMPLEMENTASI_ALGORITMA_C45_DALAM_PREDIKSI_PENYAKIT_KANKER
1% - http://mukhyi.staff.gunadarma.ac.id/Downloads/files/104121/Draft+Jurnal+Ilmiah+FIFO+AditiaDewiMukhyi.pdf
<1% - https://www.ijcaonline.org/research/volume135/number5/shah-2016-ijca-908385.pdf
<1% - https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/4629/2152/
1% - https://yankes.kemkes.go.id/view_artikel/738/kanker-berulang-recurrence-deteksi-dini-dan-pencegahan
1% - https://health.detik.com/berita-detikhealth/d-6322566/mencegah-risiko-kanker-payudara
1% - https://databoks.katadata.co.id/datapublish/2022/10/11/kanker-payudara-penyakit-kanker-paling-banyak-dialami-masyarakat-indonesia
1% - https://ugm.ac.id/id/berita/20137-kalkulasi-faktor-risiko-kanker-payudara-perlu-dilakukan
<1% - https://www.geeksforgeeks.org/kdd-process-in-data-mining/
<1% - https://core.ac.uk/download/pdf/327695069.pdf
<1% - https://www.academia.edu/75958810/Klasifikasi_Penderita_Penyakit_Diabetes_Menggunakan_Algoritma_Decision_Tree_C4_5
<1% - https://medium.com/@mimubarok.mim/decision-tree-pohon-keputusan-6484ad30c289
1% - https://jurnal.itpln.ac.id/kilat/article/download/1174/811
1% - https://jawabanapapun.com/apa-itu-akurasi-presisi-dan-recall/
<1% - https://www.esaunggul.ac.id/wp-content/uploads/2019/02/1.-Implementasi-Algoritma-Decision-Tree-C4.5-Untuk-Prediksi-Penyakit-Diabetes.pdf