Jurnal Teknik Informatika dan Sistem Informasi (JURTISI)
Vol. 5, No. 2, Desember 2025, hlm. 1 - 9

Comparison Of The Perfomance Of The C4.5 Anda *Naïve Bayes* Algorithms In Classifiying Lung Cancer

Perbandingan Kinerja Algoritma C4.5 Dan Naïve Bayes Dalam Mengklasifikasikan Penyakit Kanker Paru-Paru

Ma'ripah[1], Fitri Ayuning Tyas B[2], Ahmad Faizin C[3]

[1,2,3]Sistem Informasi, Fakultas Sains, Teknologi dan Kesehatan Universitas Muhammadiyah Brebes, Jl. Pangeran Diponegoro No. 184, Grengseng, Taraban, Kec. Paguyangan, Kabupaten Brebes, Jawa Tengah, 52276
Email: [1]iputgeureugeut@gmail.com*, [2]tyas_fa@stmikmpb.ac.id., [3]faizi.ahmad@stmikmpb.ac.id.

Abstract

Lung cancer is one of the most dangerous diseases with a high mortality rate. Early detection is crucial to increasing the chances of recovery. This study aims to evaluate the performance of the C4.5 and Naïve Bayes algorithms in classifying lung cancer cases. The dataset used is a lung disease prediction dataset from Kaggle, consisting of 30,000 records and 9 attributes. The experimental process was carried out using RapidMiner with the 10-fold cross-validation method, and evaluation was performed using a confusion matrix and t-test. The results showed that the C4.5 algorithm achieved an average accuracy of 94.44%, while Naïve Bayes achieved 87.05%. The t-test result yielded a p-value of 0.000, indicating that the performance difference between the two algorithms is statistically significant, with C4.5 proving to be superior in classifying lung cancer cases. This research is expected to serve as a reference for the development of disease classification systems, particularly in assisting with early and more accurate lung cancer detection.

Keywords: Lung Cancer, C4.5, Naïve Bayes, Classification, RapidMiner.

Abstrak

Kanker paru-paru merupakan salah satu penyakit berbahaya dengan tingkat kematian yang tinggi. Deteksi dini sangat penting dalam meningkatkan peluang kesembuhan. Penelitian ini bertujuan untuk menilai performa algoritma C4. 5 dan *Naïve Bayes* dalam pengklasifikasian penyakit kanker paru. *Dataset* yang digunakan adalah dataset predic terkena penyakit paru-paru dari Kaggle yang berisi 30.000 data dan 9 atribut. Proses eksperimen dilakukan menggunakan aplikasi *RapidMiner* dengan metode 10-*fold cross-validation* dan evaluasi menggunakan confusion matrix serta uji beda t-Test. Hasil penelitian menunjukkan bahwa algoritma C4.5 memperoleh akurasi rata-rata sebesar 94,44%, sedangkan *Naïve Bayes* sebesar 87,05%. Hasil uji t-Test, didapatkan nilai *p-value* sebesar 0.000 yang menunjukkan perbedaan kinerja kedua algoritma signifikan secara statistik, dengan C4.5 terbukti lebih unggul dalam klasifikasi penyakit kanker paru-paru. Penelitian ini diharapkan jadi acuan dalam pengembangan sistem untuk mengklasifikasi penyakit, terutama dalam membantu mendeteksi kanker paru-paru secara awal. secara lebih akurat dan efisien.

Kata kunci: Kanker Paru-Paru, C4.5, *Naïve Bayes*, Klasifikasi, *RapidMine*r.

1. PENDAHULUAN

Kesehatan merupakan aspek penting dalam mencapai kualitas hidup yang baik, sama halnya dengan fungsi vital paru-paru dalam proses bernapas [1]. Kanker paru-paru menjadi salah satu penyebab kematian tertinggi di dunia, dengan proporsi kasus global sebesar 11,6% dan angka kematian 18,4% (Globocan, 2018)[2].

Di Indonesia, insidennya mencapai 8,6% atau 30.023 kasus, dengan kematian sebesar 12,6% (26.095 jiwa). Faktor risiko utama meliputi kebiasaan merokok dan paparan polusi udara. Penyakit paru-paru lain seperti PPOK, asma, dan penyakit interstitial turut menjadi ancaman kesehatan pernapasan

Kanker paru-paru sulit dideteksi pada tahap awal, sehingga diperlukan metode deteksi yang efektif [3]. *Data mining* menjadi salah satu solusi karena mampu mengolah data dalam jumlah besar menjadi informasi yang bermanfaat [4]. Ada beragam teknik dalam *data mining*, salah satunya adalah klasifikasi, di mana terdapat berbagai metode, termasuk *Decision Tree* dan *Naïve Bayes* . Salah satu teknik dalam *data mining* adalah klasifikasi, Metode *Decision Tree* sendiri memiliki beberapa algoritma, salah satunya adalah. algoritma C4.5 yang paling umum digunakan. Sementara itu, *Naïve Bayes* adalah teknik klasifikasi yang sering digunakan karena metode ini didasarkan pada prinsip *probabilitas* dengan algoritma populer seperti C4.5 dan *Naïve Bayes*[5].

Algoritma C4.5 merupakan pengembangan dari ID3 yang membentuk model pohon keputusan, mampu menangani data numerik maupun kategorikal, memiliki mekanisme pruning, serta memberikan akurasi tinggi [6]. Selain itu, algoritma ini menunjukkan performa yang baik ketika diterapkan [7]. Algoritma ini juga mudah diinterpretasikan, menangani data yang tidak lengkap serta mekanisme pruning yang mencegah overfitting, sehingga model yang dihasilkan lebih akurat dan dapat diterapkan pada data baru[8]. Sementara itu, *Naïve Bayes* merupakan metode berbasis probabilitas yang sederhana, efisien, dan dapat diterapkan pada data kualitatif maupun kuantitatif, meskipun memiliki keterbatasan dalam pengukuran akurasi [9].

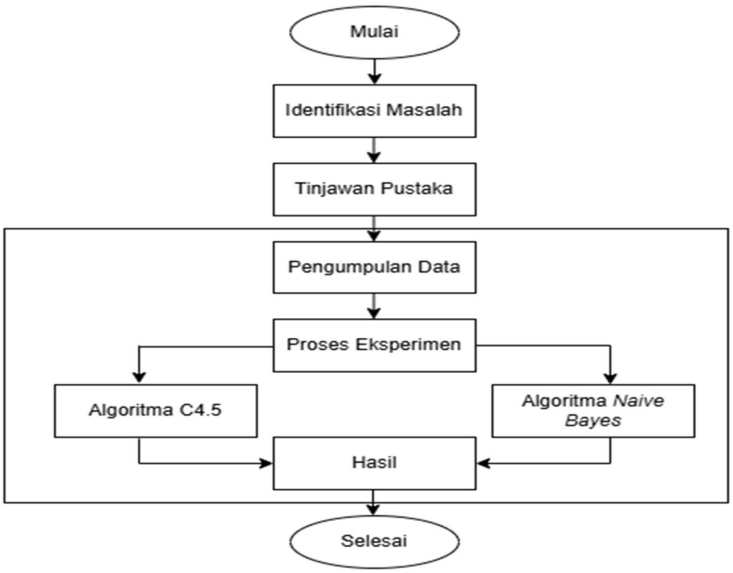
Penelitian ini membandingkan kinerja algoritma C4.5 dan *Naïve Bayes* dalam klasifikasi kanker paru-paru menggunakan dataset dari *Kaggle*. Pengujian dilakukan dengan metode 10-*fold cross-validation* pada aplikasi *RapidMiner*, serta evaluasi menggunakan *confusion matrix* dan uji t-*test.* Hasilnya diharapkan dapat menjadi acuan dalam pengembangan sistem deteksi dini kanker paru-paru secara akurat dan efisien.

2. METODE PENELITIAN

Metode yang digunakan pada penelitian ini adalah metode eksperimen, Metode eksperimen adalah metode yang dilakukan melalui proses uji coba, yang dirancang untuk menemukan jawaban yang ideal [10]. Metode eksperimen ini seirng digunakan untuk menguji keefektifan algoritma *machine learning* [11]. di mana pada penelitian ini akan dilakukan pengujian dan perbandingan kinerja dua algoritma klasifikasi, yaitu C4. 5 dan *Naïve Bayes*, dalam mengidentifikasi penyakit kanker paru-paru.

Data yang digunakan bersumber dari kumpulan data yang memprediksi risiko penyakit paru-paru yang diambil dari *Kaggle*, yang terdiri dari 30. 000 data individu dan 9 fitur serta 1 label kategori. Penelitian ini dilaksanakan melalui serangkaian langkah terstruktur untuk mengevaluasi kemampuan algoritma C4. 5 dan *Naïve Bayes* dalam mengklasifikasikan penyakit kanker paru-paru.

Dengan demikian, metode eksperimen yang diterapkan dalam penelitian ini diharapkan mampu memberikan gambaran yang jelas dan objektif mengenai perbandingan kinerja algoritma C4.5 dan *Naïve Bayes* dalam mengklasifikasikan penyakit kanker paru-paru, sehingga dapat mendukung pengambilan keputusan yang lebih akurat di bidang kesehatan berbasis data.



Gambar 1 Tahapan Penelitian

2.1 Identifikasi Masalah

Pada tahap ini, peneliti mengamati permasalahan yang terjadi di masyarakat, khususnya tingginya angka kematian akibat kanker paru-paru yang menunjukkan bahwa deteksi dini terhadap penyakit ini masih menjadi tantangan serius.

2.2 Tinjauan Pustaka

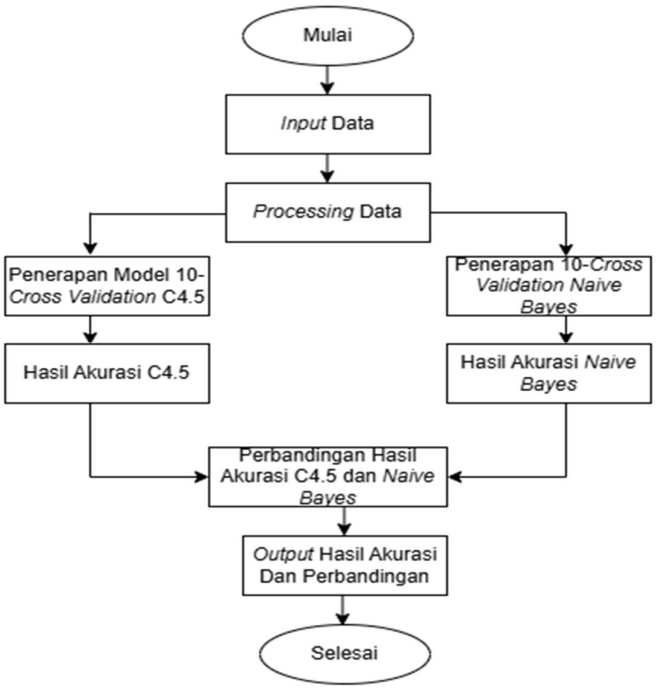
Peneliti mengumpulkan dan mempelajari literatur atau penelitian terdahulu yang relevan dengan topik penelitian. Ini mencakup teori-teori terkait *data mining*, algoritma klasifikasi (C4.5 dan Naïve Bayes), serta studi sebelumnya tentang klasifikasi penyakit paru-paru.

2.3 Pengumpulan Data

Data dikumpulkan dari sumber yang valid dan terpercaya yaitu dari *kaggle*, yaitu *dataset dataset predic* terkena penyakit paru-paru. Yang terdiri dari 30.000 entri 9 atribut dan satu label.

2.4 Proses Eksperimen

Proses ini mencakup pelatihan model (training), pengujian model (testing), serta pengolahan data menggunakan perangkat lunak *RapidMiner* Eksperimen ini dilakukan untuk membandingkan kinerja kedua algoritma dalam mengklasifikasikan data penyakit kanker paru-paru, sehingga dapat diketahui algoritma mana yang memberikan hasil lebih optimal.



Gambar 2 Tahapan Eksperimen Algoritma

2.4.1 Mulai
evaluasi algoritma dimulai dari tahap ini dengan tujuan untuk mencapai perbandingan efektivitas antara algoritma C4. 5 dan *Naïve Bayes* dalam mengelompokkan penyakit kanker paru-paru.

2.4.2 *Input* Data
*input datase*t publik yaitu *dataset predic* terkena penyakit paru-paru*, dataset* ini bersumber dari *kaggle.*

2.4.3 *Processing* Data
Langkah-langkah dalam preprocessing meliputi penanganan nilai kosong (missing values), normalisasi data numerik, encoding atribut kategorikal menjadi numerik, serta menghilangkan data outlier yang dapat mengganggu performa model.

2.4.4 Penerapan 10-*Fold Cross Validation* Pada Algoritma C4.5
Dataset dibagi menjadi sepuluh bagian (folds), di mana sembilan bagian digunakan untuk pelatihan dan satu bagian untuk pengujian. Prosedur ini dilakukan sebanyak sepuluh kali dengan bagian yang berbeda digunakan sebagai data uji di setiap iterasinya.

2.4.5 Penerapan 10-*fold cross validation* pada algoritma *naïve bayes*
Dengan menggunakan 10-*Fold Cross Validation*, dilakukan pelatihan dan pengujian model *Naïve Bayes* untuk mengevaluasi performa algoritma tersebut dalam mengklasifikasikan penyakit kanker paru-paru

2.4.6 Hasil Akurasi Pada Algoritma C4.5
Akurasi menunjukkan seberapa besar tingkat ketepatan model dalam mengklasifikasikan data uji dibandingkan dengan label sebenarnya.

2.4.7 Hail Akurasi Pada Algoritma *Naïve Bayes*
Tahap ini merupakan hasil dari proses cross validation yang dilakukan pada model *Naïve Bayes*. Sama seperti C4.5, nilai akurasi dihitung untuk mengukur kinerja model dalam memprediksi data.

2.4.8 Perbandingan Hasil Akurasi Algoritma C4.5 dan *Naïve Bayes*
Perbandingan ini bertujuan untuk menentukan algoritma mana yang memiliki performa lebih baik dalam mengklasifikasikan dataset penyakit paru-paru.

2.4.9 *Output* Hail Akurasi dan Perbandingan

Output ini dapat disajikan dalam bentuk tabel, grafik, atau narasi deskriptif, yang kemudian digunakan untuk menarik kesimpulan dalam penelitian.

2.4.10 Selesai

Semua langkah, mulai dari *input* data, *preprocessing*, pelatihan model, evaluasi, hingga analisis hasil telah diselesaikan.

3. HASIL DAN PEMBAHASAN

Hasil penelitian yang dilakukan dalam mengklasifikasi penyakit kanker paru-pparu menggunakan algoritma C.5 dan *naïve bayes* dengan metode eksperimen memiliki tahapan sebagai berikut:

3.1 Pengumpulan *Dataset*

Eksperimen dalam studi ini memanfaatkan *dataset* publik yang berasal dari *Kaggle* mengenai prediksi risiko terkena penyakit paru-paru, yang mencakup 30. 000 data dengan fitur-fitur seperti umur, gender, kebiasaan merokok, pekerjaan, kondisi rumah, kebiasaan begadang, aktivitas olahraga, perlindungan asuransi, dan riwayat penyakit bawaan

3.2 *Preprocessing* Data

Tahap ini mencakup pembersihan data untuk menghapus duplikat dan missing values, transformasi data untuk mengubah tipe atribut sesuai kebutuhan algoritma serta normalisasi jika diperlukan, seleksi atribut untuk memilih variabel yang relevan dengan penelitian, dan pemberian *role* atribut guna menentukan label (target) serta atribut prediktor (features).

3.3 Eksperimen

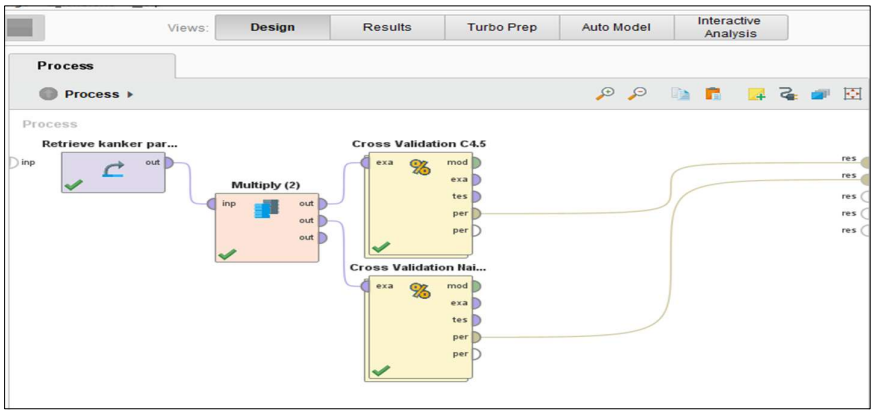
Proses tahapan eksperimen yang dilakukan:

3.3.1 *Input* Data

Proses ini dilakukan dengan menggunakan format *Microsoft Exel*. Dilanjut tahap pemberian label, proses ini bertujuan untuk mempermudah proses analisis dan klasifikasi pada tahap-tahap berikutnya dalam eksperimen.

3.3.2 Eksperimen

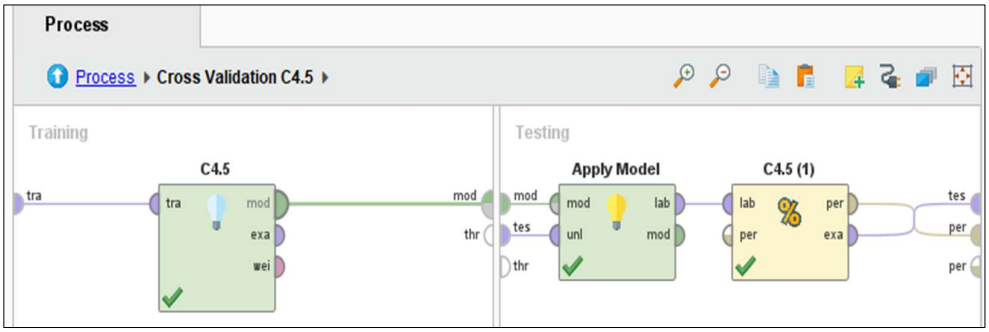
Proses eksperimen di *RapidMiner* dengan menggunakan empat operator utama. Pertama, operator *Retrieve* digunakan untuk memuat dataset Kanker Paru-Paru. Setelah proses *impor dataset* berhasil dilakukan, langkah selanjutnya adalah menduplikasi data menggunakan operator *Multipl,* Data hasil duplikasi kemudian diarahkan ke dua operator *Cross Validation*, masing-masing untuk algoritma C4.5 dan *Naïve Bayes*.



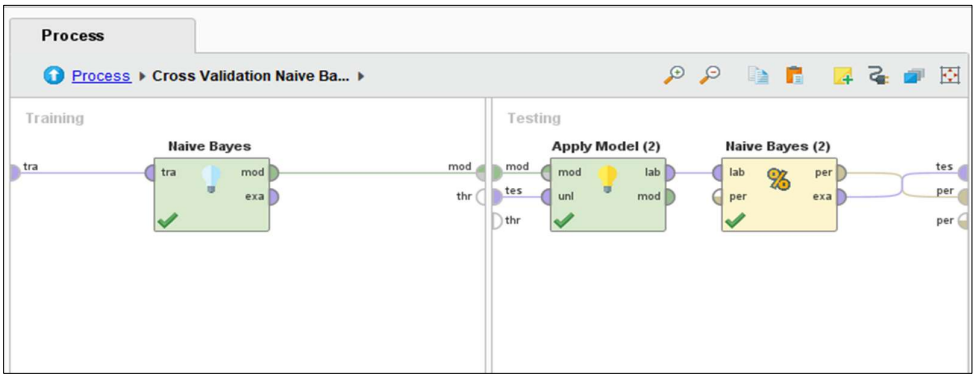
Gambar 3 Proses Eksperimen

3.3.3 *Cross-Validation*

Proses ini dibagi menjadi dua bagian utama, yaitu training (pelatihan) dan testing (pengujian). Model yang dihasilkan kemudian diteruskan ke bagian testing, di mana operator *Apply Model* digunakan untuk menerapkan model tersebut terhadap data uji. Selanjutnya, hasil prediksi dibandingkan dengan label yang sebenarnya menggunakan *operator Performance* (Classification) untuk mengukur kinerja model.



Gambar 4 *Cross-Validation* C4.5



Gambar 5 *Cross-Validation Naïve Bayes*

3.3.4 *Performance* Eksperimen

evaluasi terhadap hasil klasifikasi menggunakan algoritma C4.5 dan *Naïve Bayes*. Algoritma C4.5 menunjukkan performa terbaik dengan akurasi sebesar 94,44%, *precision* 98,37% (kelas Ya) dan 91,40% (kelas Tidak), serta *recall* 89,88% (kelas Ya) dan 98,63% (kelas Tidak). Sementara itu, algoritma *Naïve Bayes* menghasilkan akurasi 87,05%, dengan *precision* 86,34% (kelas Ya) dan 87,71% (kelas Tidak), serta *recall* 86,65% (kelas Ya) dan 87,42% (kelas Tidak).

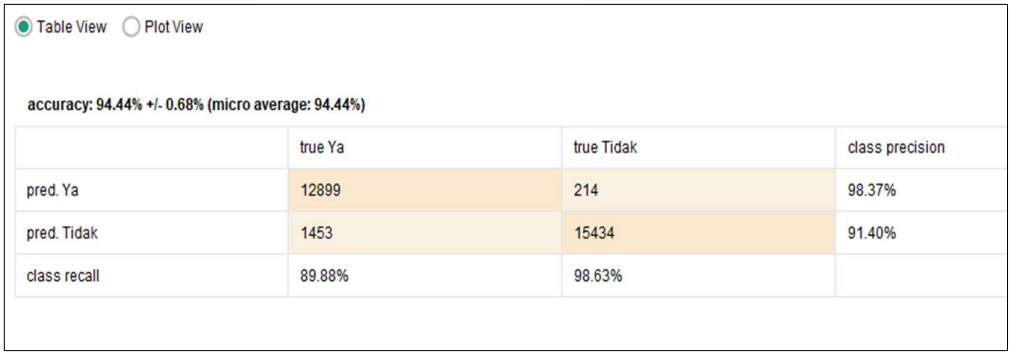


Table View Plot View

accuracy: 94.44% +/- 0.68% (micro average: 94.44%)

	true Ya	true Tidak	class precision
pred. Ya	12899	214	98.37%
pred. Tidak	1453	15434	91.40%
class recall	89.88%	98.63%	

Gambar 6 *Performance* C4.5

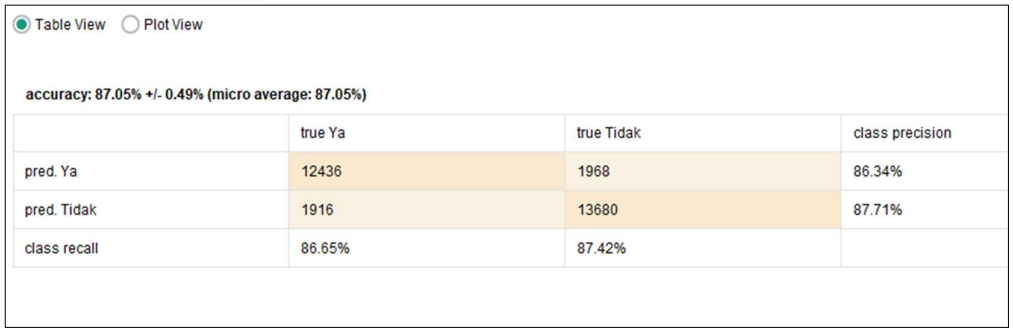


Table View Plot View

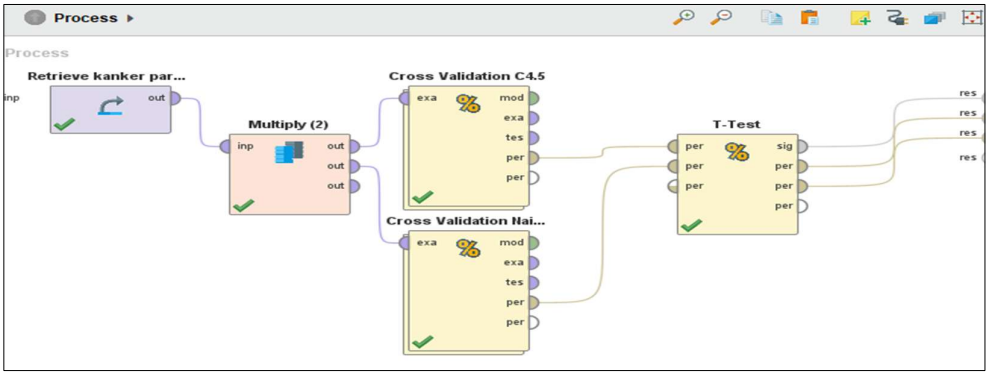
accuracy: 87.05% +/- 0.49% (micro average: 87.05%)

	true Ya	true Tidak	class precision
pred. Ya	12436	1968	86.34%
pred. Tidak	1916	13680	87.71%
class recall	86.65%	87.42%	

Gambar 8 *Performance Naïve Bayes*

3.3.5 Uji beda *T-test*

Hasil evaluasi performa seperti akurasi, presisi, dan *recall* dari kedua model kemudian dibandingkan menggunakan operator T-*Test* di *RapidMiner*. Uji ini menghasilkan *nilai p-value* yang menjadi dasar penentuan apakah perbedaan performa kedua algoritma bersifat signifikan secara statistik atau tidak.



Gambar 9 Uji beda T-*test*

3.3.6 Hasil dan Evaluasi

Berdasarkan hasil ini, dapat disimpulkan bahwa C4.5 memiliki kinerja yang lebih baik dalam mengklasifikasikan data penyakit kanker paru-paru dibandingkan *Naïve Bayes* dan perhitungan akurasi menggunakan *confusion matix.*

Tabel 1 Algoritma C4.5 Menggunakan *Confusion Matrix*

Predic	*Actual Class*		Total
	True Ya	*True* Tidak	
Ya	12,899	214	13,113
Tidak	1,453	15,434	16,887
Total *Actual*	14,352	15,648	30,000

Tabel 2 Algoritma *Naïve Bayes* Menggunakan *Confusion Matrix*

Predic	*Actual Class*		Total
	True Ya	*True* Tidak	
Ya	12,436	1,968	14,404
Tidak	1,916	13,680	15,596
Total *Actual*	14,352	15,648	30,000

Jika dilihat dari hasil evaluasi, algoritma C4.5 memiliki kinerja yang lebih baik dibandingkan dengan *Naïve Bayes*. C4.5 mencapai akurasi sebesar 94,44% dengan nilai presisi rata-rata 94,89% dan *recall* rata-rata 94,26%. Di sisi lain, algoritma *Naïve Bayes* mencapai tingkat akurasi 87,05%, dengan rata-rata presisi 87,03% dan nilai *recall* 87,04%. Hal ini menunjukkan bahwa C4.5 lebih akurat dan konsisten dalam mengklasifikasikan data pada penelitian ini.

4. KESIMPULAN

Tujuan dari penelitian ini untuk mengetahui seberapa besar perbandingan kinerja antara algoritma C4. 5 dan *naïve bayes* dalam mengklasifikasikan data mengenai penyakit kanker paru-paru dilakukan dengan metode 10-*fold x Validation*. Selain itu, ukuran kinerja dievaluasi menggunakan *confusion matrix* dan uji t-Test untuk menentukan model yang paling baik dengan menggunakan aplikasi *rapidminer* dalam mengukur tingkat akurasi yang dicapai. Hasil dari penelitian ini adalah pengolahan *dataset* yang telah dikumpulkan, dengan perhitungan yang dilakukan sesuai dengan model yang diusulkan. Berdasarkan hasil eksperimen, algoritma C4. 5 memperoleh rata-rata akurasi sebesar 0. 944 ± 0. 007, sementara algoritma *naïve bayes* mencatatkan akurasi sebanyak 0. 871 ± 0. 005. Dengan demikian, C4.5 menunjukkan performa klasifikasi yang lebih tinggi dibandingkan *naïve bayes*.

Berdasarkan hasil uji beda t-*test*, diperoleh nilai *p-value* = 0.000, yang lebih kecil dari taraf signifikansi 0.05. Hal ini menunjukkan bahwa perbedaan performa antara algoritma C4.5 dan *naïve bayes* signifikan secara statistik. Artinya, C4. 5 secara signifikan lebih baik daripada *naïve bayes* dalam mengategorikan penyakit kanker paru-paru. Oleh karena itu, algoritma C4.5 lebih direkomendasikan sebagai metode klasifikasi untuk mendeteksi penyakit kanker paru-paru karena memberikan hasil yang lebih akurat dan stabil.

DAFTAR PUSTAKA

[1] Christian, A., Yani, A., Hariyanto, H., & Sumanto, S. (2025). Analisis Machine Learning Untuk Prediksi Penyakit Paru-Paru Menggunakan Citra Paru. Jurnal Innovation And Future Technology (Iftech), 7, 122–131.

[2] Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3(1), 15–19. https://doi.org/10.57152/malcom.v3i1.591

[3] Huriah, D. A., & Nuris, N. D. (2023). Klasifikasi Penerima Bantuan Sosial UMKM Menggunakan Algoritma Naïve Bayes. JATI (Jurnal Mahasiswa Teknik Informatika), 7(1), 360–365. https://doi.org/10.36040/jati.v7i1.6300

[4] Meila Azzahra Sofyan, F., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer. JATI (Jurnal Mahasiswa Teknik Informatika), 7(2), 1409–1415. https://doi.org/10.36040/jati.v7i2.6810

Dengan Algoritma C4.5. Jurnal Teknologi Informatika Dan Komputer, 10(1), 168–182. https://doi.org/10.37012/jtik.v10i1.1985

[5] Martadiansyah, M. W., Ghufron, A., Hidayah, R. A., & Salzabila, D. (2025). Perbandingan Algoritma C4 . 5 dengan Naive Bayes Untuk Klasifikasi Curah Hujan. 3(c), 8–17..

[6] Musa, D. M., Sakti, D., Shantiony, K. A., Zega, S. K. P., Hamzah, S., Zega, Y. J., & Lubis, B. O. (2024). Penerapan Data Mining Untuk Klasifikasi Data Penjualan Pakan Ternak Terlaris Dengan Algoritma C4.5. Jurnal Teknologi Informatika Dan Komputer, 10(1), 168–182. https://doi.org/10.37012/jtik.v10i1.1985

[7] Matondang, T. R., Ramadhan Nasution, Y., Armansyah, & Furqan, M. (2024). Penerapan Algoritma C4.5 Pada Klasifikasi Status Gizi Balita. Jurnal Fasilkom, 14(1), 216–225. https://doi.org/10.37859/jf.v14i1.6941 [8]Meila Azzahra Sofyan, F., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer. JATI (Jurnal Mahasiswa Teknik Informatika), 7(2), 1409–1415. https://doi.org/10.36040/jati.v7i2.6810

[8] Maheswari, D., Anggraini, R. B., Aulia, S., Diah, Y., Lubis, B. O., Informasi, S., Informatika, B. S., Lunak, R. P., & Informatika, B. S. (2025). Implementasi algoritma c4.5 untuk klasifikasi dampak pola penggunaan media sosial terhadap kesejahteraan emosional. 9(2), 3411–3417.

[9] Dwilestari, G., & Afifah, T. A. (2025). Perbandingan Kinerja Algoritma Naive Bayes Dan Decision Tree Dalam Klasifikasi Kanker Paru-Paru. 9(1), 801–807.

[10] Fauzan, R., Vitianingsih, A. V., & Cahyono, D. (2025). Application of Classification Algorithms in Machine Learning for Phishing Detection Penerapan Algoritma Klasifikasi pada Machine Learning untuk Deteksi Phishing. 5(April), 531–540.

[11] Terapan, I., Putra, F., & Kunci, K. (2024). Penerapan Teknologi Machine Learning dalam Deteksi Dini Penyakit Pada Tanaman Pangan Sains dan Ilmu Terapan. 3, 1–5.