Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru

Dony Benaya Dinova[1], Budi Prasetiyo[2]

1,2 Jurusan Pendidikan Teknik Informatika UNNES
Email:donybenaya@stundents.unnes.ac.id

***Abstract*— Lung cancer is a serious disease that threatens global health due to its high morbidity and mortality rates. To reduce lung cancer mortality, a more precise diagnostic approach is needed. The aim of this study was to improve the sensitivity and specificity in detecting lung cancer, to support early detection efforts and more effective management. The research method involved a series of steps from data collection to model performance evaluation. Data was collected, cleaned, and analyzed for correlation before going through the preprocessing stage. Model training was conducted using the Random Forest Classifier algorithm, which proved effective in modeling the complexity of lung cancer data. The results prove that the model used can achieve an accuracy rate of 78%. The uniqueness of this research lies in the application of machine learning techniques to improve lung cancer diagnosis. By combining ADASYN oversampling with the Random Forest Classifier algorithm, this research has proven potential in the sensitivity and specificity of detecting the disease. These findings provide an important foothold in the development of more advanced and efficient diagnostic methods for lung cancer, paving the way for more effective early detection efforts and better disease management.**

***Keywords*— Machine Learning, Random Forest, ADASYN, Cancer, Lung.**

***Abstrak*— Kanker paru-paru adalah penyakit serius yang mengancam kesehatan dunia karena tingkat kesakitan dan kematian yang tinggi. Untuk mengurangi angka kematian akibat kanker paru-paru, diperlukan pendekatan diagnostik yang lebih presisi. Tujuan penelitian ini adalah meningkatkan sensitivitas dan spesifisitas dalam mendeteksi kanker paru-paru, untuk mendukung upaya deteksi dini dan manajemen yang lebih efektif. Metode penelitian melibatkan serangkaian langkah mulai dari pengumpulan data hingga evaluasi kinerja model. Data dikumpulkan, dibersihkan, dan dianalisis korelasinya sebelum melalui tahap preprocessing. Pelatihan model dilakukan menggunakan algoritma *Random Forest Classifier*, yang terbukti efektif dalam memodelkan kompleksitas data kanker paru-paru. Hasil penelitian membuktikan model yang digunakan dapat mencapai tingkat akurasi sebesar 78%. Keunikan penelitian ini terletak pada penerapan teknik machine learning untuk meningkatkan diagnosis kanker paru-paru. Dengan menggabungkan oversampling ADASYN dengan algoritma *Random Forest Classifier*, penelitian ini telah membuktikan potensi dalam sensitivitas dan spesifisitas mendeteksi penyakit tersebut. Temuan ini menjadi pijakan penting dalam pengembangan metode diagnostik yang lebih canggih dan efisien untuk kanker paru-paru, membuka jalan bagi upaya deteksi dini yang lebih efektif dan manajemen penyakit yang lebih baik.**

***Kata Kunci*— *Machine Learning*, *Random Forest*, ADASYN, Paru-Paru, Kanker.**

I. PENDAHULUAN

Kanker paru-paru menjadi penyakit serius yang mengancam kesehatan dunia karena tingkat kesakitan dan kematian yang tinggi. Pada tahun 2020, jumlah kasus kanker paru-paru mencapai angka yang mencengangkan, menyebabkan ribuan kematian setiap tahunnya. Kanker paru-paru merupakan kondisi di mana beberapa zat karsinogen memicu pertumbuhan sel-sel yang tidak dapat dikendalikan di dalam paru-paru [1]. Strategi untuk meminimalisir angka kematian akibat dari kanker ini adalah dengan fokus pada pencegahan dan deteksi dini [2].Gejala kanker paru-paru seringkali tidak spesifik, yang menyebabkan sebagian besar penderita mengabaikan gejala awalnya sebagai gangguan pernapasan umum atau penyakit ringan lainnya. Hal ini sering kali mengakibatkan penundaan diagnosis dan pengobatan yang tepat. Faktor lain yang memperburuk situasi ini adalah kurangnya kesadaran di antara petugas layanan kesehatan untuk melakukan pemeriksaan lanjutan terhadap pasien dengan gejala yang mencurigakan. Kesulitan dalam pendeteksian kanker paru telah mengakibatkan penyakit ini menjadi serius dan meningkatkan tingkat kematian yang tinggi [3]. Umumnya, kanker paru sering terjadi pada individu dengan kebiasaan merokok dan pola hidup tidak sehat telah menyebabkan kanker paru-paru menjadi jenis kanker yang sering di Indonesia, menempatkannya sebagai jenis kanker tersering ketiga di negara ini [4]. Sebagian besar kasus kematian akibat kanker paru-paru terkait dengan kebiasaan merokok dan eksposur terhadap asap rokok. Merokok adalah salah satu hal utama yang bisa membuat paru-paru rusak [5]. Merokok menjadi salah satu kebiasaan yang susah untuk dihentikan [6].Gangguan pada paru-paru dapat mengurangi efisiensi dan fungsi paru-paru dalam menyerap oksigen dari udara [7]. Strategi deteksi dini dan intervensi yang cepat sangatlah penting dalam meminimalisir angka kematian akibat dari kanker paru-paru. Salah satu pendekatan yang dapat digunakan adalah penerapan algoritma dalam bidang machine learning untuk mengidentifikasi kemungkinan adanya kanker paru-paru sejak dini.

Penelitian yang memanfaatkan metode *percentage split* dalam mendeteksi kanker paru-paru melaporkan tingkat keakuratan 86%, presisi 85%, dan *recall* 97,2% [8].

Penelitian melakukan klasifikasi menggunakan *Random Forest*, yang kemudian dievaluasi menggunakan *Confusion Matrix* mencapai 0,904. Pembagian data untuk pelatihan dan pengujian menggunakan rasio 80:20 [10].

Penerapan klasifikasi dalam bidang medis, Metode *Random Forest* memiliki kemampuan untuk mengenali kepentingan dari tiap fitur, serta memberikan kemudahan dalam interpretasi. Metode ini menonjol dengan unggulnya kinerja yang ditawarkannya. Dengan performa yang tinggi, metode ini memungkinkan pengguna untuk memahami kontribusi yang diberikan oleh setiap fitur terhadap hasil klasifikasi, sementara juga memberikan interpretasi yang mudah dimengerti [11].

Klasifikasi dalam data mining berarti pengelompokan data ke dalam kategori yang sudah ditentukan sebelumnya. Tujuannya untuk memprediksi kelas nilai dari objek-objek yang belum diketahui, dengan cara mengenali pola-pola yang cukup mewakili kelas-kelas data tersebut. Dalam proses ini, penting untuk memvalidasi data pelatihan guna menghasilkan model yang sesuai. Evaluasi model dilakukan menggunakan data uji untuk mengecek tingkat akurasi. Dengan menggunakan klasifikasi, kita dapat meramalkan nama atau nilai objek data [5].

Metode Random Forest Decision Trees merupakan pengembangan lebih lanjut dari metode CART. Salah satu perbedaan utamanya adalah dalam jumlah pohon konstruksi, di mana metode *Random Forest Decision Trees* menggunakan lebih dari satu pohon Hasil dari setiap pohon dalam klasifikasi atau regresi akan digabungkan dengan menggunakan proses voting. Proses ini bertujuan untuk menentukan kelas atau label yang paling sering muncul di antara hasil-hasil individu dari pohon-pohon tersebut. Nama "Random" pada metode ini merujuk pada proses acak dalam pembuatan pohon, di mana data yang digunakan akan diacak secara acak sebelum pembuatan setiap pohon[12].

Algoritma ADASYN mengadopsi pendekatan penyesuaian bobot distribusi pada data dari kelas minoritas, yang didasarkan pada tingkat kesulitan pembelajaran oleh model. Hal ini berarti bahwa data sintetis yang dihasilkan dari kelas minoritas, yang cenderung lebih sulit untuk dipelajari oleh model, akan diberi bobot lebih besar dibandingkan dengan data mayoritas yang relatif lebih mudah dipelajari [13].

Penelitian ini bertujuan untuk memprediksi kanker paru-paru menggunakan metode *Random Forest* serta melakukan pemilihan fitur berdasarkan gejala-gejala dan faktor risiko yang umum terjadi pada kanker paru-paru di Indonesia [14]. Diharapkan penelitian ini dapat menyumbangkan informasi yang berharga untuk meningkatkan akurasi diagnosis kanker paru-paru secara signifikan.

II. METODOLOGI

Penelitian mengumpulkan dataset yang berkaitan dengan topik, yakni kanker paru-paru (Gambar 1). Langkah berikutnya melibatkan *preprocessing* data untuk mendeteksi dan mengoreksi kesalahan dataset yang digunakan. Selanjutnya, dilakukan *balancing* data karena ketidakseimbangan kelas pada dataset. Metode ADASYN diterapkan untuk mengatasi masalah ini.

Setelah membagi dataset menjadi dua bagian, yaitu data training dan data testing, langkah selanjutnya adalah melakukan proses klasifikasi menggunakan metode *Random Forest* dan menyesuaikan parameter-parameter yang diperlukan. Setelah proses klasifikasi, dilakukan evaluasi terhadap kinerja model. Jika akurasi hasil klasifikasi memenuhi standar yang ditetapkan, maka proses dilanjutkan ke tahap berikutnya. Namun, jika tidak, dilakukan kembali dengan parameter yang sesuai agar dapat memperoleh hasil optimal. Langkah berikutnya adalah menarik kesimpulan berdasarkan hasil pengujian dan evaluasi.



Gbr. 1 Metode Penelitian.

III. *HASIL DAN PEMBAHASAN*

A. Pengumpulan Data

Penelitian dimulai dari pembuatan dataset *dummy* menggunakan Tools AI yaitu Chat-GPT (Gambar 2). Data *dummy* yang dibuat berjumlah 150 dengan imbalance dan missing value yang beragam. Data ini adalah kumpulan informasi kesehatan yang meliputi beberapa variabel seperti jenis kelamin, usia, kebiasaan merokok, keberadaan jari kuning, tingkat kecemasan, tekanan dari teman sebaya, riwayat penyakit kronis, kelelahan, alergi, gejala *wheezing*, konsumsi alkohol, batuk, sesak napas, kesulitan menelan, nyeri dada, dan riwayat kanker paru-paru.

```
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Jenis Kelamin      136 non-null    object
 1   Usia               136 non-null    float64
 2   Merokok            136 non-null    float64
 3   Jari Kuning        135 non-null    float64
 4   Kecemasan          136 non-null    float64
 5   Peer Pressure      136 non-null    float64
 6   Penyakit Kronis    135 non-null    float64
 7   Kelelahan          136 non-null    float64
 8   Alergi             138 non-null    float64
 9   Wheezing           135 non-null    float64
 10  Konsumsi Alkohol   136 non-null    float64
 11  Batuk              135 non-null    float64
 12  Sesak Napas        136 non-null    float64
 13  Kesulitan Menelan  137 non-null    float64
 14  Nyeri Dada         137 non-null    float64
 15  Kanker Paru-paru   150 non-null    object
```

Gbr. 2 Dataset

Akreditasi Sinta 5 KEMRISTEKDIKBUD No. 225/E/KPT/2022; E-ISSN 2723-7958

Penelitian ini menyimpulkan bahwa meskipun data dummy dapat dianggap sebagai *outlier* dan dapat terdeteksi dengan efektif menggunakan beberapa metode deteksi anomali, namun temuan tersebut juga menunjukkan bahwa asumsi tersebut tidak selalu akurat. Hal ini disebabkan oleh kemungkinan data palsu yang dapat disembunyikan di antara data normal jika dirancang dengan cermat. Oleh karena itu, dalam konteks penelitian ini, data palsu didefinisikan sebagai data *dummy*. Dengan demikian, penting untuk mempertimbangkan kemungkinan adanya data palsu yang disamarkan sebagai data normal dalam analisis data untuk menghindari kesalahan interpretasi.

B. Cleaning Data

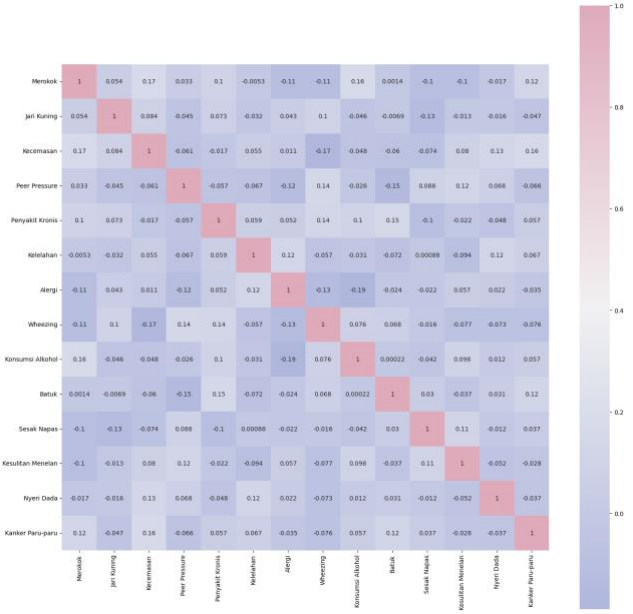
Dalam tahap pembersihan data, langkah-langkah krusial telah diambil untuk memastikan keintegritasan dan kegunaan dataset kesehatan yang dipergunakan. Awalnya, dilakukan pengecekan terhadap duplikasi data, yang kemudian dieliminasi untuk mencegah potensi pengaruh negatif pada hasil analisis. Kemudian, dilakukan inspeksi terhadap keberadaan nilai kosong dalam dataset. Untuk menangani nilai yang hilang, metode imputasi dengan menggunakan median digunakan untuk mengisi nilai numerik yang absen. Seusai proses imputasi, tidak ada lagi nilai yang kosong dalam dataset.

Langkah selanjutnya adalah fokus pada penanganan nilai kosong dalam variabel kategorikal. Dalam hal ini, baris data yang mengandung nilai kosong dihapus sepenuhnya untuk memastikan ketepatan analisis. Dengan demikian, dataset menjadi lebih bersih dan siap digunakan dalam tahapan analisis berikutnya.

Selanjutnya, dilakukan proses enkode label menggunakan metode untuk variabel kategorikal. Tindakan ini bertujuan untuk mengkonversi nilai-nilai kategorikal menjadi representasi numerik, yang akan memudahkan proses analisis oleh model machine learning di masa mendatang.

C. Analisis Korelasi

Analisis korelasi merupakan langkah penting dalam pemahaman hubungan antara berbagai fitur dalam sebuah dataset. Pada gambar heatmap korelasi (Gambar 3), nilai-nilai korelasi antara setiap pasangan fitur ditampilkan dalam bentuk warna dengan rentang dari -1 hingga 1. Korelasi positif dinyatakan dalam warna yang lebih terang, sementara korelasi negatif ditampilkan dalam warna yang lebih gelap.



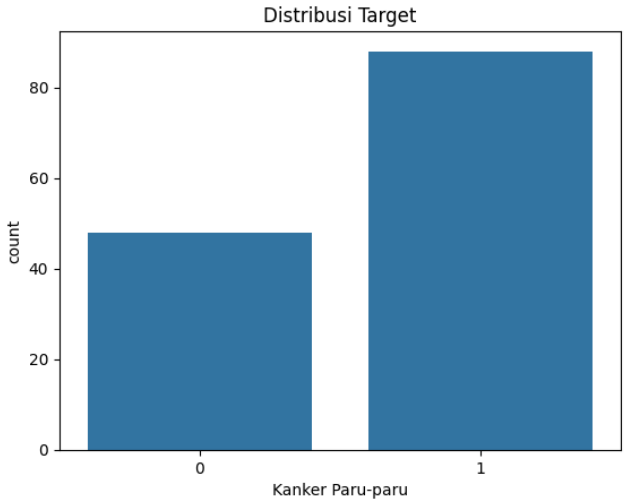
Gbr. 3 Analisis Korelasi

Dalam dataset ini beberapa fitur memiliki korelasi positif yang cukup kuat, ditandai dengan warna yang lebih cerah. Misalnya, kita dapat melihat bahwa 'Jari Kuning' dan 'Merokok' memiliki korelasi positif yang signifikan. Hal ini dapat dijelaskan dengan asumsi bahwa individu yang memiliki kebiasaan merokok cenderung memiliki paparan yang lebih tinggi.

Hasil Korelasi telah ditambahkan kolom baru 'Merokok&jarikuning', yang merupakan hasil perkalian antara 'Jari Kuning' dan 'Merokok'. Ini menunjukkan adanya usaha untuk mengeksploitasi hubungan antara kedua fitur ini dengan menggabungkannya menjadi satu fitur baru.

D. Preprocessing Data

Preprocessing Data melakukan pembagian dataset menjadi dua bagian, 75% dari data untuk latihan dan 25% untuk pengujian. Hal ini penting untuk menguji bahwa model dapat menangani data baru dengan baik dan dapat menghasilkan hasil yang dapat dipercaya. Menyesuaikan parameter dalam algoritma Random Forest merupakan tahap krusial dalam upaya menyesuaikan dan mengoptimalkan nilai-nilai parameter yang berpotensi memengaruhi kinerja keseluruhan model *Random Forest*.[15]

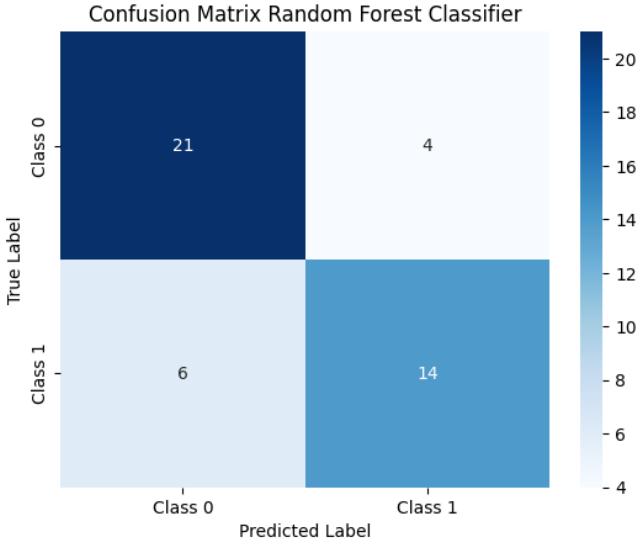


Gbr. 4 Data Imbalance.

Untuk mengatasi *imbalance* (Gambar 4), dilakukan oversampling menggunakan metode ADASYN. Metode ini bertujuan untuk meningkatkan jumlah sampel pada kelas minoritas dengan menambahkan sampel sintetis berdasarkan distribusi kepadatan lokal dari setiap titik data pada kelas tersebut. Dengan demikian, dapat memastikan bahwa model yang akan dikembangkan mampu mempelajari dengan baik kelas-kelas minoritas, sehingga meningkatkan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

E. Model

Model yang digunakan adalah *Random forest* dengan parameter yang telah disuaikan. Dalam pengujian model hasil terbaik yang dilakukan menghasilkan akurasi sebagai berikut.



Gbr. 5 Confusion Matrix

TABEL I. HASIL RANDOM FOREST

	Precision	Recall	FI-score	Support
0	0.78	0.84	0.81	25
1	0.78	0.70	0.74	20
Accuracy			0.78	45
Macro avg	0.78	0.77	0.77	45
Weighted avg	0.78	0.78	0.78	45

F. Evaluasi

Dengan akurasi sebesar 78%, model berhasil mengklasifikasikan 45 data uji dengan tepat. *Precision* untuk kelas 0, yang mengindikasikan pasien yang tidak terjangkit kanker paru-paru, adalah 0.78. Ini menunjukkan tingkat ketepatan dalam mengidentifikasi kasus negatif. Sementara itu, *recall* sebesar 0.84 menunjukan model dapat mengidentifikasi sebagian besar kasus positif yang sebenarnya. Namun, terdapat sedikit penurunan performa pada kelas 1, dengan *recall* sebesar 0.70, yang mengindikasikan bahwa sebagian kecil kasus kanker paru-paru mungkin tidak terdeteksi dengan baik oleh model. Meskipun demikian, nilai *F1-score* sebesar 0.81 menunjukkan keseimbangan yang baik antara *precision* dan *recall* untuk kedua kelas, menegaskan bahwa model telah menunjukkan kinerja yang seimbang dalam mengidentifikasi kelas positif dan negatif. Hasil evaluasi ini memberikan keyakinan tambahan bahwa model *Random Forest Classifier* dapat menjadi alat yang berguna dalam mendukung diagnosis kanker paru-paru..

IV. KESIMPULAN

Dalam penerapan model untuk mendeteksi kanker paru-paru, langkah-langkah preprocessing data, pelatihan model, dan evaluasi kinerja telah dilakukan secara cermat. Dengan menggunakan algoritma *Random Forest Classifier* dan menerapkan metode oversampling ADASYN untuk menangani ketidakseimbangan kelas, model berhasil mencapai tingkat akurasi sebesar 78%. Meskipun demikian, evaluasi lebih lanjut terhadap *precision, recall, dan F1-score* mengungkapkan adanya variasi kinerja antara kelas positif dan negatif. Dalam hal ini diperlukan peningkatan lebih lanjut dalam pengembangan model ini untuk memastikan ketepatan diagnosis dan memberikan kontribusi yang signifikan dalam bidang kesehatan.

DAFTAR PUSTAKA

[1] N. M. Aljamali, W. K. N. Al-Qraawy, and T. A. Helal, “Review on carcinogens materials in chemical laboratories,” *Int. J. Mol. Biol. Biochem.*, vol. 4, no. 1, pp. 17–25, Jan. 2022, doi: 10.33545/26646501.2022.v4.i1a.26.

[2] A. Desiani, B. Suprihatin, S. Yahdin, A. I. Putri, and F. R. Husein, “Bi-path Architecture of CNN Segmentation and Classification Method for Cervical Cancer Disorders Based on Pap-smear Images,” vol. 48, no. 3, 2021.

[3] A. Rifa’i and Y. Prabowo, “Diagnosis Kanker Paru-Paru dengan Sistem Fuzzy”.

[4] E. Priyanti, “Penerapan Algoritma Neural Network untuk Klasifikasi Kanker Paru,” vol. 9, no. 1, 2021.

[5] Arifin Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, and Ricky Firmansyah, “Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest,” *JUITIK*, vol. 3, no. 2, pp. 27–41, Jun. 2023, doi: 10.55606/juitik.v3i2.472.

[6] H. Trisnowati, O. Emildus Nabut, and U. Marlinawati, “Persepsi terhadap Peringatan Kesehatan Bergambar pada Bungkus Rokok dan Perilaku Merokok Remaja di Yogyakarta,” *JKK*, vol. 14, no. 2, p. 10, Oct. 2018, doi: 10.24853/jkk.14.2.10-20.

[7] A. A. Salsabil, M. B. Santoso, and Moch. Zainudiin, "PENANGANAN KEBIASAAN MEROKOK DI MASA PANDEMI COVID-19," *FOCUS*, vol. 3, no. 2, p. 131, Feb. 2021, doi: 10.24198/focus.v3i2.28458.

[8] E. SathiyaPriya, "A Study on Classification Algorithms and Performance Analysis of Data Mining using Cancer data to Predict Lung cancer Disease," vol. 3, no. 11, 2017.

[9] R. D. Marzuq, S. A. Wicaksono, and N. Y. Setiawan, "Prediksi Kanker Paru-Paru menggunakan Algoritme Random Forest Decision Tree".

[10] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru: Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification," *MALCOM*, vol. 3, no. 1, pp. 15–19, May 2023, doi: 10.57152/malcom.v3i1.591.

[11] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 8, pp. 1690–1692, Feb. 2018, doi: 10.1073/pnas.1800256115.

[12] M. S. Fadlilah, R. C. Wihandika, and B. Rahayudi, "Klasifikasi Penurunan Fungsi Kognitif Pasien Stroke Menggunakan Metode Klasifikasi Random Forest".

[13] F. S. Dhitama and F. A. Bachtiar, "Penentuan Kelayakan Debitur Menggunakan Metode Decision Tree C4.5 Dan Oversampling Adaptive Synthetic (ADASYN)".

[14] N. A. Khasanah, I. K. Oktaviyanti, and I. Yuliana, "HUBUNGAN RIWAYAT MEROKOK DAN TEMPAT TINGGAL DENGAN GAMBARAN SITOPATOLOGI KANKER PARU," vol. 2, no. 1.

[15] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis*, vol. 13, no. 2, pp. 67–75, Nov. 2020, doi: 10.51903/e-bisnis.v13i2.247.