INOTEK, Vol.8 Agustus 2024
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Penggunaan Algoritma KNN dalam Deteksi Awal Kanker Paru-Paru Menggunakan Data Medis

Diterima:
10 Juni 2024
Revisi:
10 Juli 2024
Terbit:
1 Agustus 2024

[1*]Mohammad Annan Makruf Mustofa, [2]Hermawan Nur Wahiid, [3]Bifadhlillah Marsheila Islami, [4]Aidina Ristyawan, [5]Erna Daniati
[1-5]Universitas Nusantara PGRI Kediri
[1]*makrufmustofa79@gmail.com*, [2]*mawan6989@gmail.com*, [3]*biff6167@gmail.com*, [4]*adinaristi@unpkediri.ac.id*, [5]*ernadaniati@unpkediri.ac.id*

Abstrak—Kanker paru-paru menjadi momok menakutkan dengan tingkat kematian tinggi. Deteksi dini menjadi kunci untuk meningkatkan peluang hidup pasien. Tujuan penelitian ini adalah mengkaji penggunaan algoritma K-Nearest Neighbors (KNN) dalam mendeteksi kanker paru-paru stadium awal melalui analisis data medis. Algoritma KNN dipilih karena kesederhanaan dan kinerjanya dalam mengklasifikasikan data kompleks. Data yang digunakan mencakup berbagai parameter medis, seperti ID pasien, umur, jenis kelamin, polusi udara, penggunaan alkohol, alergi debu, risiko genetik, dan penyakit paru-paru kronis. Hasil penelitian menunjukkan bahwa algoritma KNN mencapai tingkat akurasi tinggi dalam deteksi dini kanker paru-paru dengan pengaturan parameter K yang optimal. Temuan ini menunjukkan potensi besar algoritma KNN dalam aplikasi klinis untuk deteksi dini kanker paru-paru, yang dapat diimplementasikan dalam sistem pendukung keputusan medis untuk meningkatkan diagnosa dan intervensi dini.
Kata Kunci—*Deteksi dini kanker paru-paru; algoritma KNN; data medis; akurasi; peluang hidup pasien*

Abstract— *AbstractsLung cancer is a frightening threat with a high mortality rate. Early detection is the key to increasing the patient's chances of survival. This research examines the use of the K-Nearest Neighbors (KNN) algorithm in detecting early stage lung cancer through medical data analysis. The KNN algorithm was chosen because of its efficiency and performance in classifying complex data. The data used includes various medical parameters, such as patient ID, age, gender, air pollution, alcohol use, dust allergies, genetic risk, and chronic lung disease. The research results show that the KNN algorithm achieves a high level of accuracy in early detection of lung cancer with optimal K parameter settings. These findings demonstrate the great potential of the KNN algorithm in clinical applications for early detection of lung cancer, which can be implemented in medical decision support systems to improve early diagnosis and intervention.*
Keywords—*Early detection of lung cancer; KNN algorithm; medical data; accuracy; the patient's chance of survival*

Penulis Korespondensi:

Aidina Ristyawan,
Sistem Informasi,
Universitas Nusantara PGRI Kediri,
Email: aidinaristi@unpkediri.ac.id
ID Orcid: [https://orcid.org/0009-0003-2712-1507]
Handphone: 081232624460

I. PENDAHULUAN

Kanker paru merupakan suatu keganasan pada paru yang disebabkan oleh perubahan genetika pada sel epitel saluran nafas, sehingga terjadi proliferasi sel yang tidak terkendali. Keganasan ini dapat berasal dari organ paru itu sendiri (primer) maupun yang berasal dari luar paru (metastasis)[1]. Penelitian telah menegaskan bahwa ada keterkaitan antara kebiasaan merokok dan risiko terkena kanker paru-paru. Informasi yang tersedia menunjukkan bahwa merokok berperan sebagai faktor penyebab dalam sekitar 87% kasus kematian akibat kanker paru-paru[2]. Namun, deteksi dini seringkali terhambat oleh gejala awal yang tidak spesifik dan terbatasnya metode skrining yang tersedia.

Dalam era digital dan informasi ini, penggunaan teknologi berbasis data untuk mendukung diagnosa medis telah menjadi fokus utama dalam penelitian kesehatan. Data medis yang kaya akan informasi menawarkan peluang besar untuk diterapkan dalam metode data mining guna mengidentifikasi pola dan tren yang mungkin tidak terlihat oleh metode konvensional. Data Mining adalah proses mengumpulkan dan menganalisis data historis untuk menemukan pengetahuan, pola, atau hubungan tersembunyi dalam basis data besar yang dapat digunakan untuk pengambilan keputusan atau perbaikan keputusan di masa depan[3]. Salah satu algoritma machine learning yang banyak digunakan dalam klasifikasi dan prediksi adalah *K-Nearest Neighbors* (KNN). Algoritma ini dikenal karena kesederhanaannya dan kemampuannya dalam menangani data medis yang kompleks. Penelitian ini bertujuan untuk mengembangkan model data mining menggunakan algoritma KNN untuk deteksi awal kanker paru-paru. Dengan memanfaatkan data medis yang meliputi parameter-parameter penting seperti riwayat kesehatan, hasil laboratorium, dan pencitraan medis, kami berusaha untuk membangun model yang mampu mengklasifikasikan pasien ke dalam kategori risiko kanker paru-paru pada tahap awal. Penggunaan KNN diharapkan dapat memberikan akurasi yang tinggi dalam klasifikasi dan membantu tenaga medis dalam membuat keputusan diagnostik yang lebih cepat dan tepat.

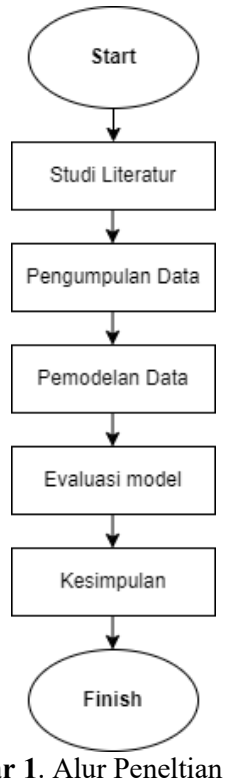
Implementasi teknologi KNN dalam memprediksi penyakit kanker paru-paru melibatkan beberapa langkah penting, mulai dari pengumpulan dan pra-pemrosesan data, pelatihan model, hingga evaluasi performa model. Prediksi adalah proses meramalkan suatu variabel di masa depan berdasarkan analisis data dari masa lalu[4]. Langkah awal dalam *preprocessing* data penting untuk mencegah duplikasi, mengatasi ketidakkonsistenan, memperbaiki kesalahan, serta menambahkan data yang diperlukan guna mendukung sistem yang sedang dikembangkan. Kemudian tahapan transformasi[5], Tahap Transformasi melibatkan pengubahan data yang telah dipilih melalui proses agregasi, sehingga data tersebut dapat diolah untuk keperluan pengujian

dalam data mining[6]. Pra-pemrosesan data diperlukan untuk memastikan bahwa data yang digunakan adalah bersih, konsisten, dan siap untuk dianalisis. Setelah data siap, tahap berikutnya adalah melatih model KNN dengan menggunakan data yang telah dikumpulkan. Model ini akan dilatih untuk mengenali pola-pola yang berhubungan dengan kanker paru-paru berdasarkan parameter yang ada. Selama proses pelatihan, model akan belajar dari contoh-contoh data sebelumnya untuk membuat prediksi yang akurat mengenai risiko kanker paru-paru pada pasien baru. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam upaya meningkatkan deteksi dini kanker paru-paru. Penggunaan teknologi KNN yang efektif dapat mendukung tenaga medis dalam membuat keputusan yang lebih informatif dan berbasis data, serta meningkatkan hasil klinis bagi pasien.

II. METODE

A. Alur Penelitian

Proses penelitian akan dimulai dengan menentukan objek penelitian, kemudian mengidentifikasi masalah-masalah yang akan diselesaikan serta tujuan penelitian. Setelah itu, dilakukan pengumpulan kemudian diproses menggunakan metode yang telah ditentukan, sehingga kesimpulan yang didapat berupa temuan pengetahuan baru atau pembuktian teori dengan metode yang digunakan dalam penelitian akan ditampilkan seperti gambar 1 [7].



Gambar 1. Alur Peneltian

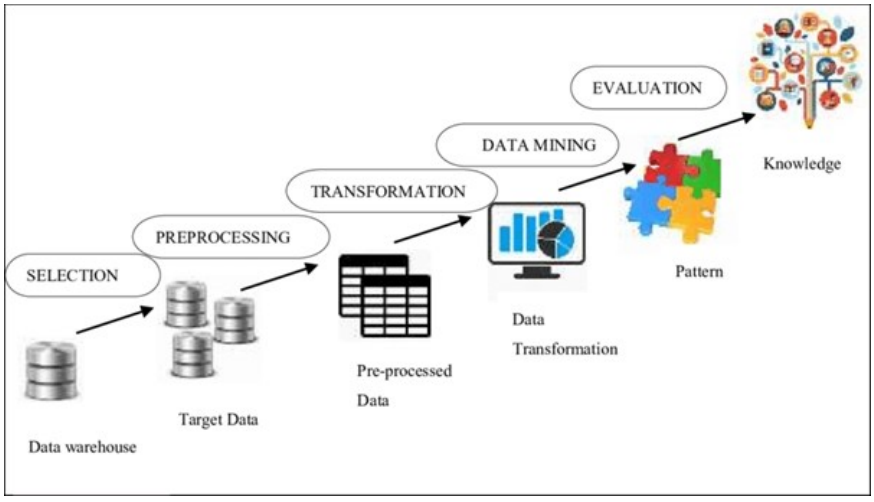
B. K-Nearest Neighbors

*K-Nearest Neighbor*s (KNN) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja berdasarkan kesamaan (*similarity*) antara data yang baru dengan data yang sudah ada. Tujuan dari metode ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel pembelajaran [8].

KNN bisa digunakan untuk memasukkan data baru (data uji) ke dalam kelompok data yang jaraknya berdekatan dengan data latih, sehingga metode ini bisa digunakan untuk mengklasifikasi data suara uji sesuai dengan kelompok data suara yang seharusnya. KNN akan mengelompokkan hasil perhitungan dengan data latih yang mempunyai kerabat terbanyak dalam nilai jangkauan yang ditentukan [9]

C. KDD

Knowledge Discovery in Database (KDD) adalah suatu proses analisis terstruktur yang bertujuan untuk memperoleh informasi yang akurat, baru, dan berguna serta untuk mengidentifikasi pola dari data yang besar dan kompleks [10] . Teknik untuk memperoleh informasi dari basis data yang sudah ada melalui beberapa tahapan, yaitu pemilihan data (*data selection*), pra-pemrosesan/pembersihan (*pre-processing/cleaning*), transformasi (*transformation*), penambangan data (*data mining*), dan interpretasi/evaluasi (*interpretation*) [11].



Gambar 2. Proses Knowledge Discovery in Database

III. HASIL DAN PEMBAHASAN

A. Data Selection

Data selection adalah proses memilih data dari suatu dataset untuk diolah lebih lanjut[12]. Data dalam penelitian ini memanfaatkan data yang diunduh dari

https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link, dengan penekanan pada kriteria kanker paru-paru. Data tersebut diakses pada bulan Mei 2024. Dataset ini mencakup berbagai parameter penting seperti ID pasien, umur, jenis kelamin, polusi udara, penggunaan alkohol, alergi debu, risiko genetik, dan penyakit paru-paru kronis. *Data selection* dilakukan untuk memilih data yang relevan dari dataset ini untuk diolah lebih lanjut.

No.	Nama Variabel	Tipe Data
1	index	integer
2	Patient Id	nominal
3	Age	integer
4	Gender	integer
5	Air Pollution	integer
6	Alcohol use	integer
7	Dust Allergy	integer
8	OccuPational Hazards	integer
9	Genetic Risk	integer
10	chronic Lung Disease	integer
11	Balanced Diet	integer
12	Obesity	integer
13	Smoking	integer
14	Passive Smoker	integer
15	Chest Pain	integer
16	Coughing of Blood	integer
17	Fatigue	integer
18	Weight Loss	integer
19	Shortness of Breath	integer
20	Wheezing	integer
21	Swallowing Difficulty	integer
22	Clubbing of Finger Nails	integer
23	Frequent Cold	integer
24	Dry Cough	integer
25	Snoring	integer
26	Level	nominal

Tabel 1. Tipe data dataset

B. Prepocessing

Preprocessing adalah langkah awal dalam memanipulasi teks asli, di mana beberapa tugas dasar dilakukan untuk mengubah atau menghilangkan elemen teks yang tidak relevan, sehingga mempersiapkan teks tersebut untuk pengolahan lebih lanjut[13]. Pada tahap

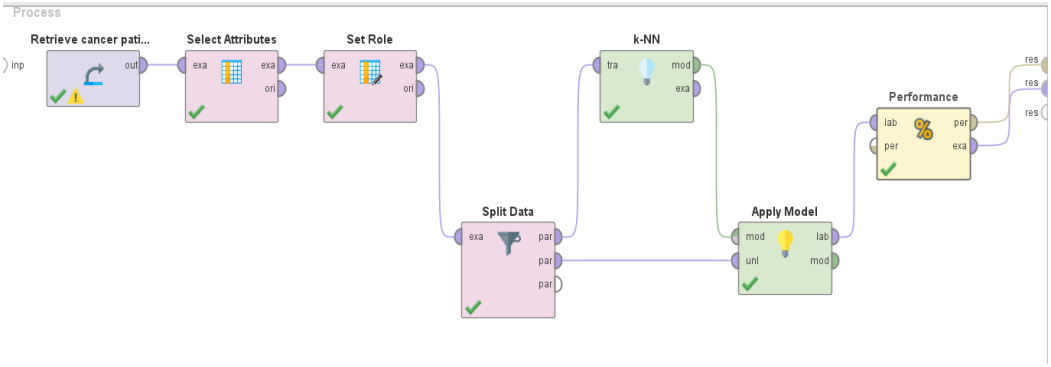
preprocessing, data yang terdapat pada dataset diolah untuk memastikan kualitas dan relevansinya. Penghapusan Kolom yang Tidak Relevan: Kolom "*index*" dan "*patient Id*" dihapus karena tidak memberikan kontribusi informasi yang signifikan untuk analisis lebih lanjut. Kolom-kolom ini lebih bersifat administratif dan tidak memiliki nilai prediktif terhadap diagnosis atau prognosis kanker paru-paru. Untuk analisis dan model prediksi di *RapidMiner,* kolom yang akan digunakan sebagai label adalah "*level*". Kolom "*level*" ini berisi informasi yang menjadi acuan utama untuk prediksi, seperti tingkat keparahan atau kategori dari kanker paru-paru. Pemilihan label ini sangat penting karena akan menjadi target yang diprediksi oleh model berdasarkan input dari fitur-fitur lainnya. Langkah-langkah tersebut memastikan bahwa data yang digunakan untuk analisis lebih lanjut adalah bersih, konsisten, dan siap untuk tahap transformasi dan data mining.

C. Transformation

Pada tahap transformasi, dilakukan proses normalisasi data, terutama untuk metode K-NN. Normalisasi ini bertujuan untuk mengurangi rentang/jarak nilai dari setiap data, karena fitur-fitur dalam data *lung cancer* memiliki skala yang berbeda, sehingga diperlukan proses normalisasi untuk mengurangi perbedaan jarak tersebut[14].

D. Data Mining

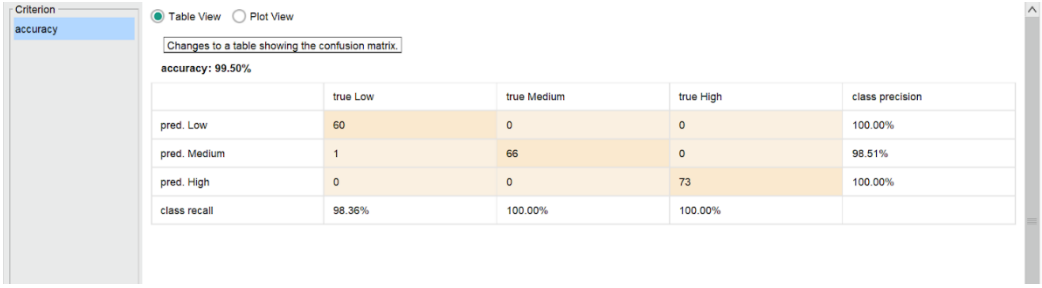
Pada tahap pertama, data pasien kanker diambil dari https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new- link. Kemudian, atribut yang relevan dipilih untuk memastikan hanya fitur-fitur yang diperlukan yang digunakan dalam analisis selanjutnya, sementara atribut yang tidak relevan atau berpotensi mengganggu dihilangkan. Selanjutnya, peran dari setiap atribut ditetapkan, termasuk mengidentifikasi mana yang akan menjadi fitur (*predictors*) dan mana yang akan menjadi target (*label*). Data kemudian dibagi menjadi dua bagian: satu untuk pelatihan (*training*) dan satu untuk pengujian (*testing*), dengan rasio *split data* 80% untuk pelatihan dan 20% untuk pengujian. Model K-NN digunakan untuk melatih data, di mana algoritma ini digunakan untuk klasifikasi berdasarkan kedekatan fitur. Model yang sudah dilatih kemudian diterapkan pada data pengujian untuk melihat seberapa baik model dapat mengklasifikasikan data baru yang belum pernah dilihat sebelumnya. Kinerja model dievaluasi menggunakan berbagai metrik, dengan hasil akhir berupa akurasi model yang mencapai 99,50%. Evaluasi ini membantu menentukan seberapa baik model dalam mengklasifikasikan data secara keseluruhan. Dalam proses ini, atribut "*Index*" dan "*Patient ID*" dibuang karena mereka tidak memberikan informasi yang relevan untuk proses klasifikasi dan hanya akan menjadi *noise* dalam analisis.



Gambar 4. Proses Data Mining

E. Evaluation

Evaluasi adalah proses untuk mengidentifikasi pola, sedangkan Penyajian Pengetahuan adalah proses di mana pemilik data memanfaatkan informasi yang telah diperoleh [15].



Criterion: accuracy

Table View / Plot View

Changes to a table showing the confusion matrix.

accuracy: 99.50%

	true Low	true Medium	true High	class precision
pred. Low	60	0	0	100.00%
pred. Medium	1	66	0	98.51%
pred. High	0	0	73	100.00%
class recall	98.36%	100.00%	100.00%	

Gambar 5. Hasil Akurasi

Dalam penelitian ini, kami menggunakan algoritma *K-Nearest Neighbors* (KNN) dalam memprediksi kanker paru-paru. Data akurasi menunjukkan tingkat akurasi yang tinggi yaitu 99.50%. Jika dibandingkan dengan penelitian terdahulu oleh Teguh et al. (2023) dengan algoritma yang sama menunjukkan bahwa berhasil mencapai akurasi sebesar 80.40%. Namun, penelitian kami menemukan bahwa metode KNN dapat meningkatkan akurasi prediksi hingga 99.50%.

Penelitian ini mengevaluasi efektivitas algoritma *K-Nearest Neighbors* (KNN) dalam mendiagnosis atau memprediksi kanker paru-paru dengan mempertimbangkan tahap *preprocessing data*. Dalam penelitian sebelumnya yang menjadi referensi, algoritma KNN diterapkan tanpa melalui tahap *preprocessing*, sedangkan dalam penelitian ini, data diolah terlebih dahulu untuk meningkatkan kualitas dan relevansinya. Tahap *preprocessing* dalam penelitian ini melibatkan penghapusan kolom yang tidak relevan seperti "*index*" dan "*patient Id*", yang tidak memiliki nilai prediktif terhadap analisis lebih lanjut. Selain itu, langkah-langkah seperti pengisian nilai yang hilang dan normalisasi data juga dilakukan untuk memastikan data yang digunakan berkualitas tinggi.

Kedua penelitian ini mengindikasikan bahwa teknik *data mining* dapat digunakan untuk mengembangkan model prediksi kanker paru-paru yang akurat. Akan tetapi, penelitian kami terbukti lebih efektif dalam mencapai akurasi prediksi yang tinggi. Temuan ini berpotensi mendukung pengembangan sistem prediksi kanker paru-paru yang lebih maju di masa depan. Peningkatan akurasi prediksi dari 87% menjadi 99.50% menunjukkan keunggulan metode KNN dalam menganalisis data dan memprediksi kanker paru-paru.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa penggunaan algoritma *K-Nearest Neighbors* (KNN) dapat mencapai tingkat akurasi yang sangat tinggi dalam memprediksi kanker paru-paru, dengan akurasi mencapai 99,50%. Hasil ini mengindikasikan bahwa teknik data mining, metode KNN, dapat digunakan secara efektif untuk mengembangkan model prediksi kanker paru-paru yang akurat. Temuan ini membuka peluang untuk pengembangan sistem prediksi kanker paru-paru yang lebih maju di masa depan, yang dapat memberikan manfaat signifikan dalam diagnosis dan pengobatan penyakit ini.

Dalam penelitian sebelumnya, algoritma KNN digunakan langsung tanpa penghapusan kolom tidak relevan atau penanganan nilai hilang, menghasilkan akurasi yang memadai tetapi berisiko terpengaruh oleh kualitas data yang kurang optimal. Sebaliknya, penelitian ini menunjukkan bahwa dengan *preprocessing*, algoritma KNN menunjukkan peningkatan akurasi dan kinerja secara keseluruhan. Data yang telah diolah menjadi lebih relevan dan berkualitas tinggi, memungkinkan algoritma KNN menghasilkan prediksi yang lebih tepat. Dari perbandingan ini, dapat disimpulkan bahwa *preprocessing* data memiliki dampak signifikan pada kinerja algoritma KNN, menekankan pentingnya *preprocessing* dalam analisis *data mining* untuk meningkatkan efektivitas algoritma prediktif.

UCAPAN TERIMAKASIH

Kami ingin mengucapkan terima kasih kepada fakultas teknik dan ilmu komputer khususnya prodi sistem informasi Universitas Nusantara PGRI Kediri, karena telah memberikan banyak dukungan terhadap penelitian kami.

DAFTAR PUSTAKA

[1] I. Buana and D. Agustian Harahap, "ASBESTOS, RADON DAN POLUSI UDARA SEBAGAI FAKTOR RESIKO KANKER PARU PADA PEREMPUAN BUKAN PEROKOK," 2022.

[2] D. Kencana Wulan, "FAKTOR PSIKOLOGIS YANG MEMPENGARUHI PERILAKU MEROKOK PADA REMAJA," 2012.

[3] R. Bahtiar, "Implementasi Data Mining Untuk Prediksi Penjualan Kusen Terlaris Menggunakan Metode K-Nearest Neighbor," 2023. [Online]. Available: https://jurnal.publikasitecno.id/index.php/jim203

[4] S. Y. , Y. M. Adiguno S, "Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda," *JURNAL SISTEM INFORMASI TGD*, 2022.

[5] I. Budiman and R. Ramadina, "Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi," *IJCCS*, vol. x, No.x, no. 1, pp. 1–5, 2015.

[6] A. Azis, A. T. Zy, and A. S. Sunge, "Prediksi Penjualan Obat Dan Alat Kesehatan Terlaris Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 1, pp. 117–124, Jan. 2024, doi: 10.47233/jteksis.v6i1.1078.

[7] Y. I. Anas, R. Firliana, and E. Daniati, "Decision Support System Pemilihan Bibit Unggul Tanaman Kelengkeng Menggunakan Metode Saw (Simple Additive Weighting)," 2020.

[8] A. Muhadi and A. Octaviano, "Penerapan Data Mining Untuk Prediksi Hasil Keuntungan Lelang Mesin X-Ray Tahun 2020 Dengan Metode K-Nearest Neighbor (Studi Kasus : PT.Ramadika Mandiri)," 2023. [Online]. Available: https://jurnal.publikasitecno.id/index.php/jim126

[9] R. A. Pangestu and S. Noris, "Analisa Data Mining Prediksi Lelang Suku Cadang Dengan Metode K-NearestNeighbor (Studi Kasus PT. Parmud Jaya Perkasa)," 2023. [Online]. Available: https://jurnal.publikasitecno.id/index.php/jim

[10] A. Karakteristik *et al.*, "Knowledge Discovery in Database Analysis of Traffic Accident Characteristic on Ahmad Yani Road Surabaya through Knowledge Discovery in Database Approach," 2018.

[11] M. Atalya, A. Leza, W. Utami, P. Anugrah, and C. Dewi, "PREDIKSI PRESTASI SISWA SMAS KATOLIK SANTO YOSEPH DENPASAR BERDASARKAN KEDISIPLINAN DAN TINGKAT EKONOMI ORANG TUA MENGGUNAKAN METODE KNOWLEDGE DISCOVERY IN DATABASE DAN ALGORITMA REGRESI LINIER BERGANDA," 2024.

[12] F. Alghifari and D. Juardi, "Fauzan Alghifari Penerapan Data Mining Pada Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *Jurnal Ilmiah Informatika (JIF)*, 2021.

[13] H. Najjichah, A. Syukur, and H. Subagyo, "PENGARUH TEXT PREPROCESSING DAN KOMBINASINYA PADA PERINGKAS DOKUMEN OTOMATIS TEKS BERBAHASA INDONESIA," 2019. [Online]. Available: http://research.

[14] R. W. Putri, A. Ristyawan, and M. N. Muzaki, "Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease," *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer*, vol. 2, no. 2, p. 143, Jul. 2022, doi: 10.32503/jtecs.v2i2.2708.

[15] G. Ramadhan *et al.*, "Penerapan Data Mining Menggunakan Algoritma C4.5 Dalam Mengukur Tingkat Kepuasan Pasien BPJS," *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, vol. 2, pp. 376–385, 2020.

[16] Kaggle. (2022). Lung Cancer Prediction [Dataset]. Kaggle. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data