

**ANALISIS ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN
PENERAPAN SMOTE DALAM DETEKSI DINI KANKER PARU-PARU**

SKRIPSI

Diajukan Untuk Memenuhi Sebagian Syarat Guna

Memperoleh Gelar Sarjana Komputer (S.Kom.)

Pada Program Studi Sistem Informasi



OLEH :

BIFADHLILLAH MARSHEILA ISLAMI

NPM: 2113030061

FAKULTAS TEKNIK DAN ILMU KOMPUTER (FTIK)
UNIVERSITAS NUSANTARA PERSATUAN GURU REPUBLIK INDONESIA
UN PGRI KEDIRI

2025

Skripsi oleh:

BIFADHLILLAH MARSHEILA ISLAMI

NPM: 2113030061

Judul:

**ANALISIS ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN
PENERAPAN SMOTE DALAM DETEKSI DINI KANKER PARU-PARU**

Telah disetujui untuk diajukan Kepada
Panitia Ujian/Sidang Skripsi Program Studi Sistem Informasi
FTIK UN PGRI Kediri

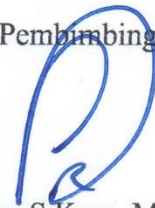
Tanggal: 2 Juli 2025

Pembimbing I



Sucipto, S.Kom, M.Kom
NIDN. 0721029101

Pembimbing II



Arie Nugroho, S.Kom, M.M, M.Kom
NIDN. 0712108103

Skripsi oleh:

BIFADHLILLAH MARSHEILA ISLAMI

NPM: 2113030061

Judul:

**ANALISIS ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN
PENERAPAN SMOTE DALAM DETEKSI DINI KANKER PARU-PARU**

Telah dipertahankan di depan Panitia Ujian/Sidang Skripsi

Program Studi Sistem Informasi FTIK UN PGRI Kediri

Pada tanggal: 11 Juli 2025

Dan dinyatakan telah memenuhi persyaratan

Panitia Penguji:

1. Ketua Penguji : Sucipto, S.Kom, M.Kom
2. Penguji I : Aidina Ristyan, S.Kom, M.Kom
3. Penguji II : Arie Nugroho, S.Kom, M.M, M.Kom

.....

.....

.....

Mengetahui,

Dekan Fakultas Teknik dan Ilmu Komputer



Dr. Sulistiono, M.Si
NIDN. 0007076801

PERNYATAAN

Yang bertanda tangan di bawah ini saya,

Nama : Bifadhlillah Marsheila Islami
Jenis Kelamin : Laki-laki
Tempat/ Tanggal Lahir : Nganjuk, 28 Januari 2002
NPM : 2113030061
Fak/ Prodi : FTIK/ S1-Sistem Informasi

Menyatakan dengan sebenarnya, bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi, dan sepanjang pengetahuan saya tidak terdapat karya tulis atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara sengaja dan tertulis diacu dalam naskah ini disebutkan dalam Daftar Pustaka.

Kediri, 2 Juli 2025

Yang menyatakan

Bifadhlillah Marsheila Islami
NPM: 2113030061

MOTTO

Sebenarnya tidak ada yang perlu dikhawatirkan, ALLAH memang tidak menjanjikan hidupmu selalu mudah. Tetapi, dua kali ALLAH berjanji bahwa:

فَإِنَّ مَعَ الْعُسْرِ يُسْرًا ۝٥

Maka, sesungguhnya beserta kesulitan ada kemudahan.

إِنَّ مَعَ الْعُسْرِ يُسْرًا ۝٦

Sesungguhnya beserta kesulitan ada kemudahan.

(Qs. Al – Insyirah 5-6)

ABSTRAK

Bifadhlillah Marsheila Islami: Analisis Algoritma K-Nearest Neighbors (KNN) Dan Penerapan Smote Dalam Deteksi Dini Kanker Paru-Paru, Skripsi, Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, UN PGRI Kediri, 2025.

Kata kunci: Kanker Paru-paru, K-Nearest Neighbors, SMOTE, Deteksi Dini, Klasifikasi, K-Fold Cross Validation, Evaluasi Model

Kanker paru-paru merupakan salah satu jenis kanker yang paling mematikan di dunia karena sering kali terdiagnosis pada stadium lanjut. Deteksi dini menjadi sangat penting untuk meningkatkan efektivitas pengobatan dan menekan angka kematian. Namun, metode skrining konvensional seperti *CT-scan* masih belum terjangkau secara luas, terutama di negara berkembang. Oleh karena itu, pemanfaatan teknologi kecerdasan buatan dan *data mining* menjadi alternatif potensial dalam pengembangan sistem deteksi dini kanker. Salah satu algoritma yang banyak digunakan untuk klasifikasi adalah *K-Nearest Neighbors* (KNN), yang dikenal efektif namun sensitif terhadap ketidakseimbangan data. Untuk mengatasi hal tersebut, diperlukan penerapan teknik oversampling seperti SMOTE (*Synthetic Minority Over-sampling Technique*) agar model dapat mengenali kelas minoritas secara lebih optimal.

Penelitian ini bertujuan untuk membangun sistem klasifikasi deteksi dini kanker paru-paru menggunakan algoritma KNN dan SMOTE berdasarkan data pasien yang diambil dari platform Kaggle. Dataset terdiri dari 1000 entri pasien dengan 26 atribut klinis dan demografis. Proses penelitian mengikuti pendekatan CRISP-DM yang mencakup business understanding, data understanding, data preparation, modelling, evaluation, dan deployment. Model KNN dibangun dengan parameter $k=15$ dan divalidasi menggunakan metode *Stratified K-Fold Cross Validation* sebanyak 10 lipatan. Evaluasi performa model dilakukan dengan metrik akurasi, presisi, *recall*, dan *F1-score*.

Hasil penelitian menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan performa klasifikasi, terutama terhadap data pada kelas minoritas. Model yang dihasilkan memiliki rata-rata akurasi sebesar 99,10%, dengan *precision*, *recall*, dan *F1-score* masing-masing sebesar 99,1%. Stabilitas antar fold menunjukkan performa yang konsisten dengan rentang akurasi antara 96% hingga 100%. Hasil ini menunjukkan bahwa kombinasi antara algoritma KNN dan teknik SMOTE sangat efektif dalam membangun sistem deteksi dini kanker paru-paru yang akurat dan andal.

Kesimpulannya, pendekatan ini memberikan solusi efisien untuk mendeteksi kanker paru secara lebih awal dengan biaya rendah dan akurasi tinggi. Namun, penelitian ini masih memiliki keterbatasan, seperti penggunaan satu jenis algoritma dan ketergantungan pada satu sumber dataset. Penelitian selanjutnya disarankan untuk membandingkan beberapa algoritma klasifikasi, menggabungkan data dari berbagai sumber, serta mengembangkan sistem berbasis *web* atau *mobile* agar dapat langsung digunakan dalam praktik klinis.

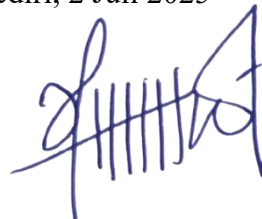
KATA PENGANTAR

Puji Syukur Kami panjatkan Kehadirat Allah Tuhan Yang Maha Kuasa, karena hanya atas perkenan-Nya penyusunan skripsi ini dapat diselesaikan. Skripsi dengan judul “Analisis Algoritma K-Nearest Neighbors (KNN) Dan Penerapan Smote Dalam Deteksi Dini Kanker Paru-Paru” ini ditulis guna memenuhi sebagian syarat untuk memperoleh gelar Sarjana Komputer, pada Program Studi Sistem Informasi Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri. Pada kesempatan ini diucapkan terima kasih dan penghargaan yang setulus-tulusnya kepada:

1. Dr. Zainal Afandi, M.Pd. selaku Rektor UN PGRI Kediri yang selalu memberikan dorongan motivasi kepada mahasiswa.
2. Dr. Sulistiono, M.Si. selaku Dekan Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri yang selalu memberikan dorongan motivasi kepada mahasiswa.
3. Sucipto, M.Kom selaku Ka Prodi Sistem Informasi Fakultas Teknik dan Ilmu Komputer UN PGRI Kediri yang selalu memberikan dorongan motivasi kepada mahasiswa.
4. Sucipto, S.Kom, M.Kom. dan Arie Nugroho, S.Kom, M.M, M.Kom. selaku Pembimbing skripsi yang dengan sabar membimbing dan memotivasi dalam menyelesaikan skripsi ini.
5. Kedua orang tua dan kakak saya yang telah mendukung dalam do’a, dana dan motivasi dalam penyelesaian skripsi ini.

Disadari bahwa skripsi ini masih banyak kekurangan, maka diharapkan tegur sapa, kritik dan saran-saran, dari berbagai pihak sangat diharapkan. Akhirnya, disertai harapan semoga skripsi ini ada manfaatnya bagi kita semua, khususnya bagi dunia pendidikan, meskipun hanya ibarat setitik air bagi samudra yang luas.

Kediri, 2 Juli 2025



Bifadhlillah Marsheila Islami
NPM: 2113030061

DAFTAR ISI

Halaman Sampul	i
Halaman Persetujuan.....	ii
Halaman Pengesahan	iii
Halaman Pernyataan.....	iv
Halaman Motto.....	v
ABSTRAK	vi
KATA PENGANTAR.....	vii
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
DAFTAR LAMPIRAN	xii
BAB I PENDAHULUAN.....	1
A. Latar Belakang	1
B. Identifikasi Masalah.....	4
C. Batasan Masalah.....	4
D. Rumusan Masalah	5
E. Tujuan Penelitian.....	5
F. Manfaat Penelitian	5
BAB II KAJIAN TEORI DAN HIPOTESIS	7
A. Kajian Teori.....	7
1. Kanker Paru-Paru	7
2. <i>Data Mining</i>	8
3. <i>K-Nearest Neighbors</i>	9
4. <i>SMOTE</i>	10
5. <i>Jupyter Notebook</i>	11
6. <i>Confusion Matrix</i>	12
7. <i>K-Fold Cross Validation</i>	14
B. Kajian Peneliti Terdahulu.....	15
C. Kerangka Berpikir.....	17
D. Hipotesis.....	18
BAB III METODOLOGI PENELITIAN	19
A. <i>Business Understanding</i>	19

B. <i>Data Understanding</i>	20
C. <i>Data Preparation</i>	20
1. Penghapusan Atribut Tidak Relevan	20
2. <i>Encoding Label</i>	20
3. Pemisahan Fitur dan Target.....	20
D. <i>Modelling</i>	21
E. <i>Evaluation</i>	21
F. <i>Deployment</i>	22
BAB IV HASIL DAN PEMBAHASAN	23
A. <i>Business Understanding</i>	23
B. <i>Data Understanding</i>	23
1. <i>Collect Initial Data</i>	23
2. <i>Describe Data</i>	24
3. <i>Explore Data</i>	49
C. <i>Data Preparation</i>	51
1. Penghapusan Atribut	51
2. <i>Encoding Label</i>	52
3. Pemisahan Fitur dan Target.....	53
D. <i>Data Modelling</i>	54
1. Validasi Model	54
2. Penanganan Ketidakseimbangan Data dengan SMOTE	55
3. Pelatihan Model	57
4. Perhitungan Manual KNN.....	57
E. <i>Evaluation</i>	60
F. <i>Deployment</i>	62
BAB V PENUTUP	64
A. Kesimpulan	64
B. Saran.....	64
DAFTAR PUSTAKA	66
LAMPIRAN	70

DAFTAR GAMBAR

Gambar 3.1 Model <i>Data Mining</i> CRISP-DM	19
Gambar 4.1 Distribusi <i>Age</i>	25
Gambar 4.2 Distribusi <i>Gender</i>	26
Gambar 4.3 Distribusi <i>Air Pollution</i>	27
Gambar 4.4 Distribusi <i>Alcohol Use</i>	28
Gambar 4.5 Distribusi <i>Dust Allergy</i>	29
Gambar 4.6 Distribusi <i>Occupational Hazards</i>	30
Gambar 4.7 <i>Genetic Risk</i>	31
Gambar 4.8 <i>Chronic Lung Disease</i>	32
Gambar 4.9 <i>Balanced Diet</i>	33
Gambar 4.10 <i>Obesity</i>	34
Gambar 4.11 <i>Smoking</i>	35
Gambar 4.12 <i>Passive Smoker</i>	36
Gambar 4.13 <i>Chest Pain</i>	37
Gambar 4.14 <i>Coughing of Blood</i>	38
Gambar 4.15 <i>Fatigue</i>	39
Gambar 4.16 <i>Weight Loss</i>	40
Gambar 4.17 <i>Shortness of Breath</i>	41
Gambar 4.18 <i>Wheezing</i>	42
Gambar 4.19 <i>Swallowing Difficulty</i>	43
Gambar 4.20 <i>Clubbing of Finger Nails</i>	44
Gambar 4.21 <i>Frequent Cold</i>	45
Gambar 4.22 <i>Dry Cough</i>	46
Gambar 4.23 <i>Snoring</i>	47
Gambar 4.24 <i>Level</i>	48
Gambar 4. 25 <i>Dataset Sebelum SMOTE</i>	55
Gambar 4. 26 <i>Dataset Setelah SMOTE</i>	56

DAFTAR TABEL

Tabel 4.1 Dataset <i>Lung Cancer</i>	24
Tabel 4.2 <i>Explore Data</i>	50
Tabel 4.3 Atribut Yang Digunakan	51
Tabel 4.4 <i>Encoding</i>	52
Tabel 4.5 Pemisahan Fitur dan Label	53
Tabel 4.6 Data Tiap Fold	55
Tabel 4. 7 Klasifikasi Data berdasarkan Jarak ke P1	60
Tabel 4.8 Akurasi Setiap Fold	61

DAFTAR LAMPIRAN

Lampiran 1 Kartu Bimbingan	70
Lampiran 2 Berita Acara Ujian Skripsi	71
Lampiran 3 Lembar Revisi Ujian Skripsi	72
Lampiran 4 Surat Keterangan Bebas Similarity	73
Lampiran 5 Bukti Halaman Awal Cek Similarity	74
Lampiran 6 Bukti Screenshot Submit Artikel/Loa Artikel/Screenshot Artikel Terbit	75
Lampiran 7 <i>Script Python</i>	76

BAB I

PENDAHULUAN

A. Latar Belakang

Kanker merupakan salah satu penyebab kematian tertinggi di dunia dan menjadi tantangan besar dalam bidang kesehatan global. Menurut Organisasi Kesehatan Dunia (WHO), setiap tahun terdapat sekitar 9,6 juta kematian akibat kanker di seluruh dunia. Kanker paru-paru menjadi salah satu jenis kanker paling mematikan karena sering kali didiagnosis pada stadium lanjut, sehingga mengurangi kemungkinan pasien untuk sembuh (Naufal et al., 2020). Berdasarkan laporan Globocan 2018, kanker paru menyumbang sekitar 11,6% dari total kasus kanker di dunia dan menyumbang 18,4% dari total kematian akibat kanker, menjadikannya penyebab kematian terbanyak di antara semua jenis kanker. Di Indonesia sendiri, prevalensi kanker paru mencapai 8,6% dengan tingkat kematian sebesar 12,6% atau sekitar 26.095 kasus (Yunianto et al., 2021). Hal ini menunjukkan urgensi akan sistem deteksi dini yang lebih efektif.

Kanker paru-paru merupakan penyakit yang ditandai oleh pertumbuhan sel-sel abnormal yang tidak terkendali di jaringan paru, yang umumnya berasal dari epitel bronkus (*bronchogenic carcinoma*). Penyakit ini biasanya berkaitan erat dengan faktor risiko seperti merokok, paparan polusi udara, bahan kimia karsinogenik, serta faktor genetik. Sayangnya, karena tidak adanya gejala yang signifikan pada tahap awal, banyak pasien yang baru mengetahui mengidap kanker paru-paru saat kondisinya sudah parah (Rifa'i & Prabowo, 2022). Deteksi dini sangat penting karena pengobatan pada tahap awal cenderung lebih efektif dibandingkan dengan pengobatan pada stadium lanjut. Sayangnya, metode skrining konvensional seperti *CT-scan* masih mahal dan belum dapat diakses secara merata oleh masyarakat luas.

Dalam era teknologi informasi dan data saat ini, penerapan kecerdasan buatan (AI) dan *machine learning* semakin menjadi perhatian dalam dunia medis untuk mendeteksi penyakit secara lebih cepat dan akurat. Salah satu algoritma yang menonjol adalah *K-Nearest Neighbors* (KNN), yang dikenal karena kesederhanaan namun keefektifannya dalam klasifikasi. KNN bekerja dengan mencari sejumlah tetangga terdekat dari data uji dalam ruang fitur dan mengklasifikasikan

berdasarkan mayoritas kelas dari tetangga tersebut (Han et al., 2017). Studi oleh Abdi Mangun et al., 2023 membuktikan bahwa KNN mampu mencapai akurasi 80,40% dalam mendeteksi kanker paru berdasarkan atribut klinis seperti usia, jenis kelamin, dan kebiasaan merokok.

Meskipun memiliki potensi tinggi, algoritma KNN juga memiliki kelemahan, salah satunya adalah sensitivitas terhadap ketidakseimbangan data. Dalam banyak kasus medis, jumlah data positif (misalnya pasien kanker) jauh lebih sedikit dibanding data negatif (pasien sehat), sehingga model cenderung bias terhadap kelas mayoritas. Oleh karena itu, diperlukan teknik penanganan ketidakseimbangan data seperti SMOTE (*Synthetic Minority Oversampling Technique*), yang dapat menciptakan data sintetis untuk kelas minoritas dan membantu memperbaiki performa model klasifikasi. SMOTE adalah teknik yang banyak digunakan dalam *machine learning* untuk meningkatkan akurasi klasifikasi pada dataset yang tidak seimbang. Teknik ini bekerja dengan membuat sampel sintetis dari data minoritas berdasarkan interpolasi antar tetangga terdekat. Dalam konteks deteksi kanker paru-paru, penerapan SMOTE terbukti meningkatkan sensitivitas model terhadap pasien yang benar-benar mengidap kanker, tanpa terlalu mengorbankan presisi (Nugroho & Harini, 2024). Hal ini penting karena dalam aplikasi medis, *false negative* (kesalahan dalam tidak mendeteksi penyakit) jauh lebih berbahaya dibanding *false positive*.

Penerapan gabungan antara algoritma KNN dan SMOTE telah diuji dalam beberapa studi mutakhir. Misalnya (Sachdeva et al., 2024) dalam penelitian menyebutkan bahwa penggunaan KNN dengan *preprocessing* SMOTE pada dataset kanker paru dapat meningkatkan akurasi dan sensitivitas model secara signifikan. Mereka melaporkan bahwa kombinasi ini mampu mengatasi ketimpangan data dan menghasilkan performa klasifikasi yang lebih stabil. Hasil ini menunjukkan bahwa pendekatan tersebut layak dipertimbangkan dalam pengembangan sistem pendukung keputusan klinis.

Selain algoritma dan teknik *oversampling*, pemilihan platform juga berpengaruh dalam implementasi. *Jupyter Notebook* menjadi pilihan karena fleksibilitas tinggi dalam *preprocessing*, tuning parameter, dan dokumentasi kode. Keunggulan ini menjadikannya alat yang ideal bagi peneliti dan pengembang

sistem deteksi penyakit berbasis *machine learning*. Jupyter juga memungkinkan visualisasi interaktif dan pelaporan hasil secara sistematis, sehingga proses analisis menjadi lebih transparan dan *reproducible*.

Data yang digunakan dalam penelitian ini diambil dari Kaggle dan terdiri atas berbagai atribut klinis seperti usia, jenis kelamin, kebiasaan merokok, konsumsi alkohol, serta kondisi medis lain seperti kelelahan, batuk darah, dan sesak napas. Atribut ini dinilai relevan dengan gejala umum kanker paru dan menjadi prediktor penting dalam model klasifikasi. Penelitian lain oleh (Paola Patricia et al., 2024) menunjukkan bahwa penggunaan atribut-atribut serupa dalam model KNN menghasilkan akurasi prediksi tinggi ketika dikombinasikan dengan teknik penyeimbangan seperti ADASYN.

Deteksi dini kanker paru bukan hanya sekedar tugas prediksi teknis, tetapi juga memiliki dampak besar terhadap peningkatan kualitas hidup pasien. Dengan diagnosis yang cepat dan akurat, terapi dapat segera diberikan, sehingga meningkatkan peluang sembuh. Oleh karena itu, integrasi antara *data mining*, algoritma klasifikasi, dan pendekatan penyeimbangan data sangat penting untuk menciptakan sistem deteksi kanker paru yang andal.

Dalam kerangka penelitian ini, evaluasi performa model tidak hanya akan dilakukan berdasarkan akurasi, tetapi juga melibatkan metrik seperti presisi, *recall*, dan *F1-score* untuk memberikan gambaran menyeluruh terhadap efektivitas model. Evaluasi dilakukan menggunakan metode validasi silang *K-Fold* untuk memastikan kestabilan model pada berbagai subset data. Evaluasi ini penting agar model yang dikembangkan tidak hanya berperforma baik pada data pelatihan, tetapi juga memiliki kemampuan generalisasi terhadap data baru.

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem diagnosis berbasis kecerdasan buatan yang dapat digunakan oleh institusi kesehatan untuk deteksi awal kanker paru. Dengan sistem ini, diharapkan keputusan medis dapat dilakukan lebih cepat, akurat, dan berbasis data. Selain itu, sistem ini juga dapat menjadi alat edukasi bagi masyarakat untuk lebih memahami faktor risiko dan gejala kanker paru secara objektif.

Manfaat penelitian ini tidak hanya terbatas pada bidang akademis, tetapi juga dapat berdampak langsung pada dunia nyata. Dalam jangka panjang, sistem

ini dapat dikembangkan menjadi aplikasi klinis yang terintegrasi dengan rekam medis digital rumah sakit, atau bahkan sebagai fitur pada aplikasi kesehatan berbasis seluler. Potensi pemanfaatan teknologi seperti ini membuka peluang baru dalam layanan kesehatan berbasis data dan otomatisasi.

Dari segi akademis, penelitian ini memperkaya literatur mengenai penerapan algoritma KNN dalam klasifikasi data kesehatan. Dengan studi kasus kanker paru, temuan penelitian ini dapat dijadikan dasar untuk pengembangan penelitian lanjutan dalam penyakit lainnya seperti kanker payudara, diabetes, dan penyakit jantung yang juga memerlukan deteksi dini berbasis data. Semakin banyak penelitian seperti ini, semakin besar peluang pengembangan sistem medis yang adaptif dan cerdas.

B. Identifikasi Masalah

Kanker paru-paru adalah salah satu penyebab utama kematian, terutama karena sering terdeteksi pada stadium lanjut. Metode deteksi dini yang ada, seperti *CT scan* cenderung mahal dan tidak selalu mudah diakses, sehingga banyak kasus yang terlambat ditangani. Selain itu, beberapa algoritma diagnostik yang digunakan saat ini memiliki keterbatasan, seperti sensitivitas terhadap *outlier* dan akurasi yang kurang optimal. Oleh sebab itu, diperlukan pendekatan baru yang lebih efektif dan efisien dalam mendeteksi kanker paru-paru secara dini.

Algoritma KNN menjadi salah satu solusi potensial karena kesederhanaannya dalam melakukan klasifikasi dan kemampuannya mengolah data. Meski begitu, penelitian terkait penerapan K-NN untuk deteksi kanker paru-paru masih terbatas. Penelitian ini bertujuan mengevaluasi kinerja K-NN dalam mendeteksi kanker paru-paru berdasarkan data klinis pasien serta membandingkan hasilnya dengan penelitian lain.

C. Batasan Masalah

Batasan masalah dalam penelitian ini sebagai berikut :

1. Penelitian ini hanya akan menggunakan algoritma *K-Nearest Neighbors* (KNN) dan tidak akan membahas algoritma lain secara mendetail.
2. Dataset yang digunakan dalam penelitian ini adalah data terkait pasien kanker paru-paru yang diambil dari *website* Kaggle dengan link sebagai

berikut: <https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link>.

3. Fokus penelitian akan terbatas pada analisis faktor-faktor yang mempengaruhi akurasi model KNN dalam mendeteksi kanker paru-paru serta evaluasi kinerja model tersebut berdasarkan metrik-metrik seperti akurasi, presisi, *recall*, dan *F1-score*.
4. Dataset penelitian ini berjenis teks dengan format tabel.

D. Rumusan Masalah

Berdasarkan batasan masalah tersebut, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Seberapa akuratkah algoritma *K-Nearest Neighbors* dalam mendeteksi dini kanker paru-paru?
2. Seberapa efektifkah penerapan SMOTE dalam meningkatkan performa model?
3. Seberapa efektifkah penghapusan fitur untuk atribut yang tidak diperlukan ?

E. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Untuk mengukur akurasi algoritma *K-Nearest Neighbors* dalam deteksi dini kanker paru-paru.
2. Menerapkan penghapusan fitur dan SMOTE untuk meningkatkan *accuracy*.

F. Manfaat Penelitian

Penelitian ini memiliki beberapa manfaat penting yang dapat memberikan kontribusi signifikan dalam bidang kesehatan, khususnya dalam deteksi dini kanker paru-paru. Pertama, penelitian ini diharapkan dapat memperkaya literatur ilmiah mengenai penggunaan algoritma *K-Nearest Neighbors* (KNN) dalam analisis data kesehatan, sehingga dapat menjadi referensi bagi peneliti yang ingin mengembangkan teknik serupa. Kedua, hasil penelitian ini dapat memberikan wawasan baru bagi praktisi medis dalam menerapkan teknologi analisis data untuk meningkatkan akurasi diagnosis dan pengobatan kanker paru-paru. Selain itu,

penelitian ini juga bertujuan untuk meningkatkan kesadaran masyarakat mengenai pentingnya deteksi dini kanker paru-paru, serta pemanfaatan teknologi dalam upaya pencegahan dan pengobatan. Dengan demikian, hasil penelitian ini tidak hanya bermanfaat bagi kalangan akademis, tetapi juga bagi masyarakat luas dan tenaga kesehatan dalam meningkatkan kualitas layanan kesehatan.

DAFTAR PUSTAKA

- Abdi Mangun, T., Nurdiawan, O., & Irma Purnamasari, A. (2023). *LUNG CANCER ANALYSIS USING K-NEAREST NEIGHBOR ALGORITHM* (Vol. 2, Issue 2). https://ejournal.ubibanyuwangi.ac.id/index.php/jurnal_tinsika
- Arifin Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, & Ricky Firmansyah. (2023). Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Jurnal Ilmiah Teknik Informatika Dan Komunikasi*, 3(2), 27–41. <https://doi.org/10.55606/juitik.v3i2.472>
- Astriani Salwa, M. (2025). *Optimasi Model Algoritma Klasifikasi Data Mining Menggunakan Metode Feature Selection Untuk Prediksi Penyakit Stroke*. <https://doi.org/10.37817/Tekinfo.v26i1>
- Banafshah Shafa, Hanny Hikmayanti Handayani, Santi Arum Puspita Lestari, & Yana Cahyana. (2024). Prediksi Kanker Paru dengan Normalisasi menggunakan Perbandingan Algoritma Random Forest, Decision Tree dan Naïve Bayes. *Decode: Jurnal Pendidikan Teknologi Informasi*, 4(3), 1057–1070. <https://doi.org/10.51454/decode.v4i3.779>
- Chen, N.-B., Xiong, M., Zhou, R., Zhou, Y., Qiu, B., Luo, Y.-F., Zhou, S., Chu, C., Li, Q.-W., Wang, B., Jiang, H.-H., Guo, J.-Y., Peng, K.-Q., Xie, C.-M., & Liu, H. (2022). CT radiomics-based long-term survival prediction for locally advanced non-small cell lung cancer patients treated with concurrent chemoradiotherapy using features from tumor and tumor organismal environment. *Radiation Oncology*, 17(1), 184. <https://doi.org/10.1186/s13014-022-02136-w>
- Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). IJCIT (Indonesian Journal on Computer and Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. In *IJCIT (Indonesian Journal on Computer and Information Technology)* (Vol. 6, Issue 2).
- Han, J., Kamber, M., & Pei, J. (2017). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.
- Hasran. (2020). *Indonesian Journal of Data and Science Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor*. 1(1), 6–10. <http://bit.ly/datasetcardio>.
- Imantata Muhammad, D., & Falih, N. (2021). *Penggunaan K-Nearest Neighbor (KNN) untuk Mengklasifikasi Citra Belimbing Berdasarkan Fitur Warna*. 1. <https://doi.org/10.52958/iftk.v17i1.2132>
- Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online

- Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 15(2), 121. <https://doi.org/10.22146/ijccs.65176>
- Joseph, J., & Rotty, L. W. A. (2020). Kanker Paru: Laporan Kasus. *Medical Scope Journal*, 2(1). <https://doi.org/10.35790/msj.2.1.2020.31108>
- Muhadi, A., & Octaviano, A. (2023). Penerapan Data Mining Untuk Prediksi Hasil Keuntungan Lelang Mesin X-Ray Tahun 2020 Dengan Metode K-Nearest Neighbor (Studi Kasus : PT.Ramadika Mandiri). In *Jurnal Informatika MULTI* (Vol. 1, Issue 2). <https://jurnal.publikasitecno.id/index.php/jim126>
- Naufal, S. A., Adiwijaya, A., & Astuti, W. (2020). Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray. *JURIKOM (Jurnal Riset Komputer)*, 7(1), 162. <https://doi.org/10.30865/jurikom.v7i1.2014>
- Nikmatul Kasanah, A., Pujiyanto, U., Elektro, T., Teknik, F., & Negeri Malang, U. (2017). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Masa Berlaku Mulai*, 1(3), 196–201. <https://doi.org/https://doi.org/10.29207/resti.v3i2.945>
- Nugroho, A., & Harini, D.** (2024). Teknik Random Forest untuk Meningkatkan Akurasi Data Tidak Seimbang. *JSITIK*, 2(2). <https://doi.org/10.53624/jsitik.v2i2.XX>
- Pangestu, R. A., & Noris, S. (2023). Analisa Data Mining Prediksi Lelang Suku Cadang Dengan Metode K-NearestNeighbor (Studi Kasus PT. Parmud Jaya Perkasa). In *Jurnal Informatika MULTI* (Vol. 1, Issue 4). <https://jurnal.publikasitecno.id/index.php/jim>
- Paola Patricia, A.-C., Marlon Alberto, P.-M., Er-nesto, B.-M., Sharith Alejandra, B.-A., Camilo, B.-C., & Fabian, R.-. (2024). Improving the Accuracy of Predictive Models in Imbalanced Lung Cancer Data. *International Conference on Swarm Intelligence*, 219–230.
- Permana, B. A. C., & Djamaluddin, M. (2023). Penerapan Python Dalam Data Mining Untuk Prediksi Kanker Paru. *Infotek: Jurnal Informatika Dan Teknologi*, 6(2), 470–477. <https://doi.org/10.29408/jit.v6i2.17816>
- Putri, R. W., **Ristyawan, A., & Muzaki, M. N.** (2022). Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease. *JTECS: Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem Dan Komputer*, 2(2), 143. <https://doi.org/10.32503/jtecs.v2i2.2708>
- Reynaldi, A., Trisyani, Y., & Adiningsih, D. (n.d.). *KUALITAS HIDUP PASIEN KANKER PARU STADIUM LANJUT*.
- Reza Muzaki, M., Melinda Nurfajriana, I., Putri Anugerah Ilahi, F., **Ristyawan, A., & Daniati, E.** (2024). Implementasi Data Mining dengan Algoritma K-

- Nearest Neighbors untuk Memprediksi Risiko Diabetes Menggunakan Chatbot Telegram. In *Agustus* (Vol. 8). Online. <https://doi.org/10.29407/inotek.v8i1.4946>
- Rifa'i, A., & Prabowo, Y. (2022). *Krea-TIF: Jurnal Teknik Informatika Diagnosis Kanker Paru-Paru dengan Sistem Fuzzy*. 10(1), 19–28. <https://doi.org/10.32832/kreatif.v10i1.6317>
- Rofiani Riska, OKtaviani Luluk, Vernanda Dwi, & Hendriawan Tri. (2024). *Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5*.
- Sachdeva, R. K., Bathla, P., Rani, P., Lamba, R., Ghantasala, G. S. P., & Sharma, R. R. (2024). Lung Cancer Classification Using Machine Learning. *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, 1–6. <https://doi.org/10.1109/ICICEC62498.2024.10808741>
- Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Irawan, Y., & Hayami, R. (2023). *Indonesian Journal of Machine Learning and Computer Science Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru*. 3, 15–19. <https://doi.org/10.57152/malcom.v3i1.591>
- Siringoringo, R. (2018). *KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR* (Vol. 3, Issue 1).
- Sucipto, S., Dwi Prasetya, D., & Widiyaningtyas, T. (2024). Educational Data Mining: Multiple Choice Question Classification in Vocational School. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, 23(2), 379–388. <https://doi.org/10.30812/matrik.v23i2.3499>
- Syifa Habibi Aulia Nur, **Nugroho Arie**, & **Firliana Rina**. (n.d.). *Habibi Aulia Nur Syifa Perbandingan Algoritma Naïve Bayes Classifier Dan K-Nearest Neighbors Untuk Analisis Sentimen Covid-19 Di Twitter*. Retrieved December 19, 2024, from <https://doi.org/10.33884/jif.v11i01.7069>
- Tempola, F., Muhammad, M., & Khairan, A. (2018). *PERBANDINGAN KLASIFIKASI ANTARA KNN DAN NAIVE BAYES PADA PENENTUAN STATUS GUNUNG BERAPI DENGAN K-FOLD CROSS VALIDATION COMPARISON OF CLASSIFICATION BETWEEN KNN AND NAIVE BAYES AT THE DETERMINATION OF THE VOLCANIC STATUS WITH K-FOLD CROSS VALIDATION*. 5(5), 577–584. <https://doi.org/10.25126/jtiik20185983>
- Wahyudinnur, R. A., Arriyanti, E., & Wahyuni. (2024). *Sistem Rekomendasi Peminatan Pada Prodi Teknik Informatika STMIK Widya Cipta Dharma Menggunakan Algoritma C5.0*.

- Wijaya, G., Fitriyah, N. Q., Izzi, Y., & Amiri, A. (2025). Optimasi Strategi Promosi Pendidikan Menggunakan Algoritma FP-Growth untuk Identifikasi Wilayah Strategis. *Journal of Information System and Technology*, 06(01), 12–17. <https://doi.org/10.37253/joint.v6i1.10295>
- Yunianto, M., Anwar, F., Nur Septianingsih, D., Dwi Ardyanto, T., & Farits Pradana, R. (2021). KLASIFIKASI KANKER PARU PARU MENGGUNAKAN NAÏVE BAYES DENGAN VARIASI FILTER DAN EKSTRAKSI CIRI GRAY LEVEL CO-OCCURANCE MATRIX (GLCM). *Indonesian Journal of Applied Physics*, 11(2).