PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 1 No. 1 (2025) pp: 118-121
P-ISSN: XXXX-XXXX, e-ISSN: XXXX-XXXX

Pengambilan Keputusan Medis Berbasis Algoritma *K-Nearest Neighbor* (KNN) Dalam Klasifikasi Pasien Stroke

[1]Risma Ananta Maulida, [2]Suci Anisa Aulia, [3]Ridho, [4]Satrio Dzulfahmi Yulianto, [5]Shania Clara Efendi, [6]Maulana Fansyuri

[1]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[2]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[3]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[4]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[5]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[6]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

[1]anantamaulidar@gmail.com, [2]sucianisaaulia@gmail.com, [3]ridho.dodo2@gmail.com, [4]golazoclaraefen08@gmail.com, [5]dzulfahmy5151@gmail.com, [6]dosen02359@unpam.ac.id

Abstract

Stroke is a non-communicable disease and one of the leading causes of death and disability worldwide. Early detection of potential stroke risk is crucial to support effective prevention and management efforts. This study aims to develop a stroke risk classification system using the K-Nearest Neighbor (KNN) algorithm implemented through the RapidMiner platform. The dataset analyzed consists of 932 patient records with various medical and demographic attributes. The research process includes data preprocessing, variable transformation, normalization, and splitting the data into training and testing sets. Model evaluation shows an accuracy rate of 82.35%; however, the model has not performed well in identifying stroke cases due to data imbalance. These findings highlight the importance of addressing class imbalance in medical data and the need to consider alternative algorithms to improve detection of minority classes.

Keywords: *Stroke, K-Nearest Neighbor, KNN, RapidMiner, Data Mining*

Abstrak

Stroke adalah salah satu penyakit tidak menular, penyebab utama kematian dan disabilitas pada seluruh dunia. Deteksi dini terhadap potensi risiko stroke sangat krusial untuk mendukung upaya pencegahan dan penanganan yang optimal. Penelitian ini bertujuan untuk membangun sistem klasifikasi risiko stroke menggunakan algoritma K-Nearest Neighbor (KNN) yang diimplementasikan melalui platform RapidMiner. Dataset yang dianalisis terdiri dari 932 data pasien dengan berbagai atribut medis dan demografis. Proses penelitian mencakup tahap pra-pemrosesan data, transformasi variabel, normalisasi, serta pembagian data menjadi data pelatihan dan pengujian. Evaluasi model menunjukkan tingkat akurasi sebesar 82,35%, namun model belum mampu mengidentifikasi kasus stroke dengan baik akibat ketidakseimbangan data. Temuan ini menyoroti pentingnya penanganan terhadap ketimpangan kelas dalam data medis dan perlunya pertimbangan algoritma alternatif untuk meningkatkan kemampuan deteksi terhadap kelas minoritas.

Kata Kunci: Stroke, *K-Nearest Neighbor*, KNN, *RapidMiner*, *Data Mining*

A. PENDAHULUAN

Gangguan pada aliran darah menuju otak dapat memicu stroke, suatu kondisi kesehatan serius yang berpotensi menyebabkan gangguan fungsi otak, kecacatan permanen, bahkan kematian. Menurut World Health Organization (WHO), lebih dari 15 juta kasus stroke terjadi setiap tahunnya di seluruh dunia, dan sepertiganya berakhir dengan kematian. Di Indonesia, data dari Riset Kesehatan Dasar (Riskesdas) menunjukkan adanya peningkatan yang signifikan dalam prevalensi stroke, menjadikannya salah satu penyebab utama kematian dan disabilitas jangka panjang.

Seiring dengan kemajuan di bidang teknologi informasi, pendekatan berbasis data menjadi semakin

Diterima Redaksi: 19-05-2025 | Selesai Revisi: 25-06-2025 | Diterbitkan Online: 30-05-2025

penting dalam dunia medis, khususnya dalam mendukung proses pengambilan keputusan klinis. Salah satu metode yang banyak digunakan adalah klasifikasi risiko stroke berdasarkan riwayat medis dan karakteristik pasien. Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode dalam *supervised learning* yang digunakan untuk mengklasifikasikan individu berdasarkan kemiripannya dengan data historis. Algoritma ini bekerja dengan cara mengidentifikasi sejumlah tetangga terdekat dari data baru untuk menentukan kategorinya, seperti apakah pasien berisiko mengalami stroke atau tidak.

Dalam penelitian ini, dilakukan implementasi dan perbandingan algoritma KNN dengan dua pendekatan berbeda, yaitu menggunakan RapidMiner sebagai platform berbasis antarmuka grafis (GUI) dan Jupyter Notebook sebagai lingkungan pemrograman berbasis Python. Tujuan dari perbandingan ini adalah untuk menilai efektivitas masing-masing pendekatan dalam hal akurasi, efisiensi proses, serta kemudahan interpretasi hasil oleh tenaga medis. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan medis berbasis data mining, khususnya dalam upaya deteksi dini risiko stroke.

B. PELAKSAAAN DAN METODE

Dataset yang digunakan dalam studi ini diperoleh dari situs Kaggle, yang berisi informasi tentang pasien termasuk usia, jenis kelamin, kondisi kesehatan, dan kebiasaan hidup. Sebanyak 932 data pasien dianalisis sebagai dasar untuk klasifikasi risiko stroke. Atribut yang digunakan meliputi: usia (dalam tahun), jenis kelamin, riwayat hipertensi, penyakit jantung, status perkawinan, jenis pekerjaan, jenis tempat tinggal (perkotaan atau pedesaan), tingkat rata-rata glukosa dalam darah, indeks massa tubuh (BMI), status merokok, dan label klasifikasi stroke sebagai variabel target (0 = tidak ada stroke, 1 = stroke). Penelitian dilakukan melalui berbagai tahap sebagai berikut:

1. **Pra-pemrosesan Data**
   Tahapan ini dimulai dengan menghapus atribut yang tidak relevan, seperti kolom ID. Untuk menangani nilai yang hilang pada atribut indeks massa tubuh (BMI), dilakukan imputasi dengan menggunakan nilai rata-rata, sehingga kualitas dan konsistensi data tetap terjaga selama proses pengolahan. Selain itu, data kategorikal seperti jenis kelamin, jenis pekerjaan, tipe tempat tinggal, dan status merokok dikonversi ke bentuk numerik menggunakan metode one-hot encoding, sehingga dapat diproses secara optimal oleh algoritma KNN.

2. **Normalisasi**
   Karena KNN sangat bergantung pada pengukuran jarak antar data, penting untuk menyamakan skala antar fitur. Oleh karena itu, normalisasi diterapkan dengan pendekatan Z-Score, di mana fitur numerik disesuaikan ke skala distribusi standar agar seluruh atribut memiliki bobot perhitungan yang seimbang, guna menghindari dominasi fitur tertentu dalam perhitungan jarak.

3. **Pembagian Dataset**
   Setelah menyelesaikan proses praproses dan normalisasi, data kemudian dibagi menjadi dua kelompok, dengan 80% untuk pelatihan dan 20% untuk pengujian. Data pelatihan digunakan untuk membangun model, sementara data uji digunakan untuk mengevaluasi sejauh mana model dapat memprediksi data baru yang belum pernah ditemui sebelumnya.

4. **Penerapan Algoritma KNN**
   Model klasifikasi dibangun menggunakan algoritma K-Nearest Neighbor dengan nilai k = 5. Prediksi kelas ditentukan berdasarkan lima tetangga terdekat menggunakan jarak Euclidean. Implementasi dilakukan dengan RapidMiner melalui operator k-NN yang telah disesuaikan pengaturannya.

5. **Evaluasi Model**
   Kinerja model dievaluasi menggunakan metrik akurasi, yang menunjukkan persentase prediksi yang benar dibandingkan dengan total prediksi yang dibuat. Selain itu, matriks kebingungan digunakan untuk menilai kemampuan model dalam membedakan antara pasien stroke dan non-stroke dengan menganalisis nilai True Positives, False Positives, True Negatives, dan False Negatives.

C. HASIL DAN PEMBAHASAN

Setelah model dilatih menggunakan algoritma K-Nearest Neighbor (KNN) di RapidMiner, dilakukan pengujian terhadap 20% data uji dari keseluruhan dataset. Hasilnya, model menunjukkan akurasi sebesar 82,35%, yang secara umum terlihat cukup baik. Namun, evaluasi lebih lanjut menggunakan confusion matrix mengungkapkan bahwa model sama sekali tidak berhasil mengidentifikasi pasien stroke dengan benar. Hal ini dibuktikan dengan nilai True Positive (TP) satu positif palsu (FP) sebesar 0 dan satu negatif palsu (FN) sebesar 28.
Situasi ini menunjukkan bahwa seluruh data pasien yang seharusnya diklasifikasikan sebagai stroke justru diprediksi sebagai non-stroke oleh sistem. Sementara data non-stroke berhasil diklasifikasikan dengan cukup baik, model sepenuhnya gagal dalam menangani kelas minoritas (stroke). Meskipun nilai akurasi cukup tinggi, hasil ini dapat menyesatkan apabila hanya dilihat secara sepintas tanpa memahami distribusi kelas dan nilai metrik lain seperti recall dan F1-score, yang dalam kasus ini bernilai **0** untuk kelas stroke**.**

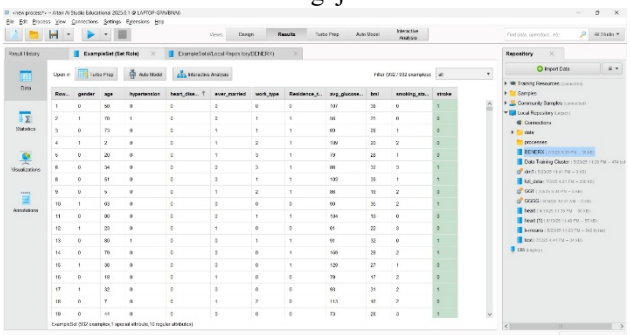
Kelemahan model ini terutama disebabkan oleh distribusi kelas yang tidak seimbang dalam dataset, di mana jumlah pasien non-stroke jauh lebih banyak dibandingkan dengan pasien stroke. Ketidakseimbangan ini menyebabkan algoritma KNN—yang bekerja berdasarkan kedekatan

antar data—lebih cenderung mengelompokkan data ke dalam kelas yang paling dominan. Dalam dunia medis, kondisi seperti ini menjadi masalah krusial karena individu yang sebenarnya berisiko tinggi terkena stroke justru berpotensi tidak teridentifikasi oleh sistem.

Kondisi ini mengindikasikan bahwa akurasi saja tidak cukup sebagai indikator utama dalam proyek klasifikasi, khususnya di ranah medis. Diperlukan evaluasi tambahan dengan metrik lain seperti precision, recall, specificity, dan AUC-ROC agar kinerja model dapat dinilai secara lebih menyeluruh. Selain itu, dibutuhkan pendekatan lanjutan untuk menangani ketidakseimbangan kelas, misalnya melalui teknik resampling, penyesuaian ambang klasifikasi, atau pemanfaatan algoritma yang lebih tangguh terhadap distribusi data yang tidak seimbang, seperti Random Forest atau XGBoost.

Dataset

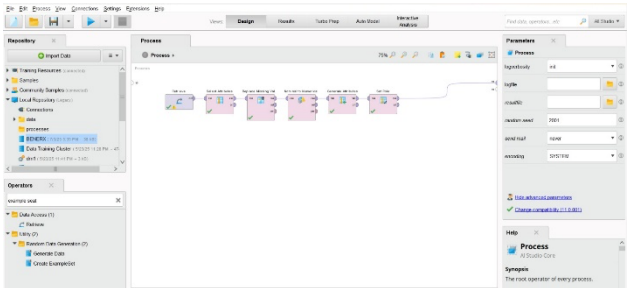
Dataset dalam penelitian ini berasal dari platform Kaggle, yang menyediakan data terkait pasien untuk tujuan analisis, yang terdiri dari 932 data pasien. Dataset ini mencakup berbagai atribut seperti usia, jenis kelamin, riwayat hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, jenis tempat tinggal, rata-rata kadar glukosa darah, BMI, gejala merokok, dan label klasifikasi yang menunjukkan apakah pasien mengalami stroke atau tidak. Data ini dianggap tidak seimbang karena jumlah pasien non-stroke jauh lebih besar dibandingkan dengan jumlah pasien stroke, yang menciptakan tantangan dalam proses klasifikasi. Dataset ini kemudian digunakan untuk melatih dan menguji model KNN.



Gambar 1. Dataset Pre-processing Data Pada RapidMiner

Implementasi Model KNN

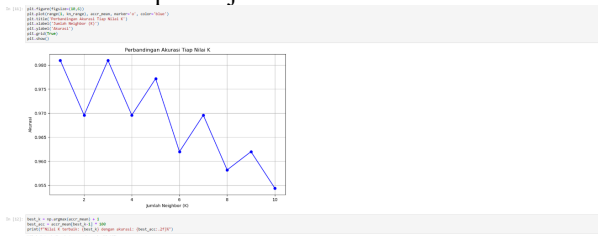
Model K-Nearest Neighbor diimplementasikan menggunakan RapidMiner. Tahapan yang dilakukan meliputi penghapusan atribut id, imputasi nilai kosong pada bmi dengan nilai rata-rata, transformasi data kategorikal menggunakan one-hot encoding, serta normalisasi fitur numerik menggunakan Z-Score. Data kemudian dibagi 80:20 untuk pelatihan dan pengujian. Model dijalankan dengan parameter k = 5, menggunakan *Euclidean Distance* sebagai metode perhitungan jarak. Proses voting dilakukan menggunakan pendekatan berbobot.



Gambar 2. Diagram Alur Proses Klasifikasi KNN di RapidMiner

Processing dan Visualisasi (Chart)

Untuk mengeksplorasi performa model dengan variasi parameter *k*, dilakukan eksperimen tambahan di Jupyter Notebook menggunakan library scikit-learn. Nilai k dicoba dari 1 hingga 11 dan hasilnya diplot dalam grafik. Ditemukan bahwa akurasi tertinggi dicapai pada nilai k = 10, dengan akurasi mencapai 95,79% pada data pelatihan. Grafik ini memberikan gambaran penting tentang pengaruh nilai k terhadap kinerja klasifikasi.



Gambar 3. Grafik Akurasi terhadap Nilai k pada Jupyter Notebook

D. PENUTUP

Kesimpulan

Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi risiko stroke berdasarkan algoritma K-Nearest Neighbor (KNN) menggunakan RapidMiner. Dataset yang digunakan berasal dari Kaggle, yang terdiri dari 932 entri pasien dengan atribut demografi dan medis. Melalui tahap pra-pemrosesan, normalisasi, dan pelatihan model, ditemukan bahwa model KNN dengan k = 5 mencapai akurasi sebesar 82,35%.

Meskipun akurasi tampak cukup tinggi, hasil evaluasi melalui confusion matrix menunjukkan bahwa model gagal mengidentifikasi pasien stroke secara benar (True Positive = 0). Hal ini disebabkan oleh ketidakseimbangan data (class imbalance) antara pasien stroke dan non-stroke dalam dataset, di mana jumlah data non-stroke mendominasi. Kondisi ini menyebabkan model hanya mampu mengenali

kelas mayoritas, dan mengabaikan kelas minoritas yang justru krusial dalam konteks medis.

Penelitian ini menegaskan bahwa akurasi bukan satu-satunya indikator performa dalam klasifikasi medis. Dalam konteks klasifikasi medis, akurasi tidak cukup menggambarkan performa model. Oleh karena itu, perlu digunakan metrik tambahan seperti sensitivitas, keseimbangan antara precision dan recall (F1-score), serta area di bawah kurva ROC. dalam mengenali kasus kritis seperti stroke. Dengan demikian, sistem klasifikasi yang dibangun belum dapat diandalkan dalam aplikasi nyata tanpa penyesuaian lebih lanjut.

Saran

1. Penanganan Ketidakseimbangan Kelas:
   Penting untuk mengatasi ketimpangan distribusi kelas dalam dataset. Salah satu solusinya adalah dengan menambahkan data sintetis pada kelas minoritas menggunakan metode seperti SMOTE, atau dengan mengurangi jumlah data dari kelas mayoritas. Pendekatan ini sangat dianjurkan guna meningkatkan kemampuan model dalam mendeteksi pasien yang berisiko stroke.
2. Penggunaan Algoritma Tambahan:
   Peneliti selanjutnya dapat membandingkan performa KNN dengan algoritma lain yang lebih kompleks dan robust terhadap class imbalance, seperti Random Forest, XGBoost, atau Support Vector Machine (SVM).
3. Peningkatan Evaluasi Model:
   Selain akurasi, evaluasi model harus mencakup metrik lain seperti presisi, recall, spesifisitas, dan F1-score untuk membuat evaluasi kinerja model lebih komprehensif dan terkontekstualisasi dengan kebutuhan medis.
4. Pengembangan Sistem Berbasis Web atau Aplikasi:
   Untuk meningkatkan dampak penerapan, hasil klasifikasi dapat dikembangkan menjadi sistem pendukung keputusan medis berbasis web atau mobile, sehingga dapat digunakan secara praktis oleh tenaga kesehatan.
5. Validasi dengan Data Medis Nyata:
   Disarankan untuk menguji model menggunakan data medis dari institusi kesehatan lokal agar hasilnya lebih representatif terhadap populasi di

E. DAFTAR PUSTAKA

1. Abiodun OJ dan Wreford AI. 2023. Stroke prediction using SMOTE for data balancing, XGBoost and KNN ensemble algorithms. *J. Appl. Phys. Sci. Int.* **15**(1): 42–53. https://doi.org/10.56557/JAPSI/2023/v15i18349
2. Asadi F dan Rahimi M. 2024. The most efficient machine learning algorithms in stroke prediction: A systematic review. *Health Sci. Rep.* **7**: e70062. https://doi.org/10.1002/hsr2.70062
3. Lavanya, J. M., & Subbulakshmi, P. (2024). Unveiling the potential... *Scientific Reports*, 14, 20053. https://doi.org/10.1038/s41598-024-70354-1
4. Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing. 2024. *BMC Bioinform.* **25**: 58. https://doi.org/10.1186/s12859-024-05866-8
5. A novel approach to predict brain stroke using KNN in machine learning. 2023. Dalam: *Proc. Int. Conf. Adv. Intell. Interact. Hum.-Comput. Interfaces (ICAIIHI)*. https://doi.org/10.1109/ICAIIHI57871.2023.10489301