seminar nasional 2024 UNIVERSITAS MA CHUNG

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 241-248, 2024

ISSN: 2808-7771 online

Prediksi Penyakit Jantung Koroner Menggunakan Metode K-NN dan Regresi Logistik Berdasarkan Kerangka Kerja CRISP-DM

Kwandy Chandra[1] dan Juan Sebastian Prasetyo[2]

[1,2] Program Studi Sistem Informasi Bisnis, Universitas Ciputra Surabaya
CitraLand CBD Boulevard, Surabaya, Indonesia, 60219

Korespondensi: Kwandy Chandra (kchandra05@student.ciputra.ac.id)

Received: 24 Juli 2024 – *Revised*: 31 Agustus 2024 - *Accepted*: 05 Sept 2024 - *Published*: 10 Sept 2024

Abstrak. Penyakit jantung koroner merupakan salah satu penyebab kematian terbanyak di seluruh dunia, menyebabkan sekitar 17,8 juta kematian setiap tahun, yang mencakup sekitar 31% dari total kematian global. Tujuan dari studi ini adalah untuk mengembangkan model prediksi digital penyakit jantung koroner menggunakan algoritma K-Nearest Neighbor (K-NN) dan regresi logistik yang diimplementasikan dengan bahasa pemrograman Python. Dataset yang digunakan mencakup sekitar 300.000 data dengan 19 variabel yang meliputi faktor-faktor seperti kondisi kesehatan umum, aktivitas fisik, riwayat merokok, dan konsumsi alkohol. Metode studi ini menggunakan kerangka kerja CRISP-DM yang mencakup enam tahap: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Pada tahap persiapan data, dilakukan pembersihan dan pemetaan variabel untuk memastikan data bebas dari nilai kosong dan siap untuk pemodelan. Dua model klasifikasi diterapkan dalam studi ini, yakni K-NN dan Regresi Logistik. Model dievaluasi menggunakan *Confusion Matrix* dan berbagai metrik seperti akurasi, presisi, *recall*, dan *F1-score*. Hasilnya menunjukkan bahwa model Regresi Logistik memiliki kinerja yang lebih baik dalam memprediksi penyakit jantung koroner dibandingkan K-NN, dengan akurasi 91,8%, presisi 88%, *recall* 92%, dan *F1-score* 89%. Model K-NN menunjukkan akurasi 91,5%, presisi 88%, *recall* 92%, dan *F1-score* 89%. Implementasi metode prediksi ini diharapkan dapat mendukung deteksi dini dan pengambilan keputusan medis yang lebih tepat, sehingga mengurangi tingkat kematian akibat penyakit jantung koroner. Simpulan studi ini adalah bahwa penggunaan algoritma pembelajaran mesin, khususnya Regresi Logistik, efektif dalam memprediksi risiko penyakit jantung koroner dan dapat berkontribusi dalam menurunkan angka kematian global akibat penyakit ini. Disarankan agar studi berikutnya memasukkan variabel tambahan yang relevan untuk meningkatkan keakuratan prediksi.

Kata kunci: jantung koroner, kematian, kesehatan, *machine learning*, prediksi

Citation Format: Chandra, K., & Prasetyo, J.S. (2024). Prediksi Penyakit Jantung Koroner Menggunakan Metode K-NN dan Regresi Logistik Berdasarkan Kerangka Kerja CRISP-DM. *Prosiding SENAM 2024: Seminar Nasional Sistem Informasi & Informatika Universitas Ma Chung*. 4, 241-248. Malang: Ma Chung Press.

PENDAHULUAN

Menurut World Health Organization (2020), penyakit jantung koroner merupakan salah satu penyebab utama kematian di dunia, menyebabkan sekitar 17,8 juta kematian per tahun atau sekitar 31% dari total kematian global. Penyakit ini umumnya terjadi di negara-

negara berpendapatan menengah, dan dari seluruh kasus penyakit jantung koroner, 85% disebabkan oleh serangan jantung dan stroke.

Mengingat tingginya angka kematian akibat penyakit jantung koroner, kami memulai proyek untuk mengembangkan alat prediksi digital menggunakan algoritma K-Nearest Neighbors (K-NN) dan *Logistic Regression* berbasis bahasa pemrograman Python. Alat ini diharapkan dapat membantu dalam deteksi dini dan pengambilan keputusan medis yang lebih tepat. Penggunaan machine learning dalam analisis data medis telah menjadi fokus studi yang populer, dengan berbagai algoritma yang telah diaplikasikan dalam bidang ini (Latif *et al*., 2019).

Tujuan dari proyek ini adalah untuk membuat model prediksi penyakit jantung koroner menggunakan algoritma K-NN dan *Logistic Regression*, serta mengevaluasi dan membandingkan kinerja kedua model dalam memprediksi penyakit tersebut. Kami bertujuan menentukan model yang memiliki akurasi dan reliabilitas terbaik untuk membantu dalam diagnosis penyakit jantung koroner. Model yang paling efektif akan digunakan sebagai alat bantu bagi staf medis dalam deteksi dini dan pengambilan keputusan terkait perawatan pasien.

Studi ini diharapkan dapat memberikan kontribusi berarti dalam bidang kedokteran, khususnya dalam upaya pencegahan dan pengobatan penyakit jantung koroner. Mitra dalam kegiatan ini adalah komunitas medis yang merawat pasien dengan risiko penyakit jantung koroner, dan target masyarakatnya adalah individu dengan risiko tinggi, termasuk mereka yang memiliki riwayat kesehatan buruk dan gaya hidup tidak sehat. Dengan mengembangkan alat prediksi yang akurat dan dapat diandalkan, diharapkan dapat mengurangi angka kematian akibat penyakit ini melalui intervensi medis yang lebih tepat waktu dan efektif.

MASALAH

Penyakit jantung koroner merupakan masalah kesehatan masyarakat yang serius di seluruh dunia, terutama di negara-negara berkembang. Tingginya angka kematian akibat penyakit ini menunjukkan tantangan besar dalam upaya pencegahan dan pengobatan yang efektif. Beberapa masalah utama yang dihadapi meliputi kurangnya deteksi dini, keterbatasan akses terhadap perawatan medis yang memadai, dan kurangnya alat prediksi yang efektif untuk diagnosis awal.

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 241-248, 2024
ISSN: 2808-7771 online

Masyarakat, terutama di wilayah dengan sumber daya kesehatan yang terbatas, sering kali mengalami kesulitan dalam mengakses layanan kesehatan yang memadai. Banyak individu yang berisiko tinggi terhadap penyakit jantung koroner tidak mendapatkan diagnosis dini yang akurat, sehingga intervensi medis sering kali terlambat. Keterlambatan ini berkontribusi pada tingginya angka kematian dan morbiditas akibat penyakit jantung koroner.

Dari perspektif ilmiah, tantangan utama adalah mengembangkan model prediksi yang akurat dan dapat diandalkan untuk mendeteksi penyakit jantung koroner pada tahap awal. Meskipun telah banyak studi dilakukan, masih ada kebutuhan signifikan untuk meningkatkan akurasi dan keandalan model prediksi. Algoritma seperti K-Nearest Neighbors (K-NN) dan *Logistic Regression* perlu dievaluasi dan ditingkatkan terus-menerus untuk memastikan hasil yang paling akurat (Nusinovici *et al*., 2020).

Ada kebutuhan mendesak untuk alat prediksi yang dapat digunakan oleh komunitas medis untuk memberikan diagnosis cepat dan akurat. Alat ini harus mudah diakses dan digunakan oleh tenaga medis di berbagai tingkat, terutama di daerah dengan akses terbatas terhadap fasilitas medis canggih. Selain itu, penting untuk menyelenggarakan program edukasi dan pelatihan bagi tenaga medis dan masyarakat tentang pentingnya deteksi dini dan pencegahan penyakit jantung koroner. Dengan mengatasi masalah-masalah ini melalui pengembangan alat prediksi berbasis algoritma K-NN dan *Logistic Regression*, diharapkan dapat mengurangi angka kematian akibat penyakit jantung koroner dan meningkatkan kualitas hidup masyarakat yang berisiko tinggi.

METODE PELAKSANAAN

Metode pelaksanaan dalam studi ini menggunakan kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*). CRISP-DM adalah metodologi yang umum digunakan dalam proyek-proyek data mining, terdiri dari enam fase utama (Chapman *et al*., 2000). Tahapan fase tersebut terdiri atas tahap pemahaman bisnis, tahap pemahaman data, tahap persiapan data, tahap pemodelan, tahap evaluasi, dan tahap implementasi. Berikut adalah penjelasan masing-masing tahap dalam konteks studi ini:

1. Pemahaman Bisnis (*Business Understanding*)

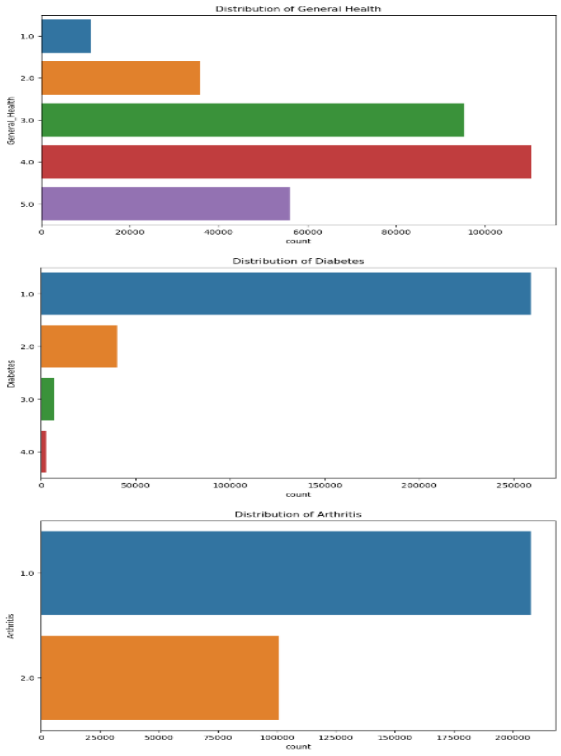
   Penyakit jantung koroner atau biasa disebut arteri koroner adalah suatu kondisi di mana pembuluh darah jantung (arteri koroner) tersumbat oleh timbunan lemak. Menurut data WHO, 17,9 juta orang di seluruh dunia meninggal karena penyakit

jantung dan pembuluh darah (penyakit kardiovaskular), termasuk penyakit jantung koroner (PJK), pada tahun 2019. Sementara di Indonesia, tercatat lebih dari 2 juta orang menderita penyakit kardiovaskular pada tahun 2018.

Tujuan utama dari fase ini adalah untuk memahami tujuan dan kebutuhan bisnis dari proyek ini. Dalam hal ini, fokusnya adalah pada pengembangan alat prediksi penyakit jantung koroner yang dapat digunakan oleh komunitas medis untuk deteksi dini dan pengambilan keputusan medis. Kebutuhan masyarakat yang mendesak untuk alat prediksi yang akurat dan mudah digunakan telah diidentifikasi.

2. Pemahaman Data (*Data Understanding*)

Pada fase ini, data yang relevan dikumpulkan dan dipahami. Data yang digunakan mencakup berbagai faktor risiko yang berkontribusi terhadap penyakit jantung koroner seperti tekanan darah, kadar kolesterol, indeks massa tubuh, riwayat kesehatan, dan lain-lain. Analisis eksplorasi data dilakukan untuk memahami distribusi dan karakteristik data. Pada studi yang akan dibuat, kami menggunakan dataset prediksi risiko penyakit jantung koroner dengan jumlah data sebanyak 309 ribu data yang memiliki 19 atribut / variabel yang akan dijabarkan di tabel dibawah ini.



Gambar 1. Distribusi Data

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 241-248, 2024
ISSN: 2808-7771 online

3. Persiapan Data (*Data Preparation*)

   Data yang telah dikumpulkan dibersihkan dan diproses untuk memastikan kualitas dan konsistensi data. Proses ini meliputi penghapusan data yang tidak lengkap, penanganan nilai yang hilang, normalisasi data, dan seleksi fitur yang relevan. Di tahap ini kami memulai dengan pengecekan dan pembersihan data untuk menghilangkan data yang memiliki nilai kosong dengan tujuan agar menghasilkan hasil prediksi yang lebih akurat. pembersihan data ini kami lakukan dengan menggunakan bahasa pemrograman *python* yang dijalankan di *Jupyter Notebook*. Selanjutnya kami melakukan pemetaan terhadap variabel yang ada di dalam dataset kami, hal ini bertujuan untuk mengubah data yang bersifat non-numerik menjadi numerik, dengan demikian proses prediksi dapat berjalan dengan lancar

   Sebelum menentukan variabel yang akan digunakan dalam model prediksi, dilakukan analisis korelasi untuk melihat hubungan antara variabel dalam dataset. Korelasi yang kuat antara variabel dapat membantu dalam memilih variabel prediktor yang relevan. Data kemudian dibagi menjadi set pelatihan dan set pengujian untuk keperluan pemodelan.

4. Pemodelan (*Modeling*)

   Pada fase ini, algoritma *K-Nearest Neighbors (K-NN)* dan *Logistic Regression* digunakan untuk membangun model prediksi penyakit jantung koroner.

   Algoritma *K-Nearest Neighbors* (K-NN) digunakan dalam studi ini karena kemampuannya dalam mengklasifikasikan data berdasarkan kedekatan dengan data lain yang sudah diketahui kategorinya (Javatpoint, n.d.).

   Logistic Regression merupakan metode yang sangat berguna dalam analisis data medis karena kemampuannya dalam memodelkan hubungan antara variabel independen dan probabilitas hasil biner, yang sering digunakan dalam prediksi dan klasifikasi data medis (Hosmer *et al*., 2013).

5. Evaluasi (*Evaluation*)

   Model yang telah dibangun dievaluasi untuk memastikan bahwa model memenuhi tujuan bisnis dan kinerja yang diharapkan. Evaluasi dilakukan dengan membandingkan kinerja model pada set pengujian. Model yang dihasilkan dievaluasi menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan skor F1. Hasil dari evaluasi dapat digunakan untuk membuat keputusan akhir tentang pemilihan model terbaik yang lebih akurat yang akan digunakan.

6. Implementasi (*Deployment*)

   Model terbaik kemudian diimplementasikan sebagai alat prediksi yang dapat digunakan oleh komunitas medis. Proses implementasi juga mencakup pengembangan dokumentasi, pelatihan untuk pengguna akhir, dan pemantauan berkelanjutan untuk memastikan kinerja model tetap optimal.

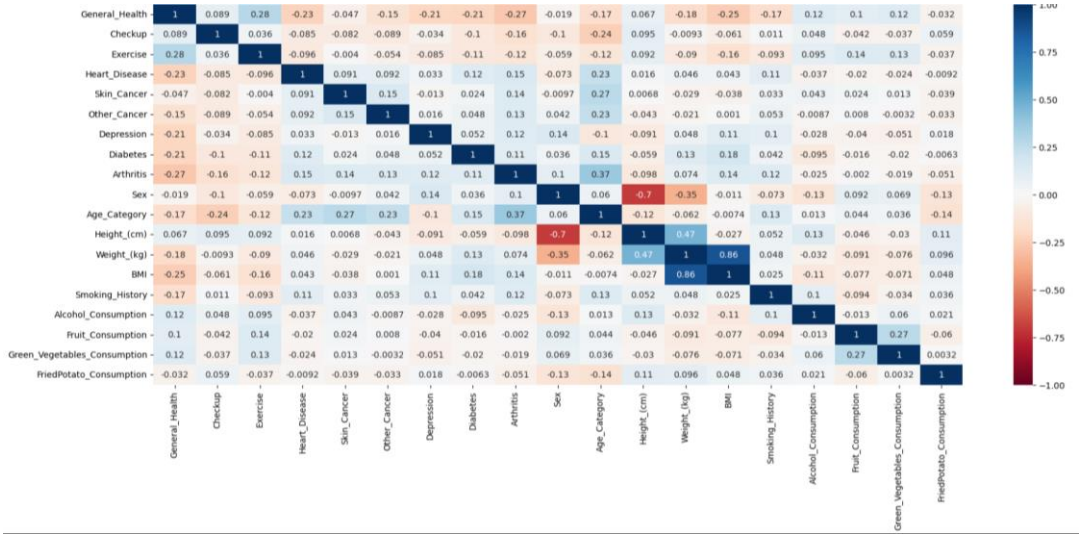
Data dikumpulkan dari sumber-sumber yang telah diidentifikasi yaitu *Kaggle*, termasuk data medis pasien dengan risiko penyakit jantung koroner. Pengumpulan data dilakukan dengan mengacu pada standar etika dan privasi data yang ketat. Data dianalisis menggunakan teknik statistik dan *machine learning* untuk mengidentifikasi pola dan hubungan antara berbagai faktor risiko dan kejadian penyakit jantung koroner. Analisis ini mencakup teknik eksplorasi data, visualisasi data, dan pemodelan prediktif.

HASIL DAN PEMBAHASAN

Perlu Setelah melakukan studi, kami menemukan beberapa hal yang merupakan hasil dari beberapa tahapan-tahapan CRISP-DM yang telah dilakukan. Pada bagian ini, temuan ini akan diuraikan dan dibahas dengan seksama sebagai bahan diskusi.

Korelasi Data

Pada tahap Persiapan Data, dari hasil korelasi yang ditemukan, kami melihat bahwa variabel yang memiliki hubungan kuat dengan penyakit jantung adalah *General_Health*, *Checkup*, *Exercise*, *Diabetes*, dan *Smoking_History*. Korelasi ini digunakan untuk membangun model prediksi.



Gambar 2. Korelasi *Heatmap* Antar Data

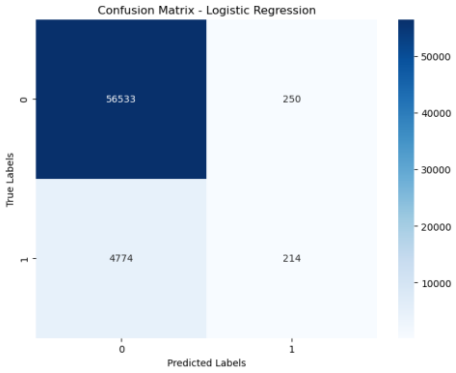
Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 241-248, 2024
ISSN: 2808-7771 online

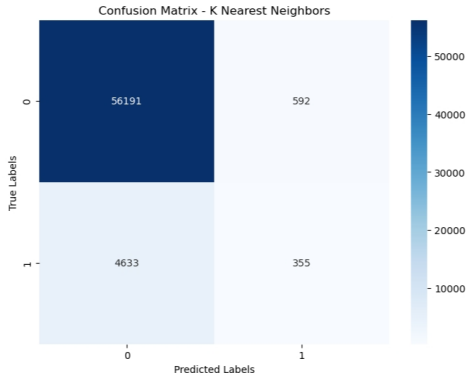
Evaluasi Model

Evaluasi model dilakukan dengan menggunakan *Confusion Matrix* dan berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan skor F1.

Hasil evaluasi kedua model adalah sebagai berikut:

1. *Confusion Matrix*:
    - *Logistic Regression*: Menunjukkan performa yang lebih baik dalam mengklasifikasikan data dibandingkan K-NN.
    - K-NN: Memiliki beberapa kesalahan klasifikasi yang lebih tinggi dibandingkan *Logistic Regression.*





Gambar 3. *Confusion Matrix Logistic Regression* **Gambar 4.** *Confusion Matrix K-NN*

2. Metrik Evaluasi:
    - *Logistic Regression* memiliki akurasi 91,8%, presisi 88%, *recall* 92%, dan skor *F1* 89%.
    - K-NN memiliki akurasi 91,5%, presisi 88%, *recall* 92%, dan skor *F1* 89%.

Analisis Perbandingan

Perbedaan utama antara kedua model adalah cara mereka mengklasifikasikan data. K-NN mengandalkan kedekatan data baru dengan data dalam dataset pelatihan, sedangkan *Logistic Regression* menggunakan hubungan linier antara variabel prediktor dan hasil. Penerapan metode klasifikasi *K-Nearest Neighbor* (K-NN) pada dataset penderita penyakit diabetes menunjukkan bahwa algoritma ini mampu memberikan hasil klasifikasi yang baik berdasarkan kedekatan dengan data yang sudah ada (Argina, 2020). Namun, pada kasus studi ini, *Logistic Regression* lebih unggul karena mampu menangkap hubungan linier yang kuat dalam data, sedangkan K-NN lebih cocok untuk data dengan pola yang lebih kompleks dan tidak linier.

KESIMPULAN

Prediksi ini mengimplementasikan algoritma *machine learning*, yaitu K-NN dan *Logistic Regression*, untuk memprediksi penyakit jantung koroner. Hasil menunjukkan bahwa *Logistic Regression* lebih akurat dalam memprediksi penyakit ini dibandingkan K-NN. Model *Logistic Regression* memiliki akurasi 91.8 %, sementara K-NN memiliki akurasi 91.5 %. Oleh karena itu, *Logistic Regression* disimpulkan sebagai model yang lebih unggul. Rekomendasi untuk studi selanjutnya adalah agar dapat mempertimbangkan untuk menambahkan variabel lain yang relevan dengan penyakit jantung koroner untuk meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

Argina, A. M. (2020). Penerapan metode klasifikasi K-Nearest Neighbor pada dataset penderita penyakit diabetes. *Jurnal Ilmiah Dasar, 1*(2), 29–33. https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/11/14

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. [Publisher not listed].

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9781118548387

JavatPoint. (n.d.). *K-Nearest neighbor (KNN) algorithm for machine learning*. www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Latif, J., Xiao, C., Imran, A., & Tu, S. (2019). Medical imaging using *machine learning* and deep learning algorithms: A review. In *Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. https://doi.org/10.1109/icomet.2019.8673502

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. (2020). *Logistic Regression* was as good as *machine learning* for predicting major chronic diseases. *Journal of Clinical Epidemiology, 122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

World Health Organization. (2021, June 11). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)