Aplikasi K-Nearest Neighbor (KNN) untuk Klasifikasi Penyakit *Kardiovaskuler*

Sri Sumarlinda[1], Wiji Lestari[2]

Sistem Informasi, Universitas Duta Bangsa Surakarta
Jl. Bhayangkara no 55-57 Kota Surakarta
[1]sri_sumarlinda@udb.ac.id
[2]wiji_lestari@udb.ac.id

Abstrak— Data mining dan machine learning adalah dua alat yang memainkan peran penting dalam studi analisis data dan sistem keputusan. Klasifikasi adalah fungsi dari data mining. Dalam fungsi klasifikasi, pengurutan atau pemetaan terjadi berdasarkan kedekatan atau kesamaan atribut data dengan label yang ditentukan. Algoritma K-Nearest Neighbor (KNN) adalah metode non-parametrik yang digunakan untuk klasifikasi dan regresi. Model prediksi penyakit kardiovaskular dengan algoritma KNN digunakan untuk mengidentifikasi dan memprediksi penyakit kardiovaskular. Algoritma KNN menggunakan jarak Euclidian untuk proses prediksi data latih. Dataset yang digunakan adalah 400 dengan 7 atribut yaitu umur, jenis kelamin, tekanan darah sistolik, kolesterol, talach, oldpeak dan slope. Hasil implementasi algoritma KNN menghasilkan performansi dengan akurasi sebesar 75,75%. Nilai presisinya adalah 76,78%, sedangkan recall menghasilkan 77,14%.

***Kata kunci*— data mining, machine learning, klasifikasi, kardiovaskuler, K-Nearest Neighbor**

I. PENDAHULUAN

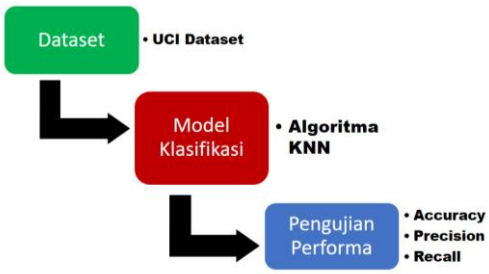
Era digital ditandai dengan peningkatan data yang cepat dan melimpah. Data sebesar itu kemudian memunculkan istilah big data. Perkembangan big data ditunjukkan dengan volume, kecepatan, kebenaran, variabilitas, keragaman dan sebagainya. Perkembangan kuantitas juga melahirkan studi-studi baru seperti data science, data engineering, data analysis, dan sebagainya. Kajian pengumpulan data merupakan kajian yang menarik dan telah banyak dilakukan. Penambangan data dan pembelajaran mesin adalah dua alat yang memainkan peran penting dalam studi analisis data dan sistem keputusan. Kumpulan data tersebut setelah diolah dengan data mining dan machine learning akan menghasilkan informasi dan kemudian sebuah keputusan [1]. Ada tiga metode pembelajaran untuk menganalisis data dengan machine learning, yaitu pembelajaran terawasi, pembelajaran tak terawasi, dan pembelajaran penguatan. Ketiga metode pembelajaran tersebut erat kaitannya dengan berbagai fungsi dalam data mining. Pembelajaran terawasi berkaitan dengan prediksi, peramalan, dan klasifikasi. Sedangkan unsupervised learning terkait dengan fungsi clustering [2]. Klasifikasi merupakan fungsi pada data mining.

Pada fungsi klasifikasi, pengurutan atau pemetaan terjadi berdasarkan kedekatan atau kemiripan atribut data dengan label yang ditentukan. Dalam klasifikasi, pelatihan biasanya diperlukan sebagai uji kinerja algoritma sebelum diimplementasikan dengan pengujian data. Salah satu algoritma yang cukup populer untuk klasifikasi dan pengenalan pola adalah K-Nearest Neighbor (KNN). Algoritma KNN merupakan metode non-parametrik yang digunakan untuk klasifikasi dan regresi [3]. Prinsip KNN adalah memiliki sekumpulan data sampel sebagai training set yang dihubungkan dengan label sehingga dapat diketahui kedekatan data tersebut [2], [4]. Jika data baru diberikan tanpa label, maka data tersebut dapat ditentukan label/kelasnya. Dari kemiripan, tetangga terdekat, dapat ditentukan label data baru [5], [6].

Penelitian ini mengimplementasikan algoritma K-Nearest Neighbor (KNN) untuk prediksi kerentanan penyakit kardiovaskular. Penyakit kardiovaskular merupakan masalah global dan penyebab utama kematian dan kecacatan. Prediksi dan klasifikasi berdasarkan faktor risiko penting untuk pencegahan dan pengobatan dini. KNN merupakan algoritma yang banyak digunakan untuk prediksi dan klasifikasi penyakit kardiovaskular. Prediksi penyakit kardiovaskular menggunakan KNN dengan parameter yang dibandingkan antara 8 dan 13 parameter [7]. KNN dikombinasikan dengan algoritma genetika untuk prediksi penyakit kardiovaskular [8]. Extended KNN digunakan untuk memprediksi penyakit jantung dengan 11 parameter [9]. Heart Disease Prediction System (HDPS) dikembangkan menggunakan algoritma Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest Classifier, dan Support Vector Machine untuk memprediksi tingkat risiko penyakit jantung[10]. Algoritma KNN dibandingkan dengan algoritma random forest, Decision tree dan Bayesian untuk prediksi penyakit kardiovaskular [11], [12], [13]. KNN berbasis web digunakan untuk memprediksi penyakit jantung [14]. KNN digunakan untuk mengidentifikasi faktor risiko utama penyakit kardiovaskular [15].

II. METODE PENELITIAN

Penelitian ini dilakukan dengan menggunakan implementasi algoritma K-Nearest Neighbor dalam prediksi dan identifikasi penyakit kardiovaskular. Dataset penelitian sebagai data pelatihan berasal dari UCI Machine Learning Repository. Algoritma KNN digunakan dengan nilai K = 5 dengan ukuran jarak Euclidian.. Pengujian kinerja KNN menggunakan akurasi, presisi, dan recall. Tahapan penelitian seperti yang ditunjukkan pada Gambar 1 di bawah ini:



Gambar 1. Tahapan Penelitian

A. *Dataset Penyakit Kardiovaskuler*

Kumpulan data penyakit jantung yang dipilih untuk proyek ini berasal dari UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/heart+disease.
Dataset terdiri dari 400 data individu, yang menggambarkan faktor kesehatan individu dan diagnosis penyakit jantung mereka. 7 faktor untuk parameter prediktif dan cardio untuk label. 8 faktor kesehatan dalam kumpulan data yang digunakan dalam proyek ini diuraikan di bawah ini.

1. **Age** – umur
2. **Sex** – jenis kelamin
   - 0 wanita
   - 1 laki-laki
3. **Sistolic Blood Pressure** – tekanan darah
4. **Cholesterol** – Kholesterol
5. **Thalach** – denyut nadi maksimal
6. **Oldpeak** – ST depression induced by exercise relative to rest
7. **Slope** – gradient dari puncak ST segment
   - 1 upsloping
   - 2 flat
   - 3 downsloping
8. **Cardio** – diagnosis penyakit kardiovaskuler
   - 0 sehat
   - 1 kardiovaskuler

B. *Algoritma K-Nearest Neighbor (KNN)*

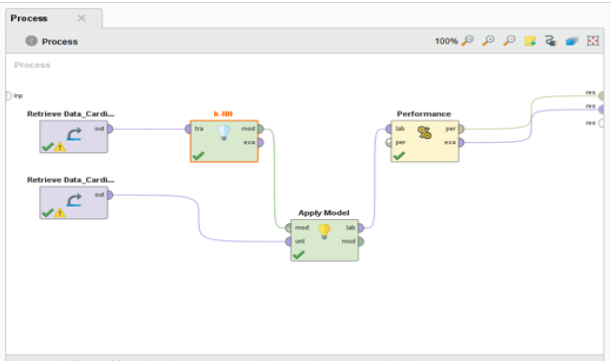
Algoritma K-Nearest Neighbor (KNN) merupakan algoritma dengan pembelajaran terawasi dan banyak digunakan untuk prediksi dan klasifikasi. Kelebihan dari algoritma KNN adalah akurasi yang tinggi, intensif pada outlier dan tidak ada asumsi tentang data. Menentukan nilai K menjadi penting. Kesamaan data dengan label digunakan jarak Euclidian dengan rumus :

$$d(x,y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2} \quad (1)$$

di sini.
d(x,y) jarak data x dan y
x_i sedang melatih data itu
y_i sedang menguji data itu
Tahapan dari algoritma KNN adalah:
• Tentukan nilai k
• Hitung jarak dengan jarak Euclidian
• Menyortir data pelatihan
• Tentukan kelas

III. HASIL DAN PEMBAHASAN

Pengembangan model klasifikasi dengan machine learning dengan algoritma KNN menggunakan software Rapidminer. Dataset menggunakan dataset UCI dengan 7 parameter yang digunakan yaitu umur, jenis kelamin, TD sistolik, kolesterol, talach, oldpeak dan slope dengan label cardio. Dataset berisi 400 data pasien yang terdiri dari penyakit kardiovaskular dan data pasien tidak sehat/tidak sehat. Implementasi algoritma KNN untuk model prediktif menggunakan Rapidminer seperti terlihat pada Gambar 2 di bawah ini:



Gambar 2. Implementasi Algoritma KNN

Implementasi algoritma KNN menggunakan K = 5, tipe pengukuran dengan MixedMeasure dan pengukuran campuran dengan Mixed Euclidian. Pengujian kinerja algoritma KNN sebagai model prediksi penyakit kardiovaskular menggunakan akurasi, presisi, dan daya ingat. Hasil pengujian kinerja implementasi algoritma KNN dan SVM seperti terlihat pada tabel 2 di bawah ini:

Table 1. Performa Algoritma KNN

Performa	KNN
akurasi	75.75%
presisi	76.78%
recall	77.14%

Dari table 1 terlihat bahwa nilai akurasi, presisi dan recall dari model klasifikasi di atas 75%. Dari hasil tersebut model klasifikasi dengan algoritma KNN cukup baik.

IV. KESIMPULAN

Model klasifikasi penyakit kardiovaskular dengan algoritma KNN digunakan untuk mengidentifikasi dan memprediksi penyakit kardiovaskular. Algoritma KNN menggunakan jarak Euclidian untuk proses prediksi data latih. Dataset yang digunakan sebanyak 400 dengan 7 atribut yaitu umur, jenis kelamin, tekanan darah sistolik, kolesterol, talach, oldpeak dan slope. Hasil implementasi algoritma KNN menghasilkan performansi dengan akurasi sebesar 75,75%. Nilai presisinya adalah 76,78%. Sedangkan recall menghasilkan 77,14% .

REFERENSI

[1] I. Hastuti, S. Purnomo, S. and W. Lestari, 2018, The Guidance of Technopreneurship Using Expert System Computing Approach Based on Entrepreneurial Values and Multiple Intelligences, *International Journal of Economics, Business and Accounting Research (IJEBAR)*, Vol.2, issue 3.

[2] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, , 2013, An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction, *Procedia - Social and Behavioral Sciences* 96 (2013) 653 – 662.

[3] D. Prasad, S.K Goyal, A. Bindal and V.S Kushwah,, 2019, System Model for Prediction Analytics Using K-Nearest Neighbors Algorithm, Journal of Computational and Theoretical Nanoscience, Vol. 16, 4425–4430, 2019.

[4] K. Alkhatib, H. Najadat, I. Hmeidi and M.K.A Shatnawi, 2013, Stock Price Prediction Using K-Nearest Neighbor (kNN) lgorithm, nternational Journal of Business, Humanities and Technology, Vol. 3 No. 3.

[5] Purwanto and D.S.S Sahid, 2021, Using KNN Algorithms for Determining the Recipient of Smart Indonesia Scholarship Program, JurnalKomputerTerapan Vol. 7, No. 2, November 2021, 163 – 173

[6] H.B Novitasari, N. Hadianto, Sfenrianto, A. Rahmawati, R. Prasetyo, J. Miharja and W. Gata, 2019, K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo), International Conference On Engineering, Technology and Innovative Researches, Journal of Physics: Conference Series 1367 (2019) 012008 IOP Publishing doi:10.1088/1742-6596/1367/1/012008.

[7] I.K.A Enriko, M. Suryanegara and D. Gunawan,, 2016, Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters, Journal of Telecommunication, Electronic and Computer Engineering, Vol. 8 No. 12

[8] M.A. Jabbar, B.I Deekshatulua and P. Chandra, 2013, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Procedia Technology 10 (2013) 85 – 94.

[9] R.S Kumar, and S.S Fatima, 2020, Heart Disease Prediction Using Extended KNN(E-KNN), International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No.5.

[10] P.K Bhunia, A. Debnath, P. Mondal, D.E Monalisa, K. Ganguly and P. Rakshit, 2021, Heart Disease Prediction using Machine Learning, International Journal of Engineering Research & Technology (IJERT), Volume 9, Issue 11.

[11] R. Hasan, 2021, Comparative Analysis of Machine Learning Algorithms for Heart DiseasePrediction, ITM Web of Conferences 40, 03007 (2021).

[12] M. Yuval, B. Yaman and O. Tosun,, 2022, Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. Mathematics 2022, 10, 311. https://doi.org/10.3390/ math10030311

[13] I. Mohit, K.S Kumar, A.U.K Reddy and B.S Kumar, 2021, An Approach to detect multiple diseases using machine learning algorithm, AMSE 2021 Journal of Physics: Conference Series 2089 (2021) 012009 IOP Publishing doi:10.1088/1742-6596/2089/1/012009.

[14] S. Kamalapurkar and S.G.H Gunjal, 2020, Web Support System for Prediction of Heart Disease using k-Nearest Neighbor Algorithm, International Journal of Computer Applications (0975 - 8887) Volume 175 - No.14.

[15] I.R Guarneros-Nolasco, N.A Cruz-Ramos, G. Alor-Hernández, L. Rodríguez-Mazahua, J.I. Sánchez-Cervantes, 2021, Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms. Mathematics 2021, 9, 2537. https:// doi.org/10.3390/math9202537.