



Jurnal Cakrawala Informasi

Journal Homepage: <http://www.itbsemarang.ac.id/sijies/index.php/jci>

e-Mail: jci@itbsemarang.ac.id



Performansi Algoritma *Decision Tree* (C4.5) untuk Prediksi Penyakit Jantung

Petra Valentino ^{1*}

Siska Narulita ²

^{1,2} Institut Teknologi dan Bisnis Semarang

INFO ARTIKEL

Histori artikel:

Diterima : 9 Desember 2022
 Revisi : 13 Desember 2022
 Disetujui : 05 Desember 2023
 Publikasi : 30 Desember 2023

Kata kunci:

Penyakit Jantung
Klasifikasi
Data Mining
Prediksi
Decision Tree
C4.5

ABSTRACT

Heart problems continue to increase and become a deadly disease. There is a need for early diagnosis of this disease, but this is very difficult to do. This is due to a lack of trained medical experts and other resources needed to carry out proper diagnosis and treatment for heart disease patients. The evaluation process using accurate prediction results regarding the risk of heart failure really helps sufferers prevent serious heart attacks and can increase the survival rate of sufferers of this disease. One of the most effective ways to identify and make predictions about heart disease is by using data mining algorithms. This research aims to create the best model for predicting heart disease using the Decision Tree (C4.5) data mining algorithm. Based on the results of the research that has been completed, it can be concluded that the best accuracy value is obtained in a prediction model that uses a comparison of training data and testing data of 90%: 10%, which produces an accuracy value of 88.35%. It is hoped that this prediction model can be a supporting tool in the diagnosis of heart disease so that serious heart attacks can be prevented and can increase the percentage of survival rates for sufferers.

ABSTRAK

Gangguan pada jantung terus meningkat dan menjadi penyakit yang mematikan. Perlunya diagnosis secara dini terhadap penyakit ini, namun hal itu sangat sulit dilakukan. Hal ini dikarenakan kurangnya tenaga ahli medis yang terlatih dan sumber daya lain yang dibutuhkan untuk melakukan diagnosis dan perawatan yang tepat bagi pasien penyakit jantung. Proses evaluasi menggunakan hasil prediksi yang akurat terhadap resiko gagal jantung sangat membantu penderita dalam mencegah serangan jantung yang parah dan dapat meningkatkan angka keselamatan dari penderita penyakit ini.

Diantara cara yang paling efektif dalam mengidentifikasi dan melakukan prediksi pada penyakit jantung adalah dengan pemanfaatan algoritma data mining. Penelitian ini bertujuan untuk membuat model terbaik untuk prediksi penyakit jantung menggunakan algoritma *data mining Decision Tree* (C4.5). Berdasarkan hasil penelitian yang telah selesai dilakukan, dapat diambil kesimpulan bahwa nilai akurasi terbaik diperoleh pada model prediksi yang menggunakan perbandingan *data training* dan *data testing* sebesar 90%:10% yang menghasilkan nilai akurasi sebesar 88,35%. Model prediksi ini diharapkan dapat menjadi alat pendukung dalam diagnosis penyakit jantung, sehingga dapat dilakukan pencegahan serangan jantung yang parah dan dapat meningkatkan persentase angka keselamatan bagi penderita.

PENDAHULUAN

Pada tahun 2015, data *World Health Organization* (WHO) menyatakan bahwa sekitar lebih dari 17 juta orang di dunia meninggal akibat penyakit jantung dan kerusakan pembuluh darah [1]. Angka kematian tersebut setara dengan 31% dari jumlah kematian yang ada di seluruh dunia [1]. Sekitar 8,7 juta orang meninggal akibat penyakit jantung [1]. Fenomena ini terjadi di negara-negara berkembang dimana rata-rata penduduknya mempunyai penghasilan menengah ke bawah [1]. Penyakit jantung itu sendiri merupakan sebutan umum untuk menggambarkan gangguan terhadap fungsi dari kinerja jantung [1]. Gangguan pada jantung atau penyakit jantung *coroner* atau *cardiovascular* (CVD) terus meningkat dan menjadi penyakit yang mematikan [2]. Penderita penyakit ini kebanyakan adalah pria terutama yang berusia paruh baya hingga tua [3]. Pasien yang didiagnosis menderita penyakit jantung, rata-rata dari setengahnya meninggal dalam kurun waktu 1-2 tahun [4].

Beberapa tes dilakukan untuk memprediksi apakah seseorang menderita penyakit jantung. Perlunya diagnosis secara dini terhadap penyakit ini, namun hal itu sangat sulit dilakukan [2]. Kurangnya tenaga ahli medis yang terlatih dan sumber daya lain yang dibutuhkan untuk melakukan diagnosis dan perawatan yang tepat bagi pasien penyakit jantung, menjadi tantangan yang sulit terutama bagi negara berkembang [5]. Proses evaluasi menggunakan hasil prediksi yang akurat terhadap resiko gagal jantung sangat membantu penderita dalam mencegah serangan jantung yang parah dan dapat meningkatkan angka keselamatan dari penderita penyakit ini [6].

Diantara cara yang paling efektif dalam mengidentifikasi dan melakukan prediksi pada penyakit jantung adalah dengan pemanfaatan algoritma *data mining*. Data mining merupakan proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis atau pembelajaran berbasis induksi (*induction based learning*), yaitu proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsep-konsep yang dipelajari [7]. Algoritma *data mining* sangat efektif ketika dilatih pada *dataset* yang sesuai [8]. Algoritma data mining mampu mengatasi kerumitan dalam diagnosis penyakit jantung dengan menggunakan model prediksi, seperti algoritma *Decision Tree*, *Naive Bayes*, *Logistic Regression*, *K-Nearest Neighbor*, dan lain sebagainya [4].

Penelitian ini bertujuan untuk membuat model terbaik untuk prediksi penyakit jantung menggunakan algoritma *data mining Decision Tree* (C4.5). Dalam penelitian ini, digunakan

dataset publik yang diambil dari *Kaggle Datasets* (<https://www.kaggle.com/datasets>). *Dataset* ini terdiri dari 1025 *record* data dan 14 belas atribut. Dari model prediksi yang dihasilkan tersebut, akan dilakukan perbandingan terhadap nilai akurasi yang dihasilkan dari pembagian *data training* dan *testing* pada *dataset*. Berdasarkan hasil perbandingan tersebut akan diperoleh model terbaik yang dapat menghasilkan prediksi yang lebih akurat.

TINJAUAN PUSTAKA

A. *Data Mining*

Definisi *data mining* adalah melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data [9]. Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar [10]. *Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data* [11].

Data mining merupakan proses untuk menemukan pola (*pattern*) dari suatu data. Pola (*pattern*) yang ditemukan harus memiliki arti atau mengandung informasi penting [9]. *Data mining* adalah proses analisis dari *dataset* (berukuran besar) untuk menemukan suatu hubungan tak terduga serta meringkas data tersebut dengan cara baru yang dapat dimengerti dan bermanfaat bagi pemilik data [7]. Menurut Larose (2006), *data mining* merupakan proses analisis dari sekumpulan data untuk menemukan pengetahuan yang tidak terduga dan merangkum data dengan menggunakan cara yang berbeda dari sebelumnya, sehingga dapat dipahami dan berguna bagi pemilik data [12]. Kata *mining* itu sendiri berasal dari kata

mine dari bahasa Inggris yang artinya menambang sumber daya yang tersembunyi [13]. Definisi *data mining* lainnya adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis atau pembelajaran berbasis induksi (*induction based learning*), yaitu proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsep-konsep yang akan dipelajari [14]. *Knowledge Discovery in Databases* (KDD) adalah penerapan metode *scientific* pada *data mining* [14]. Dalam konteks ini, *data mining* merupakan satu langkah dari proses KDD [14].

B. *Classification*

Classification adalah sebuah model dalam *data mining* dimana *classifier* dikonstruksi untuk memprediksi *categorical label* [15]. Kategori tersebut dapat direpresentasikan dengan nilai yang sesuai dengan kebutuhannya, dimana pengaturan dari nilai tersebut tidak memiliki tertentu [11]. Klasifikasi mempunyai ciri data yang bersifat atribut *categorical* dan atribut *categorical* sebagai label yang merupakan keputusan dari atribut data yang lain [15]. Klasifikasi data adalah proses dua langkah, yang terdiri dari langkah belajar (model klasifikasi didekonstruksi) dan langkah klasifikasi (model digunakan untuk memprediksi label kelas untuk data yang diberikan) [11]. Proses data klasifikasi mempunyai dua tahapan, yaitu *learning* (*training data* menggunakan algoritma klasifikasi) dan *classification* (*test data* digunakan untuk mengestimasi ketepatan dari *classification rules*) [15].

C. *Decision Tree* (C4.5)

Algoritma *decision tree* (C4.5) adalah algoritma klasifikasi yang memiliki struktur seperti pohon terdiri atas *node* internal (akar) dan *node* cabang. *Node* internal (akar) merupakan atribut dan sebuah *node* cabang mewakili *class* [16]. Algoritma *decision tree* (C4.5) merupakan algoritma klasifikasi yang mengklasifikasikan sampel data secara *top-down* mulai dari simpul akar dan bergerak sesuai dengan hasil pengujian dari *node* internal sampai *node* cabang dicapai dan *class* ditetapkan [17]. Tahapan-tahapan algoritma *decision tree* (C4.5):

1. Memilih atribut sebagai akar yang didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

- S : himpunan kasus
A : atribut
n : jumlah atribut A
|S_i| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

Nilai entropi dapat dihitung dengan persamaan:

$$\text{Entropy}(S) = \sum_{i=1}^n - p_i * \log_2 p_i$$

- S : himpunan (*dataset*) kasus
n : banyaknya partisi
p_i : probabilitas yang didapat dari jumlah kasus pada partisi ke-i dibagi total kasus

2. Membuat cabang untuk setiap nilai didalam akar tersebut,
3. Membagi atribut dalam cabang,

4. Ulangi proses untuk setiap atribut sampai semua atribut pada cabang memiliki *class* yang sama.

D. Confusion Matrix

Hasil pengujian akan diukur menggunakan *confusion matrix* atau matrik konfusi. *Confusion matrix* memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah [7].

E. RapidMiner



Gambar 1. *RapidMiner*

RapidMiner merupakan *software open source*. *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining*, dan analisis prediksi. *RapidMiner* menggunakan berbagai metode atau teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna, sehingga dapat membuat keputusan yang terbaik [13]. *Software* ini memiliki kurang lebih 500 operator *data mining*, termasuk operator untuk *input*, *output*, *data preprocessing*, dan visualisasi. *RapidMiner* ditulis menggunakan bahasa pemrograman *Java*, sehingga dapat bekerja di semua sistem operasi.

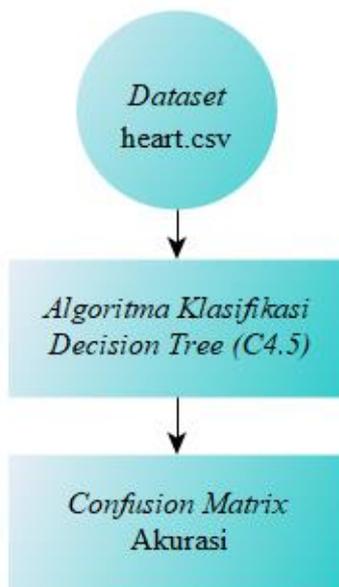
RapidMiner menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah *pipeline* analitis. GUI ini akan menghasilkan *file XML* (*Extensible Markup Language*) yang mendefinisikan proses analitis keinginan pengguna untuk diterapkan ke data. *File* ini kemudian dibaca oleh *RapidMiner* untuk menjalankan analisis secara

otomatis. *RapidMiner* memiliki beberapa sifat sebagai berikut [13]:

1. Ditulis menggunakan bahasa pemrograman *Java*, sehingga dapat dijalankan di berbagai sistem operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator *trees*.
3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa *scripting* memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
5. Konsep *multilayer* untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki GUI, *command line mode* dan *Java* API yang dapat dipanggil dari program lain.

METODE PENELITIAN

Tahapan penelitian yang dilakukan mulai dari *import dataset*, implementasi algoritma klasifikasi C4.5, dan evaluasi hasil menggunakan *confusion matrix* ditunjukkan pada gambar berikut:



Gambar 2. Model Tahapan Penelitian

A. Dataset

Dataset yang digunakan dalam penelitian ini diambil dari <https://www.kaggle.com/datasets>, yaitu dataset heart.csv yang terdiri dari 14 atribut dan 1025 *record* data. *Dataset* ini yang akan diimport ke dalam tool *RapidMiner*. Atribut yang berfungsi sebagai label adalah atribut target. Adapun sebagian dari *record data* pada *dataset* yang digunakan pada penelitian ini ditunjukkan pada tabel berikut:

Tabel 1. *Dataset*

| age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|-----|-----|----|----------|------|-----|---------|---------|-------|---------|-------|----|------|--------|
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1 | 2 | 2 | 3 | 0 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0 | 2 | 1 | 3 | 0 |
| 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |
| 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1 | 1 | 0 | 2 | 1 |
| 58 | 1 | 0 | 114 | 318 | 0 | 2 | 140 | 0 | 4.4 | 0 | 3 | 1 | 0 |
| 55 | 1 | 0 | 160 | 289 | 0 | 0 | 145 | 1 | 0.8 | 1 | 1 | 3 | 0 |
| 46 | 1 | 0 | 120 | 249 | 0 | 0 | 144 | 0 | 0.8 | 2 | 0 | 3 | 0 |
| 54 | 1 | 0 | 122 | 286 | 0 | 0 | 116 | 1 | 3.2 | 1 | 2 | 2 | 0 |
| 71 | 0 | 0 | 112 | 149 | 0 | 1 | 125 | 0 | 1.6 | 1 | 0 | 2 | 1 |
| 43 | 0 | 0 | 132 | 341 | 1 | 0 | 136 | 1 | 3 | 1 | 0 | 3 | 0 |
| 34 | 0 | 1 | 118 | 210 | 0 | 1 | 192 | 0 | 0.7 | 2 | 0 | 2 | 1 |
| 51 | 1 | 0 | 140 | 298 | 0 | 1 | 122 | 1 | 4.2 | 1 | 3 | 3 | 0 |
| 52 | 1 | 0 | 128 | 204 | 1 | 1 | 156 | 1 | 1 | 1 | 0 | 0 | 0 |
| 34 | 0 | 1 | 118 | 210 | 0 | 1 | 192 | 0 | 0.7 | 2 | 0 | 2 | 1 |
| 51 | 0 | 2 | 140 | 308 | 0 | 0 | 142 | 0 | 1.5 | 2 | 1 | 2 | 1 |
| 54 | 1 | 0 | 124 | 266 | 0 | 0 | 109 | 1 | 2.2 | 1 | 1 | 3 | 0 |
| 50 | 0 | 1 | 120 | 244 | 0 | 1 | 162 | 0 | 1.1 | 2 | 0 | 2 | 1 |
| 58 | 1 | 2 | 140 | 211 | 1 | 0 | 165 | 0 | 0 | 2 | 0 | 2 | 1 |

B. Implementasi Algoritma C4.5

Dalam penelitian ini, peneliti mengimplementasikan algoritma klasifikasi C4.5. Algoritma C4.5 merupakan salah satu algoritma dalam penggalian data yang berfungsi sebagai pengklasifikasi suatu kelas. Peneliti menggunakan algoritma C4.5 untuk memprediksi penyakit jantung.

C. Evaluasi Menggunakan *Confusion Matrix*

Hasil pengujian dalam penelitian ini akan diukur menggunakan *confusion matrix* atau matrik konfusi untuk mendapatkan nilai akurasi.

PEMBAHASAN DAN HASIL

A. Dataset

Dalam penelitian ini diambil sampel sebanyak 1025 data dimana dilakukan pembagian *data training* dan *data testing* dengan persentase 90% : 10%, 80% : 20%, 70% : 30%, dan 60% : 40%. Untuk pembagian *data training* dan *data testing* ini dilakukan melalui komponen operator *split data*. Namun sebelumnya, *dataset* yang bertipe *file .csv* akan diimport ke dalam *RapidMiner* melalui operator *Read CSV*. Berikut ini merupakan *dataset* yang sudah diimport ke *RapidMiner* dan sudah dilakukan penentuan tipe data untuk masing-masing atribut:

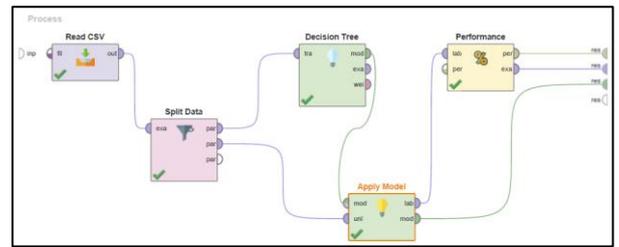
| row No. | target | age | sex | rrb | heartRate | cholest | dia | resting | blood |
|---------|--------|-----|-----|-----|-----------|---------|-----|---------|-------|
| 1 | 0 | 57 | 1 | 0 | 576 | 212 | 0 | 1 | 166 |
| 2 | 0 | 65 | 1 | 0 | 140 | 200 | 1 | 0 | 148 |
| 3 | 0 | 70 | 1 | 0 | 145 | 174 | 0 | 1 | 105 |
| 4 | 0 | 81 | 1 | 0 | 148 | 203 | 0 | 1 | 161 |
| 5 | 0 | 82 | 0 | 0 | 138 | 204 | 1 | 1 | 106 |
| 6 | 1 | 58 | 0 | 0 | 109 | 248 | 0 | 0 | 102 |
| 7 | 0 | 58 | 1 | 0 | 174 | 218 | 0 | 2 | 140 |
| 8 | 0 | 55 | 1 | 0 | 102 | 200 | 0 | 0 | 140 |
| 9 | 0 | 46 | 1 | 0 | 123 | 140 | 0 | 0 | 144 |
| 10 | 0 | 54 | 1 | 0 | 122 | 200 | 0 | 0 | 110 |
| 11 | 1 | 71 | 0 | 0 | 112 | 140 | 0 | 1 | 100 |
| 12 | 0 | 43 | 0 | 0 | 132 | 241 | 1 | 0 | 150 |
| 13 | 1 | 34 | 0 | 0 | 118 | 218 | 0 | 1 | 102 |
| 14 | 0 | 51 | 1 | 0 | 140 | 200 | 0 | 1 | 102 |

Gambar 3. Import Dataset

Dalam proses tersebut juga sudah dilakukan penentuan labelnya, yaitu atribut target.

B. Implementasi Algoritma C4.5

Setelah *dataset* diimport ke dalam *RapidMiner*, selanjutnya dilakukan pembagian *data training* dan *data testing* dari *dataset* tersebut menggunakan operator *split data*. Baru kemudian *data training* digunakan untuk pengimplementasian algoritma C4.5, sedangkan *data testing* untuk pengujian model yang dihasilkan. Gambar proses pada *RapidMiner*nya ditunjukkan berikut ini:



Gambar 4. Implementasi Algoritma C4.5

C. Evaluasi Menggunakan Confusion Matrix

Dengan menggunakan berbagai nilai perbandingan antara *data training* dan *data testing* pada saat implementasi algoritma C4.5, maka diperoleh nilai akurasi dari model yang dihasilkan sebagai berikut:

Tabel 2. Nilai Akurasi

| Data Training : Data Testing | Akurasi |
|------------------------------|---------|
| 90% : 10% | 88,35% |
| 80% : 20% | 82,93% |
| 70% : 30% | 84,09% |
| 60% : 40% | 82,20% |

KESIMPULAN

Berdasarkan pembahasan dan hasil dari penelitian yang telah selesai dilakukan, dapat diambil kesimpulan bahwa nilai akurasi terbaik diperoleh pada model prediksi yang menggunakan perbandingan *data training* dan *data testing* sebesar 90% : 10% yang menghasilkan nilai akurasi sebesar 88,35%. Model prediksi ini diharapkan dapat menjadi alat pendukung dalam diagnosis penyakit jantung, sehingga dapat dilakukan pencegahan serangan jantung yang parah dan dapat meningkatkan persentase angka keselamatan bagi penderita.

DAFTAR PUSTAKA

- [1] A. Riani, Y. Susianto, and N. Rahman, "Implementasi Data Mining untuk Memprediksi Penyakit Jantung Menggunakan Metode Naive Bayes," *JINITA J. Innov. Inf. Technol. Appl.*, vol. 1, no. 1, pp. 25–34, 2019.
- [2] W. Nugraha, "Prediksi Penyakit Jantung

- Cardiovaskular Menggunakan Model Algoritma Klasifikasi,” *J. Sigmata*, vol. 9, no. 2, pp. 78–84, 2021.
- [3] M. S. Oh and M. H. Jeong, “Sex Differences in Cardiovascular Disease Risk Factors among Korean Adults,” *Korean J. Med.*, vol. 95, no. 4, pp. 266–275, 2020.
- [4] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease using Machine Learning Algorithms,” *Hindawi Mob. Inf. Syst.*, vol. Wearable T, pp. 1–21, 2018.
- [5] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis,” *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, pp. 176–183, 2013.
- [6] Q. K. Al-Shayea, “Artificial Neural Networks in Medical Diagnosis,” *IJCSI Int. J. Comput. Sci.*, vol. 8, no. 2, pp. 150–154, 2011.
- [7] F. Gorunescu, *Data Mining Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer, 2011.
- [8] I. D. Mienye, P. Y. Sun, and P. Z. Wang, “An Improved Ensemble Learning Approach for The Prediction of Heart Disease Risk,” *Informatics Med. Unlocked*, pp. 1–5, 2020.
- [9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier Inc., 2011.
- [10] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.
- [11] J. Han, M. Kamber, and J. Pei, *Data Mining, Concepts and Techniques*, Third Edit. Massachusetts: Morgan Kaufmann Publishers, 2012.
- [12] D. T. Larose, *Data Mining Methods and Models*. New Jersey: John Wiley & Sons, Inc, 2006.
- [13] Dennis Aprilla C, Donny Aji Baskoro, L. Ambarwati, and I. W. S. Wicaksana, *Belajar Data Mining dengan Rapid Miner*. Jakarta: Open Content Model, 2013.
- [14] F. Riandari and A. Simangunsong, “Penerapan Algoritma C4.5 untuk Mengukur Tingkat Kepuasan Mahasiswa,” *J. Mantik Penusa*, vol. 3, no. 2, pp. 1–7, 2019.
- [15] A. Wanto *et al.*, *Data Mining: Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis, 2020.
- [16] T. W. Liao and E. Triantaphyllou, *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*. World Scientific, 2007.
- [17] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [18] M. Zed, *Metode Penelitian Kepustakaan*. Jakarta: Yayasan Obor Indonesia, 2008.
- [19] Z. I. Alfianti, “Pengelompokan Wilayah Penyebaran Covid-19 di Kabupaten Karawang Menggunakan Algoritma K-Means,” *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 111–122, 2021.
- [20] N. Mirantika, A. Tsamratul’ain, and F. D. Agnia, “Penerapan Algoritma K-Means Clustering untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat,” *Nuansa Inform. Technol. Inf. J.*, vol. 15, no. 2, pp. 92–98, 2021.
- [21] R. Meri, “Metode K-Means Clustering dalam Persebaran Covid-19 di Sumatra Barat,” *J. Edik Inform.*, vol. 8, no. 1, pp. 1–12, 2021.
- [22] E. Juliana, V. N. Aleyda, and Yuliana, “Penerapan Metode Clustering K-Means untuk Membantu Menentukan Tingkatan Status Daerah Dampak Covid-19,” *J. Media TIK*, vol. 4, no. 3, pp. 112–114, 2021.
- [23] N. Nugroho and F. D. Adhinata, “Penggunaan Metode K-Means dan K-Means++ sebagai Clustering Data Covid-19 di Pulau Jawa,” *Tek. J. Teknol. Inf. dan Komun.*, vol. 11, no. 3, pp. 170–179, 2022.