

# KLASIFIKASI FAKTOR-FAKTOR PENYEBAB PENYAKIT DIABETES MELITUS DI RUMAH SAKIT UNHAS MENGGUNAKAN ALGORITMA C4.5\*

Dewi Rahma Ente<sup>1</sup>, Sri Astuti Thamrin<sup>2</sup>, Hedi Kuswanto<sup>3</sup>, Samsul  
Arifin<sup>4</sup>, and Andreza<sup>5</sup>

<sup>1</sup>Departemen Statistika, FMIPA, Universitas Hasanuddin, Indonesia, dewi01136@gmail.com

<sup>2</sup>Departemen Statistika, FMIPA, Universitas Hasanuddin, Indonesia, tuti@unhas.ac.id

<sup>3</sup>Departemen Statistika, FMIPA, Universitas Hasanuddin, Indonesia, hedikuswanto454@gmail.com

<sup>4</sup>Departemen Statistika, FMIPA, Universitas Hasanuddin, Makassar, Indonesia, zula4717@gmail.com

<sup>5</sup>Pendidikan Dokter, Fakultas Kedokteran, Universitas Hasanuddin, Indonesia,  
andrezafauzialghifary@gmail.com

**Indonesian Journal of Statistics and Its Applications (eISSN 2599-0802)**  
**Vol 4 No 1 (2020), 80 - 88**

Copyright © 2020 Dewi Rahma Ente, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin, and Andreza  
This is an open-access article distributed under the Creative Commons Attribution License, which  
permits unrestricted use, distribution, and reproduction in any medium, provided the original work is  
properly cited.

## **Abstract**

Diabetes mellitus (DM) is one of the chronic and deadly diseases that are widely observed in various countries today. This disease continues and is increasing to a very alarming stage. This study aims to identify and see the relationship between factors that influence DM disease. The method used in this research is C4.5 algorithm which is one of the algorithms used to make predictive classifications. Classification is one of the processes in data mining that aims to find patterns in relatively large data that use the representations in the form of decision trees. This method is applied to data from medical records of patients with DM in 2014-2018 taken from the Hasanuddin University Teaching Hospital. The results obtained indicate that there are four factors that influence the prediction of a patient's DM status namely, Fasting Blood Glucose (GDP), LDL Cholesterol, Triglycerides, and Body Weight.

**Keywords** C4.5 algorithm, classifications, data mining, decision trees, diabetes mellitus

---

Received Jun 2019, Accepted Feb 2020, Published online on Feb 2020

## 1. Pendahuluan

Diabetes Melitus (DM) merupakan penyakit kronis yang ditandai dengan hiperglikemia dan intoleransi glukosa yang terjadi karena kelenjar pankreas tidak dapat memproduksi insulin secara adekuat atau karena tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif (Kurniawaty & Yanita, 2016) WHO memprediksi kenaikan jumlah penyandang DM di Indonesia dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030 International Diabetes Federation (IDF) pada tahun 2009 juga memprediksi kenaikan jumlah penderita DM dari 7,0 juta menjadi 12,0 juta pada tahun 2030. Meskipun terdapat perbedaan angka prevalensi, laporan keduanya menunjukkan adanya peningkatan jumlah penderita DM sebanyak 2–3 kali lipat pada tahun 2030. Hal ini menjadikan Indonesia menduduki rangking ke-4 dunia setelah Amerika Serikat, India dan Cina (PERKENI, 2011) Berdasarkan hasil Riset Kesehatan Dasar tahun 2018 melalui pemeriksaan gula darah menunjukkan prevalensi DM di Indonesia naik dari 6,9% di tahun 2013 menjadi 8,5% di tahun 2018 (RISKESDAS, 2018). Ini merupakan suatu jumlah yang besar untuk dapat ditangani sendiri oleh para ahli DM. Tingginya angka statistik itu, tentunya patut diantisipasi oleh pihak penyedia layanan kesehatan seperti rumah sakit untuk mencegah timbulnya ledakan pasien DM (Rodiyatul *et al* , 2010).

Dalam bidang kedokteran terdapat banyak catatan penderita penyakit DM. Data yang sangat banyak belum dapat digunakan apabila tidak ada informasi atau kesimpulan dari data tersebut. Bahkan data yang banyak justru dapat menjadi sampah dan tidak berguna. Oleh karena itu perlu dilakukan proses ekstraksi untuk mencari informasi dalam data yang belum diketahui sebelumnya. Salah satu metode yang dapat digunakan untuk proses ekstraksi ini melalui pembelajaran mesin (*machine learning*)

Penelitian bidang kesehatan telah banyak yang menggunakan pendekatan pembelajaran mesin. Menurut Selya & Anshutz (2018), pendekatan pembelajaran mesin dapat meningkatkan resiko prediksi pada luaran kesehatan daripada pendekatan konvensional. Sementara itu, Karisma & Otok (2017) mengatakan bahwa metode *Clasificassion and Regression Tress* (CART) merupakan salah satu metode yang dapat digunakan untuk mengetahui faktor yang mempengaruhi penderita DM. Yusa & Sindu (2017) juga menggunakan model *decision tree* dengan algoritma C4.5 untuk klasifikasi terjadinya obesitas. DeGregory *et al* (2018) telah membuktikan bahwa algoritma pembelajaran mesin menyediakan *overview* unik pada tahapan aplikasi analisis data pada obesitas. Rohman (2013) dan Ente *et al* (2019) juga telah menggunakan metode algoritma C4.5 untuk memprediksi penyakit jantung dan penyakit DM.

Berkaitan dengan penyakit DM, status DM penderita penting untuk diketahui sebelum penderita DM mengalami komplikasi serius. Algoritma C4.5 telah populer digunakan untuk memprediksi status penyakit. Oleh karena itu dalam tulisan ini akan digunakan algoritma C4.5 sebagai salah satu implementasi data mining untuk mengklasifikasi penyakit DM.

## 2. Metodologi

### 2.1 Sumber Data

Data yang digunakan pada tulisan ini berasal dari hasil rekam medis pasien DM di Rumah Sakit Pendidikan Universitas Hasanuddin di Kota Makassar. Jumlah pasien DM sebanyak 127 orang. Atribut yang digunakan adalah jenis kelamin (Jk), usia, berat

badan (Bb), tinggi badan (Tb), glukosa darah puasa (GDP), kolesterol HDL, kolesterol LDL, kolesterol total (Kol Tot), trigliserida (Tg) dan status DM (Hasil) Interval usia penderita DM berkisar pada usia 26-82 tahun dengan rata-rata usia pasien adalah 53,03 tahun

Metode analisis yang digunakan pada tulisan ini yaitu teknik data mining dengan algoritma C4.5 dengan perangkat lunak R. Adapun langkah-langkah yang dilakukan adalah melakukan pengumpulan data, melakukan seleksi atribut, penentuan data latih dan data uji dengan validasi silang 10 lipat (*ten-fold cross validation*), menerapkan algoritma C4.5, interpretasi hasil dan validasi hasil

## 2.2 Data Mining

Data mining didefinisikan sebagai proses menemukan pola dalam data. Proses ini harus otomatis atau biasanya secara semi-otomatis. Pola tersebut diidentifikasi, divalidasi, dan digunakan untuk membuat sebuah prediksi (Witten *et al*, 2011).

Data mining dapat diartikan sebagai proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu (Mardi, 2016). Data mining juga merupakan proses seleksi, eksplorasi, dan pemodelan dari sejumlah besar data untuk menemukan pola atau kecenderungan yang biasanya tidak disadari keberadaannya. Data mining dapat dikatakan sebagai proses mengekstrak pengetahuan dari sejumlah besar data yang tersedia (Pramadhani & Setiadi, 2014). Data mining merujuk untuk menemukan pola baru dari kekayaan data dalam database dengan berfokus pada algoritma untuk mengekstrak suatu pengetahuan yang berguna. Data mining memiliki dua tujuan utama yaitu prediksi dan deskripsi. Prediksi melibatkan menggunakan beberapa atribut dalam dataset untuk memprediksi nilai yang tidak diketahui dari atribut lain yang relevan. Sementara itu, deskripsi melibatkan penemuan pola dan tren dalam suatu data (Silwattananusarn & Tuamsuk, 2012). Dengan bantuan teknik data mining dapat diketahui pola suatu penyakit berdasarkan data yang sudah ada seperti nama pasien, usia, jenis kelamin, dan lainnya. Dengan demikian, jika sudah diketahui faktor-faktor yang mempengaruhi suatu diagnose penyakit, maka memudahkan untuk klasifikasi penyakit (Mardi, 2016).

## 2.3 Klasifikasi

Klasifikasi data merupakan suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan obyek di dalam sebuah basis data dan mengklasifikasikannya ke dalam kelas-kelas yang berbeda menurut model klasifikasi yang ditetapkan. Tujuan dari klasifikasi adalah untuk menemukan model dari data latih yang akan membedakan atribut ke dalam kategori atau kelas yang sesuai model.

## 2.4 Pohon Keputusan

Pohon keputusan atau biasa juga disebut dengan *decision tree* merupakan salah satu metode yang dapat digunakan dalam pengklasifikasian dan prediksi yang representasinya menggunakan node dan ruas (Iyer *et al*, 2015). Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon (Mardi, 2016). Struktur sebuah pohon keputusan sederhana yaitu non-terminal node merupakan tes pada satu atau lebih atribut dan terminal node mencerminkan hasil keputusan (Karegowda *et al*, 2012).

## 2.5 Algoritma C4.5

Konsep dasar dari algoritma C4.5 yaitu mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*). Algoritma C4.5 memetakan atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru (Novandya & Oktria, 2017). Kelebihan dari algoritma C4.5 yaitu mudah dimengerti, fleksibel dan menarik karena mampu divisualisasikan dalam gambar. Secara umum langkah-langkah algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Widayu *et al*, 2017):

1. Memilih atribut sebagai akar dengan berdasarkan nilai gain tertinggi pada masing-masing atribut yaitu dengan cara

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

dimana  $S$  adalah himpunan kasus,  $A$  adalah atribut,  $n$  merupakan jumlah partisi atribut  $A$ ,  $|S_i|$  adalah jumlah kasus pada partisi ke- $i$  dan  $|S|$  adalah jumlah kasus dalam  $S$ . Sementara itu, untuk menghitung nilai entropy dapat digunakan rumus:

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i * \log_2 p_i \quad (2)$$

dimana  $S$  adalah himpunan kasus,  $A$  adalah fitur,  $n$  adalah jumlah partisi  $S$ ,  $P_i$  adalah proporsi dari  $S_i$  terhadap  $S$

2. Membuat cabang pada masing-masing nilai
3. Membagi kasus dalam cabang
4. Mengulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

## 2.6 Validasi Silang

Validasi silang merupakan metode statistika yang digunakan untuk melakukan evaluasi dan komparasi terhadap sebuah dataset dengan cara membagikan data tersebut menjadi dua bagian, yaitu data latih dan data uji. Salah satu jenis validasi silang adalah validasi silang lipat 10. Validasi ini dilakukan dengan cara membagi dataset menjadi sepuluh segmen  $d_1 - d_{10}$  yang berukuran sama besar dengan cara melakukan pengacakan data. Kemudian  $d_1$  akan digunakan terlebih dahulu untuk proses latih dan dilakukan validasi menggunakan sisa dari data selain  $d_1$ . Setelah itu  $d_2$  akan digunakan untuk latih, sementara sisa data selain  $d_2$  digunakan untuk validasi, dan seterusnya. Dengan melakukan validasi seperti ini maka akurasi yang akan didapatkan akan lebih tinggi (Refaeilzadeh *et al*, 2009). Interpretasi nilai akurasi dapat diklasifikasikan menjadi lima bagian yang berbeda, yaitu 50%–60% (akurasi sangat lemah), 60%–70% (tingkat akurasi lemah), 70%–80% (tingkat akurasi sedang), 80%–90% (tingkat akurasi tinggi), dan 90%–100% (tingkat akurasi sangat tinggi) (Gorunescu, 2011).

## 3. Hasil dan Pembahasan

Bagian ini menjelaskan mengenai hasil dari klasifikasi algoritma C4.5 serta akurasi aplikasi yang dibangun. Jumlah data yang digunakan pada penelitian ini sebanyak 127 data dengan 9 atribut. 8 atribut yang digunakan sebagai atribut prediktor dan 1 atribut respon dengan menggunakan metode pengukuran validasi silang lipat 10.

Data pada Tabel 1 yang diperoleh di Rumah Sakit tidak disertai keterangan yang secara rinci menjelaskan maksud dari setiap atribut yang diperoleh. Hal ini menjadi acuan bagi peneliti untuk dijadikan sebagai langkah awal yaitu menganalisa maksud dari data dengan pencarian informasi. Adapun informasi yang didapatkan disajikan pada Tabel 2.

Tabel 1 Beberapa sampel data rekam medis DM di Rumah Sakit Pendidikan Universitas Hasanuddin

Jk	Usia	Bb	Tb	GDP	HDL	LDL	Kol Total	Tg	Hasil
1	35	40.5	145	84	43	93	165.6	148	Normal
0	41	38.25	145	95	44	93	163.4	132	Normal
0	52	40	148	230	41	146	231	151	DM
0	41	41.65	149	86	42	97	163.4	122	Normal
1	53	44.1	149	90	42	94	161.8	129	Normal
0	64	52	150	414	35	191	281	165	DM
0	73	50	150	54	67	1254	1077	374	DM
0	71	45	150	212	48	136	258	182	DM
0	68	65	150	229	38	162	207	118	DM
0	47	60	150	212	22	138	167	106	DM
0	53	60	150	352	32	170	205	100	DM
0	67	50	150	137	45	135	201	88	DM
0	52	55	150	150	79.8	112	261	391	DM
0	62	50	150	132	32	99	169	197	DM
0	50	43	150	420	64	169	266	166	DM
0	59	70	150	135	21	147	199	128	DM
0	38	45	150	368	56	118	203	143	DM
0	53	60	150	154	28	111	149	113	DM
0	54	79	150	225	29	174	209	247	DM
0	76	60	150	356	26	102	175	172	DM
0	68	48	150	132	34	154	221	207	DM
1	28	45	150	80	40	96	164	140	Normal
0	54	42.5	150	103	44	96	167.4	137	Normal
1	54	45.9	151	74	41	98	164.2	126	Normal
1	69	61	152	184	52	103	175	81	DM

### 3.1 Seleksi Atribut

Banyaknya atribut pada data yang digunakan pada penelitian ini dapat mengakibatkan terjadinya *overfitting*. Untuk mengatasi masalah ini, maka dilakukan seleksi atribut. Proses seleksi atribut dilakukan dengan menggunakan Uji *Chi-Square* ( $\chi^2$ ). Uji ini berguna untuk melihat atribut prediktor yang berpengaruh terhadap atribut respon. Uji *Chi-Square* diterapkan pada setiap atribut, dan diukur dari nilai *p-value*. Atribut yang paling informatif akan diidentifikasi dengan menyortir masing-masing atribut berdasarkan nilai *p-value*. Penentuan seleksi atribut yaitu dengan membandingkan nilai *p-value* dengan taraf signifikan sebesar 5%.

Adapun hipotesis yang digunakan yaitu:

$H_0$  : Tidak ada pengaruh antara atribut – n dengan penyakit DM

$H_1$  : Ada pengaruh antara atribut – n dengan penyakit DM

Dari Tabel 3 dapat dilihat bahwa terdapat 7 atribut yang berpengaruh yaitu umur, berat badan, tinggi badan, GDP, LDL, kolestrol total dan trigleserida

Tabel 2. Atribut dan deskripsi dari data rekam medis DM di Rumah Sakit Pendidikan Universitas Hasanuddin

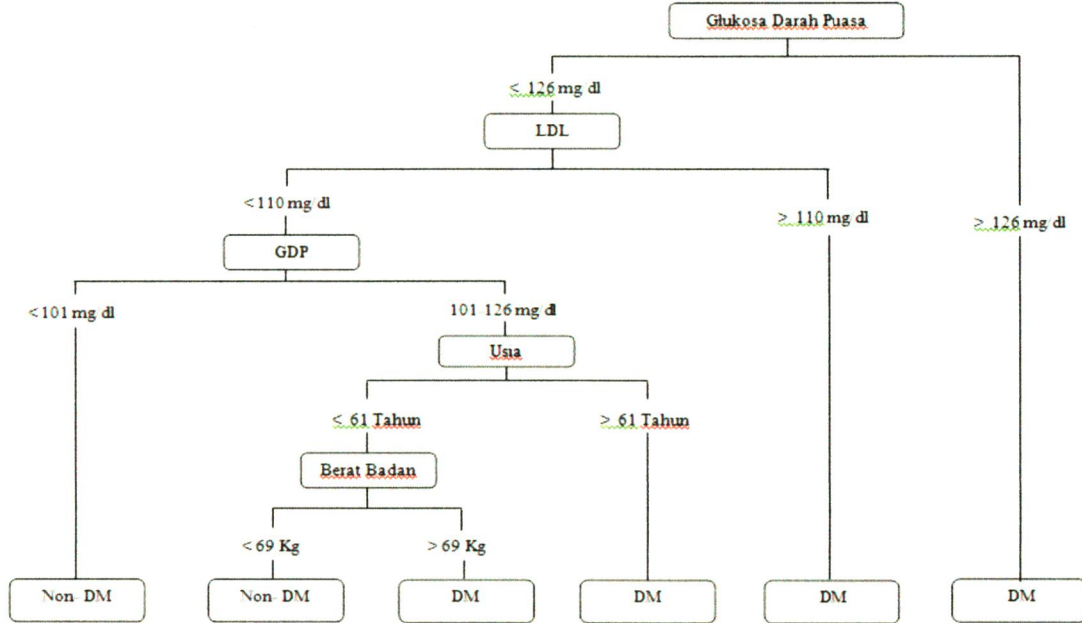
No	Atribut	Skala	Keterangan
1	Jenis Kelamin	Nominal	0 Perempuan 1 Laki-Laki
2	Usia	Ordinal	1 26 – 32 Tahun 2 33 – 39 Tahun 3 40 – 46 Tahun 4 47 – 53 Tahun 5 54 – 60 Tahun 6 61 – 67 Tahun 7 68 – 74 Tahun 8 75 – 81 Tahun
3	Berat Badan	Rasio	-
4	Tinggi Badan	Rasio	-
5	GDP	Ordinal	1 Rendah, bila kadar < 70 mg/dL 2 Normal, bila kadar 70 – 100 mg/dL 3 PraDM, bila kadar 101-126 mg/dL 4 Tinggi, bila kadar > 126 mg/dL
6	HDL	Ordinal	1 Normal, bila kadar > 65 mg/dL 2 Rendah, bila kadar < 65 mg/dL
7	LDL	Ordinal	1 Normal, bila kadar < 110 mg/dL 2 Tinggi, bila kadar > 110mg/dL
8	Kolestrol Total	Ordinal	1 Normal, bila kadar < 200 mg/dL 2 Tinggi, bila kadar > 200 mg/dL
9	Trigliserida	Ordinal	1 Normal, bila kadar < 150 mg/dL 2 Tinggi, bila kadar > 150 mg/dL

Tabel 3: Nilai *p-value* untuk setiap atribut data DM

Atribut	<i>p-value</i>	Hasil
Jenis Kelamin	0,7014	Tidak Berpengaruh
Usia	1 01E-06	Berpengaruh
Berat Badan	6 52E-03	Berpengaruh
Tinggi Badan	0,004644	Berpengaruh
GDP	2,2E-16	Berpengaruh
HDL	0,08883	Tidak Berpengaruh
LDL	6 76E-13	Berpengaruh
Kolestrol Total	8 49E-06	Berpengaruh
Trigliserida	8 49E-06	Berpengaruh

### 3.2 Penerapan Algoritma C4.5

Penerapan algoritma C4.5 pada penelitian ini dilakukan dengan membagi data latih dan data uji menggunakan validasi silang lipat 10. Data latih merupakan data yang digunakan untuk membentuk model dan data uji merupakan data yang digunakan untuk mengevaluasi keakuratan model.



Gambar 1 Model pohon keputusan yang terbentuk dari data DM

Adapun aturan atau *rule* yang terbentuk berdasarkan Gambar 1 yaitu jika  $GDP > 126$  mg/dl maka seseorang positif berpeluang mengidap DM. Jika  $GDP < 126$  mg/dl dan kolesterol  $LDL > 110$  mg/dl maka seseorang positif berpeluang mengidap DM. Jika kolesterol  $LDL < 110$  mg/dl dan  $GDP < 101$  maka seseorang negatif berpeluang DM. Jika kolesterol  $LDL < 110$  mg/dl,  $GDP$  berada antara 101-126 mg/dl dan usia  $< 66$  tahun maka peluang seseorang akan menderita penyakit DM itu kecil. Jika kolesterol  $LDL < 110$  mg/dl,  $GDP$  berada antara 101-126 mg/dl dengan usia  $> 66$  tahun maka seseorang berpeluang besar akan menderita penyakit DM.

Berdasarkan dari *rule* yang didapatkan dengan algoritma C4.5 maka terdapat empat atribut yang dapat digunakan untuk mengidentifikasi faktor-faktor yang substansial mempengaruhi seseorang menderita penyakit DM yaitu kolesterol GDP, LDL, usia, dan berat badan.

Pengukuran akurasi data latih dan data uji dari algoritma C4.5 dengan validasi silang lipat 10 setelah proses seleksi atribut dapat dilihat pada Tabel 4. Dari Tabel 4 dapat diketahui nilai akurasi. Nilai akurasi memiliki rentang antara 50% sampai dengan 100% dengan tingkat akurasi rata-rata prediksi yaitu 98,5%. Ini berarti model yang didapatkan sangat baik dengan tingkat akurasi sangat tinggi.

Tabel 4 Hasil pengukuran akurasi dengan validasi silang lipat 10

Akurasi	Nilai	Persentase
1st fold	0.9942308	99,42%
2nd fold	0.9942308	99,42%
3rd fold	0.9942308	99,42%
4th fold	0.9942308	99,42%
5th fold	0.9846154	98,46%
6th fold	0.9807692	98,07%
7th fold	0.9769231	97,69%
8th fold	0.9769231	97,69%
9th fold	0.9769231	97,69%
10th fold	0.9769231	97,69%

#### 4. Kesimpulan

Algoritma C4.5 merupakan salah satu metode pembelajaran mesin yang menghasilkan pohon keputusan yang optimal dan banyak digunakan dalam klasifikasi. Algoritma C4.5 telah diaplikasikan pada data rekam medis penderita DM di Rumah Sakit Pendidikan Universitas Hasanuddin untuk mengidentifikasi faktor-faktor yang mempengaruhi status penyakit DM seseorang. Faktor-faktor yang mempengaruhi status DM secara substansial adalah glukosa darah puasa (GDP), kolesterol LDL, usia dan berat badan. Dengan mengetahui faktor-faktor yang mempengaruhi status DM penderita maka komplikasi serius akibat DM ini dapat dicegah sedini mungkin.

**Ucapan Terima Kasih.** Terima kasih penulis sampaikan kepada Kemenristekdikti yang telah membiayai penelitian ini melalui hibah Program Kreativitas Mahasiswa (PKM) skema penelitian (PKM-P) tahun 2019.

#### Daftar Pustaka

- DeGregory, K., Kuiper, P., DeSilvio, T., Pleuss, J., Miller, R., Roginski, J., Fisher, C. B., Harness, D., Viswanath, S., Heymsfield, S. B., Dungan, I., & Thomas, D. M. (2018). A review of machine learning in obesity. *Obesity Reviews*, 19(5), 668–685.
- Ente, D., Arifin, S., Andreza, & Thamrin, S. (2019). Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar). *Journal of Physics Conference Series*, 1341(9), 1–8. IOP Publishing.
- Gorunescu, F. (2011). *Data Mining Concepts, models and techniques* (Vol. 12). German: Springer Science & Business Media.
- Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *Journal of Data Mining & Knowledge Management Process*, 5(1), 1–14.
- Karegowda, A. G., Punya, V., Jayaram, M., & Manjunath, A. (2012). Rule based classification for diabetic patients using cascaded k-means and decision tree C4.5. *International Journal of Computer Applications*, 45(12), 45–50.

- Karisma, R. D. L. N., & Otok, B. W. (2017). Model Machine Learning CART Diabetes Melitus. *Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika Dan Nilai-Nilai Islami)*, 1(1), 485–491.
- Kurniawaty, E., & Yanita, B. (2016). Faktor-faktor yang berhubungan dengan kejadian Diabetes Melitus tipe II. *Jurnal Majority*, 5(2), 27–31.
- Mardi, Y. (2016). Data Mining. Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, 2(2), 213–219.
- Novandya, A., & Oktria, I. (2017). Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi. *Jurnal Ilmiah Teknik Informatika*, 6(2), 98–106.
- [PERKENI] Perkumpulan Endokrinologi Indonesia. (2011). *Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia*. Jakarta (ID): Perkumpulan Endokrinologi Indonesia.
- Pramadhani, A. E., & Setiadi, T. (2014). Penerapan Data Mining untuk Klasifikasi Prediksi Penyakit ISPA (Infeksi Saluran Pernapasan Akut) dengan Algoritma Decision Tree (ID3). *Jurnal Sarjana Teknik Informatika*, 2(1), 831–839.
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. LIU & M. T. OZSU (Eds.), *Encyclopedia of Database Systems* (pp. 532–538). [https://doi.org/10.1007/978-0-387-39940-9\\_565](https://doi.org/10.1007/978-0-387-39940-9_565)
- [RISKESDAS] Riset Kesehatan Dasar. (2018). Hasil Utama RISKESDAS 2018. Retrieved June 20, 2019, from [http://www.kemkes.go.id/assets/upload/dir\\_519d41d8cd98f00/files/Hasil-riskesdas-2018\\_1274.pdf](http://www.kemkes.go.id/assets/upload/dir_519d41d8cd98f00/files/Hasil-riskesdas-2018_1274.pdf)
- Rodiyatul, F. S., Tama, B. A., & Mulya, M. (2010). Pengembangan Perangkat Lunak Diagnosa Penyakit Diabetes Mellitus Tipe II Berbasis Teknik Klasifikasi Data. *Prosiding Seminar Nasional Hasil-Hasil Penelitian 2010 UNSRI*
- Rohman, A. (2013). Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung. *Dinamika Sains*, 11(26), 40–49.
- Selya, A. S., & Anshutz, D. (2018). Machine Learning for the Classification of Obesity from Dietary and Physical Activity Patterns. In *Advanced Data Analytics in Health* (pp. 77–97). Springer.
- Silwattananusarn, T., & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: a literature review from 2007 to 2012. *International Journal of Data Mining and Knowledge Management Process*, 2(5), 13–24.
- Widayu, H., Nasution, S. D., Silalahi, N., & Mesran. (2017). Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5. *Media Informatika Budidarma*, 1(2), 32–37.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques (3rd ed)*. New York (US): Elsevier.
- Yusa, M., & Sindu, W. (2017). Evaluasi Model Decision Tree C4.5 Guna Prediksi Posibilitas Resiko Obesitas. *Seminar Nasional Informatika (SNIf)*, 1(1), 147–152.