e-ISSN 2962-1968

Data Mining Untuk Klasifikasi Diagnosis Tingkat Keparahan Penyakit Diabetes Dengan Algoritma Logistik Regresi

Neha Poetri Setiawati[1], Bagus Adi Nugroho[2], Ardhi Tiya Setiawan[3*], Dwi Hartanti[4]

[1,2,3,4]*Informatika*
Univesitas Duta Bangsa Surakarta
[1]*202020853@mhs.udb.ac.id,* [2]*202020121@mhs.udb.ac.id,* [3*]*202020379@mhs.udb..ac.id,* [3]*dwihartanti@udb.ac.id*

Abstrak— **Jurnal ini membahas penerapan data mining dengan menggunakan algoritma logistik regresi untuk mengklasifikasikan tingkat keparahan penyakit diabetes. Penelitian ini menggunakan dataset yang mencakup atribut medis, seperti kadar glukosa darah, tekanan darah, dan indeks massa tubuh. Metode data mining digunakan untuk mengidentifikasi pola dan hubungan dalam dataset, sedangkan algoritma logistik regresi digunakan untuk mengembangkan model klasifikasi. Hasil penelitian menunjukkan bahwa algoritma logistik regresi efektif dalam memprediksi tingkat keparahan diabetes berdasarkan atribut medis yang diuji. Penelitian ini memberikan kontribusi dalam meningkatkan diagnosis dan perawatan pasien diabetes serta mengurangi risiko komplikasi yang terkait dengan penyakit ini.**
Kata kunci— Data Mining, Diabetes, Logistik Regresi, Tingkat Keparahan, Klasifikasi.

Abstract— **This journal discusses the application of data mining using logistic regression algorithm to classify the severity of diabetes. The study utilizes a dataset that includes medical attributes, such as blood glucose levels, blood pressure, and body mass index. Data mining methods were used to identify patterns and relationships in the dataset, while the logistic regression algorithm was used to develop a classification model. The results showed that the logistic regression algorithm was effective in predicting the severity of diabetes based on the tested medical attributes. This research contributes to improving the diagnosis and treatment of diabetic patients and reducing the risk of complications associated with the disease.**
Keywords— Data Mining, Diabetes, Logistic Regression, Severity, Classification.

I. PENDAHULUAN

Diabetes mellitus, suatu kondisi kronis yang ditandai dengan kadar glukosa darah yang tinggi, merupakan masalah kesehatan yang signifikan yang mempengaruhi jutaan orang di seluruh dunia. Penanganan dan pengobatan yang tepat sangat penting karena dapat menyebabkan komplikasi serius jika tidak ditangani atau tidak terkontrol dengan baik. Tingkat keparahan diabetes dapat bervariasi dari ringan hingga berat, dan penilaian tingkat keparahan yang akurat sangat penting untuk penanganan dan perawatan yang tepat.

Dalam beberapa tahun terakhir, data mining telah mendapatkan popularitas sebagai bidang penelitian dalam analisis data medis. Data mining melibatkan proses penggalian pengetahuan atau informasi yang berguna dari kumpulan data yang besar dan kompleks. Dalam konteks diabetes, data mining dapat digunakan untuk mengidentifikasi pola dan hubungan dalam data pasien, sehingga memungkinkan penerapan algoritma klasifikasi yang efektif untuk mengklasifikasikan tingkat keparahan diabetes.

Salah satu algoritma klasifikasi yang umum digunakan dalam data mining adalah regresi logistik. Algoritma ini memodelkan hubungan antara variabel dependen (tingkat keparahan diabetes) dan variabel independen (faktor-faktor yang berhubungan dengan diabetes) menggunakan fungsi logistik. Regresi logistik telah terbukti berhasil dalam berbagai aplikasi klasifikasi, termasuk dalam bidang kesehatan.

Tujuan dari penelitian ini adalah untuk mengintegrasikan data mining dan algoritma regresi logistik untuk mengklasifikasikan tingkat keparahan diabetes. Dengan menggunakan teknik data mining, penelitian ini akan mengeksplorasi dataset yang mencakup berbagai atribut medis seperti kadar glukosa darah, tekanan darah, indeks massa tubuh (BMI), riwayat diabetes dalam keluarga, dan lain-lain. Melalui pemanfaatan algoritma regresi logistik, penelitian ini akan mengembangkan model klasifikasi yang mampu memprediksi tingkat keparahan diabetes berdasarkan atribut-atribut tersebut. Diharapkan hasil dari penelitian ini dapat memberikan kontribusi dalam bidang kesehatan dengan menyediakan metode yang lebih akurat dan efektif dalam mengklasifikasikan tingkat keparahan penyakit diabetes. Pengetahuan yang dihasilkan dari penelitian ini dapat digunakan oleh para profesional

e-ISSN 2962-1968

medis untuk membantu dalam diagnosis dini dan manajemen yang tepat, sehingga meningkatkan perawatan pasien dengan diabetes dan mengurangi risiko komplikasi potensial.

II. METODOLOGI PENELITIAN

Penelitian ini dilakukan dengan cara mengumpulkan data dari pasien yang bertujuan untuk mengklasifikasikan diagnosis tingkat keparahan penyakit diabetes dengan algoritma logistik regresi seperti gambar 1 yang terdiri dari 5 tahap yaitu (1) Pengumpulan Data (2) Persiapan Data (3) Implementasi Algoritma (4) Hasil Penelitian dan Pengujian (5) Kesimpulan.

Gambar 1. Metode penelitian

Tahap pengumplan data merupakan tahap pengumpulna data, data dari pasien pada data yang diperoleh dari data publik *diabetes prediction dataset* dengan jumlah data sebanyak dua ratus data pasien yang kami gunakan dan terdapat 8 atribut yaitu gender, age, hypertension, heart disease, smoking history, bmi, HbA1 level, blood glucose level, diabetes. Tahap Persiapan Data Langkah ini melibatkan pembersihan dan pengolahan awal data untuk memastikan kualitas dan kecocokan yang tepat. Hal ini dapat mencakup penghilangan data yang hilang atau tidak lengkap, normalisasi atau standarisasi atribut, dan penanganan outliers jika diperlukan. Tahap implementasi algoritma yaitu tahap dimana algoritma diimplementasikan pada data yang diperoleh sebelumnya.

Model persamaan aljabar layaknya OLS yang biasa kita gunakan adalah berikut: Y = B0 + B1X + e. Dimana e adalah error varians atau residual. Dengan model regresi ini, tidak menggunakan interpretasi yang sama seperti halnya persamaan regresi OLS. Model Persamaan yang terbentuk berbeda dengan persamaan OLS

$$1n + \frac{P'}{-p'} + B_0 + B_1 x$$

- Ln : Logaritma Natural
- B0 + B1X: Persamaan yang biasa dikenal Dalam OLS

Sedangkan P Aksen adalah probabilitas logistik yang didapat rumus sebagai berikut

$$P' = \frac{\exp\,(B_0 + B_1 x)}{1 + \exp\,(B_0 + B_1 x)} = \frac{e^{B_{0+B1x}}}{1 + e^{B_{0+B1x}}}$$

- Dimana exp atau ditulis “e” adalah fungsi exponen.

(Perlu diingat bahwa exponen yaitu kebalikan dari algoritma natural. Sedangkan logaritma natural adalah bentuk logaritma namun dengan nilai konstan 2,71828182845904 atau bisa dibulatkan menjadi 2,27).

III.HASIL PEMBAHASAN

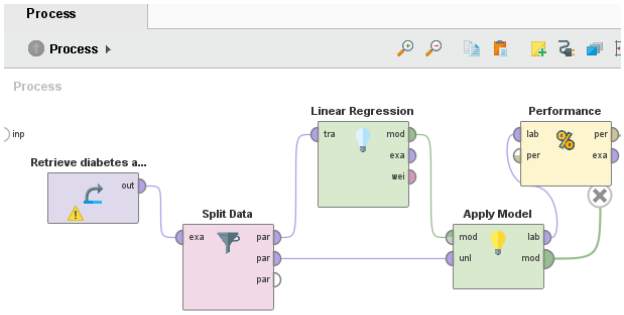
Pada penilitian ini menggunakan dataset yang diambil dari *kaggle* data publik *diabetes prediction dataset*. variabel diantaranya gender, age, hypertension, heart disease, smoking history, bmi, HbA1 level, blood glucose level, diabetes. Dari delapan variabel tersebut menjadi penentu klasifikasi tingkat keparahan penyakit diabetes yang di derita. Dalam kasus kami diagnose diabetes termasuk dalam Regresi Logistik Biner. Karena

memiliki dua kemungkinan . Yaitu 1 dan 0. Jika ya bernilai 1 dan tidak bernilai 0.

A. Pengujian Model Logistik Regresi dengan RapidMIner

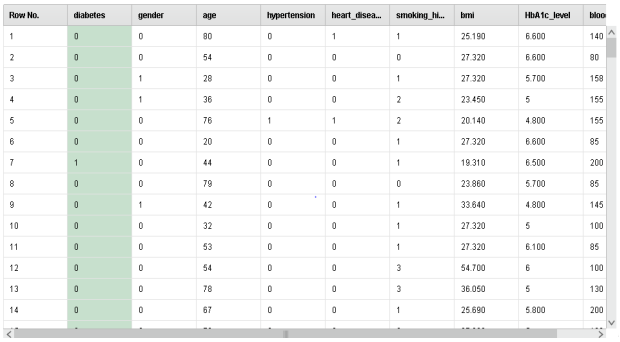
Regresi logistik ialah metode analisis data yang memanfaatkan konsep matematika untuk mengidentifikasi keterkaitan antara dua faktor data. Dengan menggunakan keterkaitan ini, metode ini dapat memprediksi nilai salah satu faktor berdasarkan faktor yang lain. Prediksi yang dihasilkan umumnya bersifat terbatas, seperti hanya dapat memberikan jawaban ya atau tidak.

Pada penilitian ini menggunakan dataset yang diambil dari *kaggle*. Pengujian dilakukan dengan model data mining algoritma logistik regresi menggunakan tools excel dan *rapidminer* versi 10.1 (gambar 2). Variabel yang dimaksukan terdiri dari 8 variabel diantaranya gender, age, hypertension, heart disease, smoking history, bmi, HbA1 level, blood glucose level, diabetes. Dari delapan variabel tersebut menjadi penentu klasifikasi tingkat keparahan penyakit diabetes yang di derita.



Gambar 2. Rancangan Model Pengujian Logistik Regresi

Gambar 2, terdapat proses pengambilan data yang menggunakan pembatasan jumlah data sebanyak 200 data. Langkah berikutnya adalah membagi data, di mana data akan dibagi menjadi dua bagian. Sebanyak 70% data akan digunakan untuk pelatihan menggunakan algoritma Regresi Logistik, sementara 30% data akan digunakan sebagai data untuk pengambilan keputusan.



Row No.	diabetes	gender	age	hypertension	heart_disea...	smoking_hi...	bmi	HbA1c_level	bloo
1	0	0	80	0	1	1	25.190	6.600	140
2	0	0	54	0	0	0	27.320	6.600	80
3	0	1	28	0	0	1	27.320	5.700	158
4	0	1	36	0	0	2	23.450	5	155
5	0	0	76	1	1	2	20.140	4.800	155
6	0	0	20	0	0	1	27.320	6.600	85
7	1	0	44	0	0	1	19.310	6.500	200
8	0	0	79	0	0	0	23.860	5.700	85
9	0	1	42	0	0	1	33.640	4.800	145
10	0	0	32	0	0	1	27.320	5	100
11	0	0	53	0	0	1	27.320	6.100	85
12	0	0	54	0	0	3	54.700	6	100
13	0	0	78	0	0	3	36.050	5	130
14	0	0	67	0	0	1	25.690	5.800	200

Gambar 3. Example Dataset

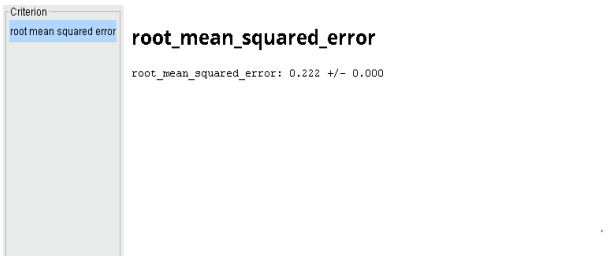
Pada gambar 3 menjelaskan data pasien yang berjumlah 200 . untuk tabel berwarana hijau sebagai labelnya. Dan gender, age, hypertension, heart disease, smoking history, bmi, HbA1 level, blood glucose level, diabetes sebagai atributnya. Selain keterkaitan sebagai penentu tingkat keparahan penyakit diabtes .

ExampleSet (Apply Model) | LinearRegression (Linear Regression) | PerformanceVector (Performance)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
gender	0.031	0.040	0.052	1.000	0.768	0.444	
age	0.002	0.001	0.135	0.955	1.883	0.062	*
heart_disease	-0.089	0.107	-0.059	0.959	-0.835	0.405	
smoking_history	0.027	0.014	0.133	0.991	1.932	0.056	*
bmi	-0.002	0.003	-0.040	0.977	-0.580	0.563	
HbA1c_level	0.113	0.017	0.457	0.988	6.738	0.000	***
blood_glucose_l...	0.003	0.000	0.359	0.990	5.243	0.000	***
(Intercept)	-0.986	0.132	?	?	-7.441	0.000	***

Gambar4. Hasil operasi logistik regresi

Pada gambar 4 menjelaskan bahwa nilai koofisien yang sangat berpengaruh untuk penyakit diabetes ini yaitu blood_glucose_level danHbA1c level karena semakin banyak nilai bintang pada code yang tertera semakin tinggi kemungkinan terkena penyakit diabetes.



Gambar 5. Menunjukkan Hasil Nilai error

Pada gambar 5 menjelaskan Fungsi **root mean squared error** menghitung akar kuadrat rata-rata dari perbedaan antara nilai sebenarnya dan nilai yang diprediksi oleh model. Secara matematis.

Gambar 6. Description

Pada gambar 6 menjelaskan tentang formula rumus yang dapat menghasilkan prediksi kemungkinan penyakit diabetes.

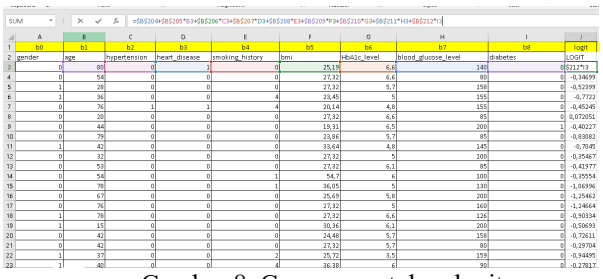
B. Pengujian Model Logistik Regresi dengan Excel

Perhitungan model logistik regresi dengan menggunakan excel sedikit rumit. Pada perhitungan excel kita harus mengansumsikan semua nilai variabel dengan 0,001. Seperti pada gambar 6.

b0	0,001
b1	0,001
b2	0,001
b3	0,001
b4	0,001
b5	0,001
b6	0,001
b7	0,001
b8	0,001

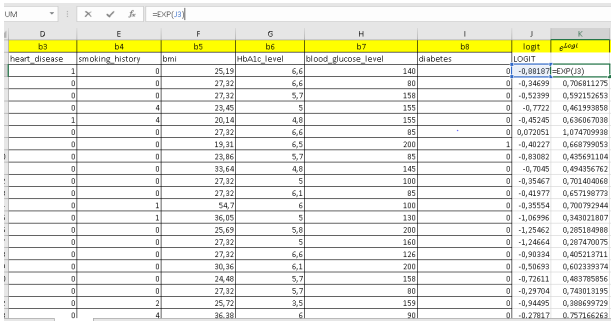
Gambar 7. Asumsi nilai variabel

Untuk selanjutnya evaluasi nilai logit. Nilai logit berfungsi untuk menggambarkan hubungan variabel dependen dan variabel independen dalam bentuk logaritma odds. Untuk menentukan logit dengan rumus seperti pad gambar 7.



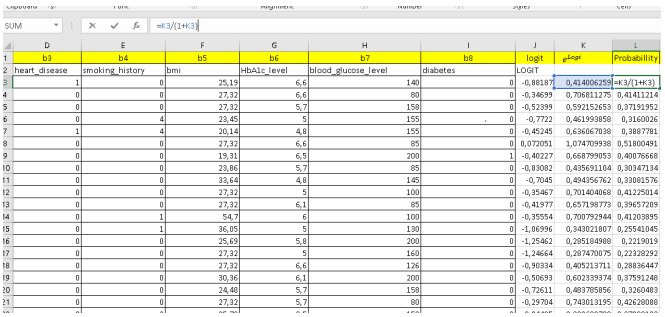
Gambar 8. Cara menentukan logit

Menentukan Eksponensial Logit pada regresi logistik, menggunakan fungsi invers logit atau sigmoid function. Fungsi tersebut mengubah nilai logit menjadi probabilitas antara 0 dan 1. Cara menentukan Eksponensial Logit dengan rumus seperti pada gambar 8.



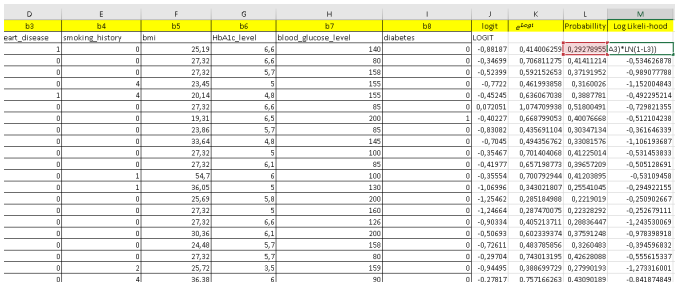
Gambar 9. Cara menentukan e-logit

P(X) merupakan nilai probabilitas disebabkan karena terjadinya peristiwa X. Probabilitas kejadian X dapat dijelaskan dengan $P(x) = \frac{e^x}{1+e^x}$. bisa dilihat pada gambar 9.



Gambar 10. Menetukan nilai probabilitas

Menentukan Evaluasi Jumlah Nilai Log-Kemungkinan dengan rumus pada gambar 10.



Gambar 11. Menentukan evaluasi jumlah log-kemungkinan

Evaluasi jumlah nilai log-kemungkinan dalam algoritma logistik regresi berperan dalam

mengevaluasi performa dan kesesuaian model logistik regresi terhadap data yang digunakan untuk pelatihan. Nilai log-kemungkinan merupakan hasil dari logaritma fungsi kemungkinan logistik, yang menggambarkan probabilitas kelas target (contohnya, kelas biner 0 atau 1) untuk setiap sampel dalam dataset tersebut.

b0	-0,18550575	solver Decision Variables
b1	-0,01309839	
b2	0,753993708	
b3	0,13537012	
b4	-0,01256578	
b5	0,006600206	
b6	0,119145128	
b7	-0,0052606	
b8	0,509790265	

Gambar 12. Solver decision variables

Pada gamabr 11 menjelaskan langgkah selanjutnya adalah Analisis Solver. dalam logistik regresi merupakan metode yang digunakan untuk mencari solusi numerik yang optimal dengan tujuan memaksimalkan fungsi likelihood atau log-likelihood pada model logistik regresi. Fungsi ini memiliki peran penting dalam menentukan estimasi parameter model yang paling cocok dengan data pelatihan.

spesific case	
0	gender
43	age
0	hypertension
0	heart_disease
0	smoking_history
26,71	bmi
6,5	HbA1c_level
300	blood_glucose_level
188,3214757	X
0,420465948	e-Logit
29,60%	probabilitas

Gambar 13. Spesific case

Pada gambar 12 kami mengambil contoh data pada nomor 200 untuk menjadi spesific case. Spesific case dijelaskan sebagai Analisis Risiko Logistik regresi digunakan untuk menganalisa risiko atau kemungkinan kejadian tertentu. Misalnya, dalam studi kesehatan, logistik regresi dapat digunakan untuk memprediksi kemungkinan seseorang mengalami penyakit yang didasarkan pada faktor risiko tertentu.

III. KESIMPULAN

Studi ini menunjukkan bahwa pemanfaatan data mining, khususnya dengan menggunakan algoritma regresi logistik, efektif dalam analisis dan klasifikasi tingkat keparahan penyakit diabetes. Temuan penelitian menunjukkan bahwa algoritma regresi logistik berhasil memprediksi tingkat keparahan penyakit diabetes dengan akurasi tinggi.

Metode data mining yang digunakan dalam penelitian ini melibatkan pemrosesan dan analisis data yang terkait dengan gejala diabetes dan variabel yang relevan dalam diagnosis penyakit tersebut. Pengumpulan data dilakukan dari sampel pasien yang terdiagnosis diabetes dengan tingkat keparahan yang berbeda. Data tersebut kemudian disaring, diproses, dan dianalisis menggunakan algoritma regresi logistik.

Hasil analisis menunjukkan bahwa algoritma regresi logistik mampu mengidentifikasi pola-pola dalam data dan menghasilkan model prediksi yang mampu mengklasifikasikan tingkat keparahan penyakit diabetes dengan akurasi yang baik. Temuan ini mendukung penggunaan data mining sebagai alat yang berharga dalam membantu diagnosis penyakit diabetes dengan lebih akurat dan efisien.

Dalam konteks penggunaan algoritma regresi logistik, penelitian ini menunjukkan bahwa variabel-variabel yang relevan dalam diagnosis tingkat keparahan penyakit diabetes dapat diidentifikasi dan digunakan sebagai faktor prediktif dalam model. Algoritma ini dapat memperhitungkan hubungan antara variabel-variabel tersebut dan menghasilkan estimasi probabilitas terjadinya tingkat keparahan penyakit diabetes pada pasien tertentu.

Namun, perlu diingat bahwa hasil penelitian ini bergantung pada kualitas dan representativitas data yang digunakan. Selain itu, algoritma regresi logistik memiliki kelemahan dan batasan, seperti asumsi linearitas dan independensi yang perlu diperhatikan.

Secara keseluruhan, penelitian ini menyimpulkan bahwa pemanfaatan data mining dengan algoritma

regresi logistik merupakan pendekatan efektif dalam klasifikasi diagnosis tingkat keparahan penyakit diabetes. Namun, penelitian lanjutan dengan sampel yang lebih besar dan variasi variabel diperlukan untuk memvalidasi dan meningkatkan keandalan model prediksi yang dihasilkan.

REFERENSI

[1] Ginting, J. A. (2019). Data mining untuk analisa pengajuan kredit dengan menggunakan metode logistik regresi. *Jurnal Algoritma, Logika dan Komputasi*, *2*(2). https://doi.org/10.30813/j-alu.v2i2.1845

[2] Sa'diah, C., Widiharih, T., & Hakim, A. R. (2021). Klasifikasi PEMBERIAN kredit Sepeda motor MENGGUNAKAN METODE REGRESI LOGISTIK BINER Dan CHI-squared automatic interaction detection (CHAID) DENGAN GUI R (Studi Kasus: Kredit Sepeda motor Di PT X). *Jurnal Gaussian*, *10*(2), 159-169. https://doi.org/10.14710/j.gauss.v10i2.29923S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[3] Ghozi, S., Ramli, R., & Setyani, A. (2018). Analisis keputusan nasabah dalam memilih jenis bank: Penerapan model regresi logistik biner (Studi kasus pada bank Bri cabang balikpapan). *MEDIA STATISTIKA*, *11*(1), 17-26. https://doi.org/10.14710/medstat.11.1.17-26R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[4] Sari, M., & Purhadi, P. (2021). Pemodelan indeks pembangunan manusia provinsi jawa barat, jawa Timur Dan jawa tengah tahun 2019 dengan menggunakan metode regresi logistik ordinal. *Jurnal Gaussian*, *10*(1), 149-158. https://doi.org/10.14710/j.gauss.v10i1.30022M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[5] *Innassuraiya, S., Widiharih, T., & Utami, I. T. (2022). Analisis KLASIFIKASI MENGGUNAKAN METODE REGRESI LOGISTIK BINER Dan bootstrap aggregating classification and regression trees (Bagging cart) (Studi Kasus: Nasabah Koperasi Simpan Pinjam Dan Pembiayaan Syariah (KSPPS)). Jurnal Gaussian, 11(2), 183-194. https://doi.org/10.14710/j.gauss.v11i2.35458* "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.