

KLASIFIKASI BREAST CANCER MENGGUNAKAN METODE LOGISTIC REGRESSION

Andita Dani Achmad

Program Studi Teknik Elektro, Universitas Fajar
e-mail: anditadaniachmad@gmail.com

Abstrak

Kanker payudara merupakan jenis kanker dengan kejadian tertinggi bagi wanita Indonesia. Kanker payudara yang dideteksi karena adanya gejala cenderung lebih sulit disembuhkan dibandingkan didiagnosis setelah dilakukan pemeriksaan lebih dini. Salah satu pengujian yang dilakukan untuk mencari sel-sel kanker payudara adalah mamografi. Penelitian ini bertujuan untuk mengidentifikasi apakah terdapat indikasi kanker payudara dari citra hasil mamografi. Dengan menggunakan metode logistic regression diperoleh akurasi klasifikasi ata latih sebesar 76.04% dan akurasi data uji sebesar 83.33%.

Kata kunci: Klasifikasi, Kanker, Kanker Payudara, Mamografi, Logistic Regression

Abstract

Breast cancer is the type of cancer with the highest incidence in Indonesian women. Breast cancer that is detected because of symptoms tends to be more difficult to cure than diagnosed after an early examination. One of the tests performed to look for breast cancer cells is mammography. This study aims to identify whether there are indications of breast cancer from mammographic images. By using the logistic regression method, the classification accuracy of training data is 76.04% and the accuracy of test data is 83.33%.

Keywords: Classification, Cancer, Breast Cancer, Mammography, Logistic Regression

1. PENDAHULUAN

Kanker adalah pertumbuhan yang tidak normal dari sel-sel jaringan tubuh yang berubah menjadi ganas. Sel-sel tersebut dapat tumbuh lebih lanjut serta menyebar ke bagian tubuh lainnya serta menyebabkan kematian. Sel tubuh yang mengalami mutase (perubahan_ dan mulai tumbuh serta membelah lebih cepat dan tidak terkendali seperti sel normal. Sel kanker tidak mati setelah usianya cukup melainkan tumbuh terus dan bersifat *invasing* (menyerang) sehingga sel normal yang ter umbuh dapat terdesak atau malah mati [1].

Kanker dengan angka kejadian tertinggi pada penduduk wanita adalah kanker payudara. Kanker. Penyakit ini menyebabkan perempuan di Indonesia meninggal duni dengan angka kematian 21,5 per 100.000 penduduk [2].

Penelitian terdahulu telah dilakukan dengan teknik pembelajaran mesin untuk memprediksi penyakit payudara menggunakan algoritma Gain Ratio [3], support vector machine [4], dan koparasi algoritma naïve bayes, decision tree dan support vector machine [5]. Penelitian ini akan mengklasifikasi *breast cancer* dengan menggunakan metode *logistic regression*.

2. LANDASAN TEORI

2.1 Kanker Payudara

Kanker payudara adalah pertumbuhan sel yang abnormal pada jaringan payudara seseorang. Sebagian besar kanker payudara bermula pada sel-sel yang melapisi ductus (kanker ductal), beberapa bermula di lobules (kanker lobular), serta sebagian kecil bermula di jaringan lain [6].

Istilah "*benign*" atau "jinak" memiliki arti ringan atau tidak progresif. Dikatakan tumor jinak jika gangguan *proliferasi* sel-sel tidak menyerang jaringan yang berada di dekatnya atau tidak menyebar ke bagian lain dari tubuh. Tumor jinak biasanya dikelilingi oleh selubah fibrosa (kapsul) yang menghambat kemampuan meraka untuk berperilaku ganas [7].

Istilah "*malign*" atau "ganas" yang berarti buruk dan berpotensi mematikan. Tumor ganas adalah kanker, tetapi kaner belum tentu berbentuk tumor. Sel-sel kanker dapat menyerang dan merusak jaringan dan organ di dekat tumor. Sel-sel kanker dapat melepaskan diri dari tumor ganas dan memasuki aliran serta menyebar dari tumor asli untuk membentuk tumor baru dari bagian tubuh lainnya. Kebanyakan tumor ganas tidak memiliki kapsul seperti tumor jinak [7].

2.2 Klasifikasi

Klasifikasi adalah proses mengelompokkan data dengan karakteristik yang sama kedalam kelas-kelas yang telah ditentukan terlebih dahulu. Label kelas dalam klasifikasi digunakan untuk menamai kelompok yang memiliki pola yang serupa [8].

2.3 Logistic Regression

Logistic regression merupakan salah satu jenis regresi yang menghubungkan antara satu atau beberapa variabel independen (variabel bebas) dengan variabel dependen yang berupa kategori, biasanya 0 dan 1. *Logistic regression* banyak digunakan pada ilmu biologi dan social. *Logistic regression* digunakan manakal variabel yang ada berjenis kategori. Contoh penerapan *logistic regression* :

1. *Spam detection* : memprediksi apakah email termasuk kategori *spam* (1) dan bukan (0).
2. *Medical* : menentukan apakah jaringan berjenis tumor (*malignant*) atau bukan (*benign*).
3. *Credit card fraud* : memprediksi apakah transaksi *credit card* termasuk kategori *fraud* (1) atau bukan (0).
4. *Banking* : memprediksi apakah calon customer layak mendapatkan *credit/loan* (1) atau tidak (0).

Contoh permasalahan di atas berhubungan dengan probabilitas di mana *logistic regression* banyak digunakan dalam penentuan regresi binomial.

Kata binomial artinya hanya ada dua kemungkinan yang menjadi *outcome*, tidak bisa lebih dari dua. Misalkan 1 atau 0, ya atau tidak, hitam atau putih, dan seterusnya. Sehingga *logistic regression* dapat digunakan untuk berbagai permasalahan yang berkaitan dengan klasifikasi (*classification*).

Tabel 1. Perbedaan *Classification* dan *Regression*

<i>Classification</i>	<i>Regression</i>
<i>Input variable</i> bersifat diskrit (<i>discrete</i> atau <i>categorical</i>)	<i>Input variable</i> bersifat kontinu (<i>continuous</i>)
Tujuan atau <i>outcome</i> berupa <i>class</i>	Tujuan atau <i>outcome</i> berupa bilangan riil
Digunakan untuk memprediksi <i>class</i> di masa depan	Digunakan untuk memprediksi nilai di masa depan
Metode kalkulasi dengan mengukur akurasi	Metode kalkulasi dengan menghitung nilai <i>root mean square</i>
Algoritma : <i>classification tree</i> , <i>logistic regression</i> , dan lain-lain.	Algoritma : <i>regression tree</i> , <i>linear regression</i> , dan lain-lain.

Sigmoid Function

Pada *logistic regression*, komputer belajar menemukan model yang disebut sigmoid function atau logistic function. Rumus *sigmoid function* yang umum digunakan, ialah:

$$f(x) = \frac{1}{1+e^{-x}} \quad [9]$$

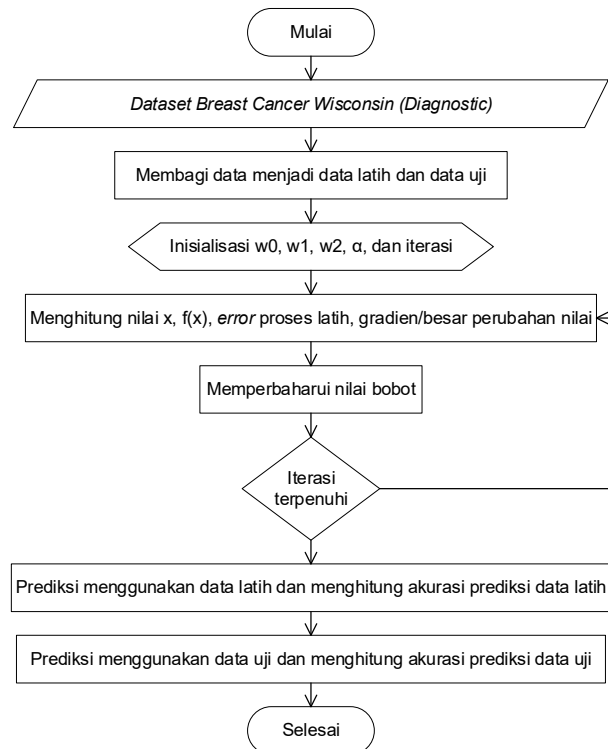
3. METODE PENELITIAN

3.1. Jenis Data dan Sumber Data

Penelitian ini menggunakan data sekunder yang bersumber dari UCI *Machine Learning Repository*. Data yang digunakan adalah *Breast Cancer Wisconsin (Diagnostic)* dari total data set 569 dan 10 atribut, digunakan sebanyak 120 data set dan 2 atribut yaitu radius dan tekstur.

3.2. Alur Klasifikasi

Tahapan klasifikasi *breast cancer* digambarkan pada diagram alir pada gambar 1.



Gambar 1. Tahapan Klasifikasi *Breast Cancer* Menggunakan Metode *Logistic Regression*

4. PEMBAHASAN

Pengklasifikasian *breast cancer* menggunakan metode *logistic regression* menggunakan 120 data pasien dengan 2 atribut, yaitu atribut radius dan tekstur. Data tersebut dibagi menjadi 80% data untuk data latih dan 20% data untuk data uji seperti pada pada tabel 2, data 1–96 digunakan untuk data latih sedangkan data 97–120 digunakan sebagai data uji.

Tabel 2. Dataset *Breast Cancer Wisconsin (Diagnostic)*

Radius (X)	Tekstur (X)	Status (Y)	Radius (X)	Tekstur (X)	Status (Y)	Radius (X)	Tekstur (X)	Status (Y)
17.99	10.38	0	9.787	19.94	1	12.98	19.35	1
20.57	17.77	0	11.6	12.84	1	11.22	19.86	1
19.69	21.25	0	6.981	13.43	1	11.25	14.78	1
11.42	20.38	0	12.18	20.52	1	12.3	19.02	1
20.29	14.34	0	9.876	19.4	1	17.06	21	0
12.45	15.7	0	10.49	19.29	1	12.99	14.23	1
18.25	19.98	0	11.43	17.31	1	18.77	21.43	0
13.71	20.83	0	12.9	15.92	1	11.51	23.93	1
13	21.82	0	10.75	14.97	1	14.05	27.15	1
12.46	26.04	0	11.9	14.65	1	11.2	29.37	1
17.14	16.4	0	14.95	18.77	1	15.22	30.62	0
14.58	21.53	0	14.44	15.18	1	20.92	25.09	0
18.61	20.25	0	13.74	17.91	1	21.56	22.39	0
15.3	25.27	0	13	20.78	1	20.13	28.25	0
17.57	15.05	0	8.219	20.7	1	16.6	28.08	0
18.63	25.11	0	9.731	15.34	1	20.6	29.33	0
11.84	18.7	0	11.15	13.08	1	7.76	24.54	1
17.02	23.98	0	13.15	15.34	1	20.64	17.35	0
19.27	26.47	0	12.25	17.94	1	13.69	16.07	1
16.13	17.88	0	16.78	18.8	0	16.17	16.07	1
17.2	24.52	0	17.47	24.68	0	10.57	20.22	1
13.8	15.79	0	13.43	19.63	0	13.46	28.21	1
18.05	16.15	0	15.46	11.89	0	13.66	15.15	1
20.18	23.97	0	13.77	22.29	0	11.08	18.83	0
25.22	24.91	0	18.08	21.84	0	14.68	20.13	0
19.1	26.29	0	19.18	22.49	0	16.13	20.68	0
10.82	24.21	1	18.22	18.87	0	19.81	22.15	0
10.86	21.48	1	9.042	18.9	1	13.54	14.36	1
11.13	22.44	1	12.43	17	1	13.08	15.71	1
12.77	29.43	1	10.25	16.18	1	9.504	12.44	1
9.33	21.94	1	20.16	19.66	0	16.5	18.29	1
12.88	28.92	1	12.86	13.32	1	13.4	16.95	1
10.29	27.61	1	20.34	21.51	0	20.44	21.78	0
10.16	19.59	1	12.2	15.21	1	20.2	26.83	0
9.423	27.88	1	12.67	17.3	1	12.21	18.02	1
14.59	22.68	1	14.11	12.88	1	21.71	17.25	0
12.36	21.8	1	12.03	17.93	1	11.95	14.96	1

14.64	15.24	1	16.27	20.71	0	11.66	17.07	1
14.62	24.02	1	16.26	21.88	0	15.75	19.22	0
12.18	17.84	1	16.03	15.51	0	25.73	17.46	0

Tabel 3. Hasil Pengklasifikasian *Breast Cancer* Menggunakan Metode *Logistic Regression*

Radius (X)	Tekstur (X)	Status (Y)	Radius (X)	Tekstur (X)	Status (Y)	Radius (X)	Tekstur (X)	Status (Y)
17.99	10.38	0	9.787	19.94	1	12.98	19.35	1
20.57	17.77	0	11.6	12.84	1	11.22	19.86	1
19.69	21.25	0	6.981	13.43	1	11.25	14.78	1
11.42	20.38	1	12.18	20.52	1	12.3	19.02	1
20.29	14.34	0	9.876	19.4	1	17.06	21	1
12.45	15.7	1	10.49	19.29	1	12.99	14.23	1
18.25	19.98	0	11.43	17.31	1	18.77	21.43	0
13.71	20.83	1	12.9	15.92	1	11.51	23.93	1
13	21.82	1	10.75	14.97	1	14.05	27.15	1
12.46	26.04	1	11.9	14.65	1	11.2	29.37	1
17.14	16.4	0	14.95	18.77	1	15.22	30.62	1
14.58	21.53	1	14.44	15.18	1	20.92	25.09	0
18.61	20.25	0	13.74	17.91	1	21.56	22.39	0
15.3	25.27	1	13	20.78	1	20.13	28.25	0
17.57	15.05	0	8.219	20.7	1	16.6	28.08	1
18.63	25.11	0	9.731	15.34	1	20.6	29.33	0
11.84	18.7	1	11.15	13.08	1	7.76	24.54	1
17.02	23.98	1	13.15	15.34	1	20.64	17.35	0
19.27	26.47	0	12.25	17.94	1	13.69	16.07	1
16.13	17.88	1	16.78	18.8	1	16.17	16.07	1
17.2	24.52	1	17.47	24.68	1	10.57	20.22	1
13.8	15.79	1	13.43	19.63	1	13.46	28.21	1
18.05	16.15	0	15.46	11.89	1	13.66	15.15	1
20.18	23.97	0	13.77	22.29	1	11.08	18.83	1
25.22	24.91	0	18.08	21.84	0	14.68	20.13	1
19.1	26.29	0	19.18	22.49	0	16.13	20.68	1
10.82	24.21	1	18.22	18.87	0	19.81	22.15	0
10.86	21.48	1	9.042	18.9	1	13.54	14.36	1
11.13	22.44	1	12.43	17	1	13.08	15.71	1
12.77	29.43	1	10.25	16.18	1	9.504	12.44	1
9.33	21.94	1	20.16	19.66	0	16.5	18.29	1
12.88	28.92	1	12.86	13.32	1	13.4	16.95	1
10.29	27.61	1	20.34	21.51	0	20.44	21.78	0
10.16	19.59	1	12.2	15.21	1	20.2	26.83	0
9.423	27.88	1	12.67	17.3	1	12.21	18.02	1
14.59	22.68	1	14.11	12.88	1	21.71	17.25	0
12.36	21.8	1	12.03	17.93	1	11.95	14.96	1
14.64	15.24	1	16.27	20.71	1	11.66	17.07	1
14.62	24.02	1	16.26	21.88	1	15.75	19.22	1

12.18	17.84	1	16.03	15.51	1	25.73	17.46	0
-------	-------	---	-------	-------	---	-------	-------	---

Persentase hasil akurasi pengklasifikasi data latih 76.04% sedangkan presentase hasil akurasi data uji 83.33% yang digambarkan pada gambar 2.

Akurasi Proses Latih : 76.04 %
 [0 0 0 1 0 1 0 1 1 1 0 1 0 1 0 0 1 1 0 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
 1 0 0 0 1 1 1 0 1 0 1
 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 0 1 0]
 Akurasi Proses Uji : 83.33 %
 [1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 1 0 1 1 1 0]

Gambar 2. Akurasi Pengklasifikasian Data Latih dan Akurasi Pengklasifikasian Data Uji.

5. KESIMPULAN

Klasifikasi *breast cancer* ke dalam kategori benign dan malign menggunakan 120 data pasien dengan 2 atribut, yaitu radius dan tekstur memiliki presentase akurasi untuk data latih 76.04% dan data ujinya 83.33%.

DAFTAR PUSTAKA

- [1] Kementerian Kesehatan Republik Indonesia. 2016. INFODATIN Bulan Peduli Kanker Payudara (Pusat Data dan Informasi Kementerian Kesehatan RI). Jakarta Selatan, Indonesia : Kementerian Kesehatan Republik Indonesia.
- [2] Artha, Dheby, dkk. 2019. Klasifikasi Pengidap Kanker Payudara Menggunakan Metode Voting Based Extreme Learning Machine (V-ELM). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol 3 No. 3 hlm 2180-2186. e-ISSN : xxx-xxx
- [3] Aisyah, B., & Sulisty, Y. 2016. Klasifikasi Kanker Payudara Menggunakan Algoritma Gain Ratio. 8(2), 2–5.
- [4] Chazar, C., & Widhiaputra, B. E. 2020. Informasi (Jurnal Informatika dan Sistem Informasi) Volume 12 No. 1/Mei/2020. 12(1), 67–80.
- [5] Prahartiwi, L., I., & Dari, W. 2021. Jurnal Teknik Komputer AMIK BSI. Volume 7, No. 1. Januari 2021. P-ISSN 2422-2436. E-ISSN: 2550-0120.
- [6] Octa Herlina, Thomas Sri Widodo, dan Indah Soesanti. 2012. Klasifikasi Nomsupervisde Citra Thermal Kanker Payudara berbasis Fuzzy C-Means. JNTETI, Vol.1.
- [7] Fourina Ayu Novianti dan Santi Wulan Purnami. 2012. Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regesi Logistik dan Support Vector Machine (SVM) berdasarkan hasil Mamografi. Institut Teknologi Sepuluh Nopember : Jurnal SAINS dan Seni ITS Vo. 1
- [8] Sugianto, N. A., Cholissodun, I., & Widodo, A. W., 2018. Klasifikasi Keminatan Menggunakan Algoritma Extreme Learning Machine dan Particle Swarm Optimization untuk Seleksi Fitur. Jurnal Pengembangan Teknologi Informasi dan Komputer, 2(5), p.1856-1865.
- [9] Purnama, Bedy. 2019. Pengantar Machine Learning Konsep dan Praktikum dengan Contoh Latihan Berbasis R dan Python. Bandung : Informatika Bandung.