Prediksi Jenis Kanker Payudara Menggunakan Model Regresi Logistik

Jason Hadinata/10819013
Program Studi Sarjana Aktuaria, Institut Teknologi Bandung

AK4082: Model Linier Lanjut
Dra. Dumaria Rulina Tampubolon, M.Sc., Ph.D.

23 Mei 2022

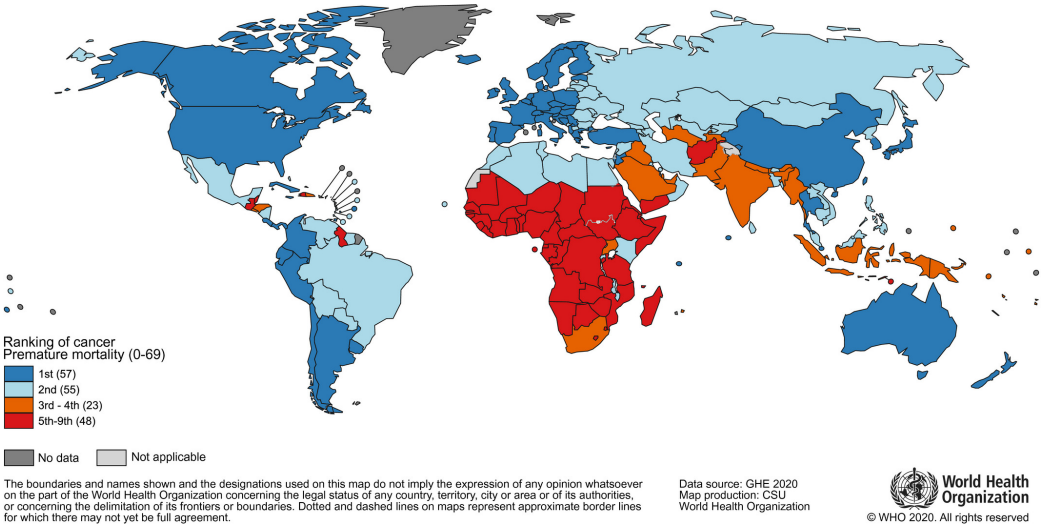
1 Pendahuluan

Beberapa bulan lalu lalu, salah seorang teman dekat saya tiba-tiba mempublikasikan foto dirinya pergi ke suatu tempat menggunakan pesawat. Saya tidak tahu ke mana ia pergi. Ia juga tidak memberi tahu ke mana ia pergi ketika saya tanya lewat pesan instan. Tentunya tidak memberi tahu saya adalah haknya. Saya pun kemudian tidak terlalu memikirkannya.

Beberapa minggu setelah itu, saya berkesempatan untuk bertemu dengannya. Saat itulah akhirnya ia menceritakan mengapa ia pergi. Teman saya bercerita bahwa ia pergi ke Penang untuk berobat. "Kamu sakit apa?" saya tanya. "Kemarin aku baru saja operasi pengangkatan tumor di dada."

Menurut website Alodokter [1], tumor adalah benjolan yang muncul akibat sel-sel tubuh tumbuh secara berlebihan. Kondisi ini terjadi ketika sel lama yang seharusnya mati masih terus bertahan hidup, sementara pembentukan sel-sel baru terus terjadi. Tumor dibedakan menjadi dua, yaitu tumor jinak dan tumor ganas. Tumor jinak tumbuh dengan lambat dan tidak menyebar sedangkan tumor ganas tumbuh lebih cepat dan menyebar ke jaringan lainnya [3]. Tumor ganas biasa disebut juga kanker.

Kanker adalah salah satu penyakit penyebab kematian prematur. Gambar 1 menunjukkan peringkat kanker di berbagai negara. Karena kanker bersifat menyebar, deteksi dini dapat membantu mempercepat bahkan mencegah kanker. Dengan demikian, suatu model yang bisa memprediksi jenis kanker menjadi penting dalam membantu pengambilan keputusan.



Gambar 1: Peringkat kanker sebagai penyebab kematian prematur (0-69 tahun) di 183 negara. Sumber: World Health Organization.

Secara khusus, akan dibahas kanker payudara dalam pemodelan ini. Kanker payudara adalah jenis kanker yang paling banyak diderita wanita. Pada tahun 2020, kanker ini menjadi jenis kanker yang paling banyak memakan korban jiwa wanita di 110 dari 183 negara [7]. Hal ini menimbulkan pertanyaan, "Apa saja variabel yang menentukan jenis suatu kanker payudara?" dan "Bagaimana model yang bisa memprediksi jenis suatu kanker payudara?"

2 Metodologi

Penelitian ini bersifat kuantitatif. Data yang digunakan adalah *Breast Cancer Wisconsin (Diagnostic) Data Set* yang bisa diakses di UCI Machine Learning Repository [2]. Dataset ini berisi 569 kasus tumor di payudara yang dilengkapi dengan jenis tumor dan sepuluh variabel bebas numerik yang sudah diagregat. Nilai dari variabel-variabel numerik ini dihitung dari gambar digital fine needle aspirate (FNA) dari massa tumor. Tidak terdapat entri kosong di data. Nama dan deskripsi tiap kolom untuk suatu observasi dapat dilihat pada tabel 1.

No.	Kolom	Deskripsi
1	`id`	Nomor ID pasien
2	`diagnosis`	`B` = tumor jinak, `M` = tumor ganas
3	`radius_*`	Rata-rata jarak dari pusat ke ujung
4	`texture_*`	Simpangan baku dari nilai gray-scale
5	`perimeter_*`	Keliling dari tumor
6	`area_*`	Luas dari tumor
7	`smoothness_*`	Variasi lokal dari `radius`
8	`compactness_*`	$\texttt{perimeter}^2/\texttt{area} - 1.0$)
9	`concavity_*`	Keparahan dari bagian cekung kontur
10	`concave points_*`	Banyak bagian cekung kontur
11	`symmetry_*`	Kesimetrian bentuk tumor
12	`fractal dimension_*`	"Coastline approximation" -1

Tabel 1: Daftar kolom dan deskripsi dari dataset *Wisconsin Breast Cancer*. Tiap kolom numerik diagregat menjadi statistik `*mean` (rata-rata), `*se` (standard error), dan "`*worst`" (rata-rata tiga nilai "terburuk").

Model yang digunakan adalah model regresi logistik dengan fungsi link logit. Misal $Y \sim B(1, p)$ adalah peubah acak yang akan diregresi oleh prediktor-prediktor dengan matriks desain $\mathbf{X}$ dan vektor parameter $\boldsymbol{\beta}$. Model regresi logistik dengan link logit memodelkan p dapat dilihat pada persamaan (1).

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}'\boldsymbol{\beta} \iff p = \frac{1}{1-\exp\left(-\mathbf{X}'\boldsymbol{\beta}\right)} \tag{1}$$

Vektor parameter $\boldsymbol{\beta}$ tersebut dicari menggunakan metode maximum likelihood secara numerik dengan metode Newton-Raphson. Model terbaik yang akan dipilih adalah yang meminimumkan nilai Akaike Information Criterion (AIC) dengan cara stepwise regression dua arah dari "model kosong". Terakhir, performa model akan dilihat dari tabel kontingensi dan nilai luas di bawah kurva (*area under curve*/AUC) *receiver operating characteristic* (ROC). [4]

Tabel kontingensi dari model logistik adalah sebagai berikut

		Actual	
		F	T
Predicted	F	TN	FN
	T	FP	TP

Tabel 2: Tabel kontingensi model logistik.

Actual menandakan nilai sebenarnya sedangkan predikted menandakan hasil prediksi model. F menandakan tidak terjadinya kejadian sedangkan T menandakan terjadinya kejadian. TN adalah true negative atau kejadian tidak terjadinya kejadian yang berhasil diprediksi tidak terjadi. TP adalah true positive dengan interpretasi kejadian terjadi yang berhasil diprediksi terjadi. FP adalah false positive atau galat tipe 1 sedangkan FN adalah false negative atau galat tipe 2.

Taraf signifikansi yang akan digunakan dalam penelitian ini adalah 5%. Beberapa metrik yang akan ditinjau juga adalah Sensitivity $= Se = \frac{TP}{FN+TP}$, Specificity $= Sp = \frac{TN}{TN+FP}$, dan Accuracy $= \frac{TN+TP}{TN+FN+FP+TP}$. Kurva ROC kemudian bisa didapat dengan memetakan $1 - Sp$ di sumbu x dan Se di sumbu y untuk berbagai nilai threshold.

Bila prediksi peluang untuk observasi ke-i $\hat{p}_i$ lebih besar dari suatu threshold, observasi tersebut diklasifikasikan sebagai kejadian terjadi (predicted T). Jika ternyata nilai tersebut lebih rendah dari threshold, observasi tersebut diklasifikan sebagai kejadian tidak terjadi (predicted F). Nilai threshold optimal akan ditentukan menggunakan statistik Youden $J = Se + Sp - 1$. Threshold yang dipilih adalah yang memaksimumkan nilai J [9].

3 Analisis

Analisis data dilakukan menggunakan Python[1]. Secara spesifik, data cleaning dilakukan menggunakan package Pandas [8] dan pemodelan dilakukan menggunakan package Statsmodels [6]. Cuplikan dari data dapat dilihat pada tabel 3.

id	diagnosis	radius_mean	texture_mean	perimeter_mean
87930	B	12.47	18.60	81.09
859575	M	18.94	21.31	123.60
8670	M	15.46	19.48	101.70
907915	B	12.40	17.68	81.47
921385	B	11.54	14.44	74.65

Tabel 3: Lima observasi sembarang dan empat kolom pertama dari dataset *Wisconsin Breast Cancer*.

Untuk melakukan regresi terhadap `diagnosis`, entri kolom tersebut diubah dengan memetakan `M` ke 1 dan `B` ke 0. Dengan demikian, entri ke-i kolom `diagnosis` bisa dipandang berdistribusi $B(1, p_i)$ dengan kejadian `diagnosis` $= 1$ sama dengan kejadian tumor ganas dan kejadian `diagnosis` $= 0$ sama dengan kejadian tumor jinak. Cuplikan hasil pemetaan tersebut dapat dilihat pada tabel 4.

id	diagnosis	radius_mean	texture_mean	perimeter_mean
87930	0	12.47	18.60	81.09
859575	1	18.94	21.31	123.60
8670	1	15.46	19.48	101.70
907915	0	12.40	17.68	81.47
921385	0	11.54	14.44	74.65

Tabel 4: Lima observasi sembarang `diagnosis` yang sudah dipetakan ke 0 atau 1.

[1]Source code dapat diakses dari GitHub repository https://github.com/jasonhadiputra/wisconsin-breast-cancer

Tidak terdapat entri kosong di dataset. Dengan demikian, model regresi logistik akan langsung digunakan pada data. Hasil stepwise regression yang meminimumkan nilai AIC dapat dilihat pada tabel 5. Nilai AIC model tersebut adalah 73.097. Variabel `texture_se` tidak signifikan pada taraf kepercayaan 5%. Namun, p-value dari variabel tersebut cukup dekat dengan 5% dan mengeluarkannya dari model akan meningkatkan nilai AIC sehingga variabel tersebut akan tetap dianggap dalam model. Taksiran vektor parameter $\boldsymbol{\beta}$ dapat dilihat pada kolom `coef`.

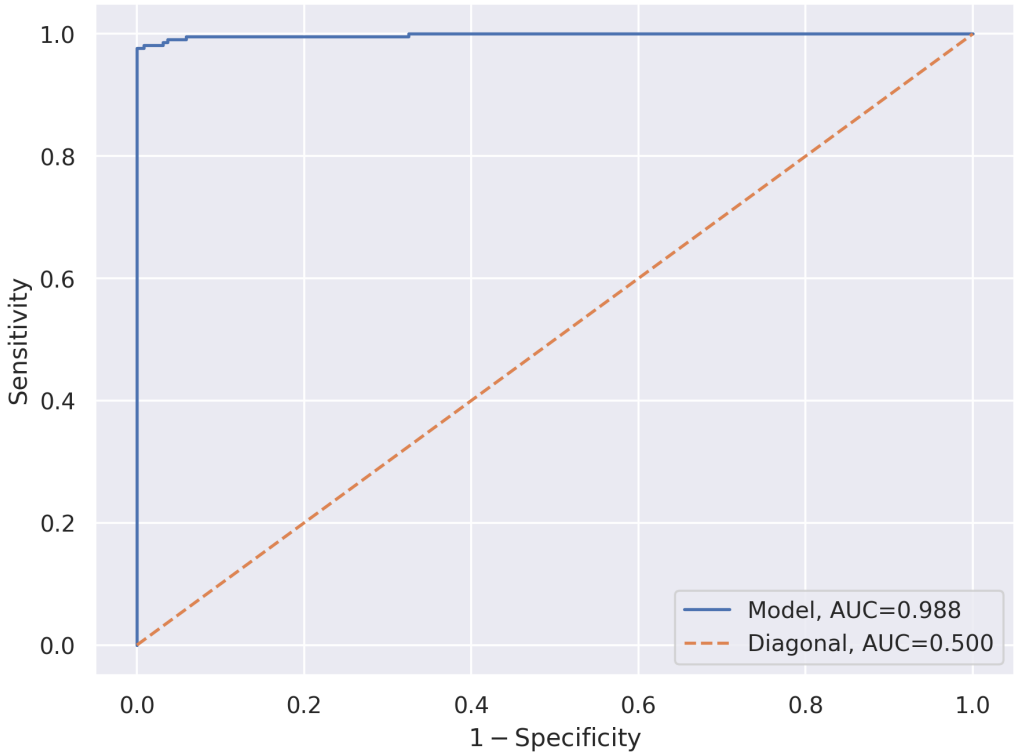
Dep. Variable:	diagnosis	**No. Observations:**	569
Model:	Logit	**Df Residuals:**	559
Method:	MLE	**Df Model:**	9
Date:	Fri, 06 May 2022	**Pseudo R-squ.:**	0.9293
Time:	09:14:20	**Log-Likelihood:**	-26.548
converged:	True	**LL-Null:**	-375.72
Covariance Type:	nonrobust	**LLR p-value:**	1.571e-144

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-44.3757	8.931	-4.969	0.000	-61.880	-26.871
Q('smoothness_worst')	65.5417	30.007	2.184	0.029	6.729	124.354
Q('texture_worst')	0.5412	0.134	4.031	0.000	0.278	0.804
Q('symmetry_worst')	15.6127	7.473	2.089	0.037	0.966	30.259
Q('compactness_se')	-116.5703	36.229	-3.218	0.001	-187.578	-45.563
Q('concavity_mean')	37.3743	14.197	2.632	0.008	9.548	65.201
Q('texture_se')	-2.6112	1.472	-1.774	0.076	-5.496	0.274
Q('area_se')	0.2265	0.068	3.309	0.001	0.092	0.361
Q('concave points_worst')	39.1800	19.514	2.008	0.045	0.934	77.426
Q('area_worst')	0.0096	0.003	2.771	0.006	0.003	0.016

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

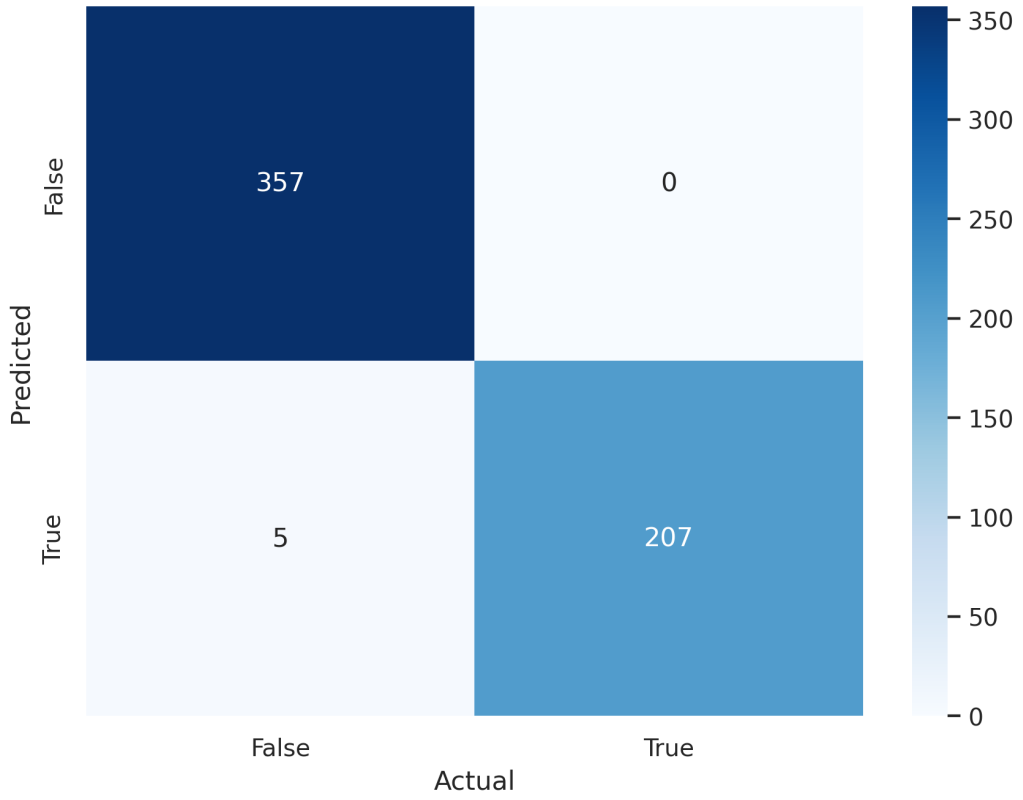
Tabel 5: Hasil fitting model regresi final.

Kurva ROC dan nilai AUC yang bersesuaian dapat dilihat pada gambar 2. Dengan demikian, model bisa dianggap sebagai pembeda jenis kanker yang luar biasa (outstanding) [5]. Didapatkan pula nilai threshold yang memaksimumkan $Se + Sp - 1$ adalah 0.6754665480564337.



Gambar 2: Kurva ROC dan nilai AUC dari model final.

Tabel kontingensi dari model dengan threshold sekitar 0.68 diberikan pada gambar 3. Akurasi dari model mencapai 99.12% dengan $Se = 97.64\%$ dan $Sp = 100.00\%$.



Gambar 3: Tabel kontingensi hasil prediksi (confusion matrix).

4 Kesimpulan

Variabel yang menentukan jenis suatu kanker payudara adalah `smoothness_worst`, `texture_worst`, `symmetry_worst`, `compactness_se`, `concavity_mean`, `texture_se`, `area_se`, `concave points_worst`, dan `area_worst`. Model terbaik yang meminimumkan nilai AIC didapat dapat dilihat pada persamaan 2.

$$
\begin{aligned}
&Diagnosis \sim B(1,p)\\
&\text{logit}(p) = -44.3757 + 65.5417(\texttt{smoothness\_worst}) + 0.5412(\texttt{texture\_worst})\\
&\qquad\quad + 15.6127(\texttt{symmetry\_worst}) - 116.5703(\texttt{compactness\_se})\\
&\qquad\quad + 37.3743(\texttt{concavity\_mean}) - 2.6112(\texttt{texture\_se}) + 0.2265(\texttt{area\_se})\\
&\qquad\quad + 39.18(\texttt{concave points\_worst} + 0.0096(\texttt{area\_worst})\\
&\text{logit}(p) = \log\left(\frac{p}{1-p}\right)
\end{aligned}
\tag{2}
$$

Nilai threshold terbaik yang memaksimumkan statsitik Youden $J = Se + Sp - 1$ adalah 67.55%. Model ini dapat diklasifikasikan sebagai pembeda yang luar biasa (outstanding) [5]. Source code dapat diakses dari GitHub repository https://github.com/jasonhadiputra/wisconsin-breast-cancer.

Beruntungnya, tumor yang diangkat dari teman saya adalah tumor jinak. Semua tumornya sudah diangkat dan potensi kesembuhannya baik.

Referensi

[1] Alodokter. *Tumor*. Mar. 2, 2022. URL: https://www.alodokter.com/tumor.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[3] Lisa Fayed. *Differences Between a Malignant and Benign Tumor*. Dec. 12, 2021. URL: https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240.

[4] Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data*. New York: Cambridge University Press, 2008.

[5] Jayawant N. Mandrekar. "Receiver Operating Characteristic Curve in Diagnostic Test Assessment". In: *Journal of Thoracic Oncology* 5.9 (2010), pp. 1315–1316. ISSN: 1556-0864. DOI: https://doi.org/10.1097/JTO.0b013e3181ec173d. URL: https://www.sciencedirect.com/science/article/pii/S1556086415306043.

[6] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[7] Hyuna Sung et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249. DOI: https://doi.org/10.3322/caac.21660. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

[8] The pandas development team. *pandas-dev/pandas: Pandas*. Version 1.3.5. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[9] W. J. Youden. "Index for rating diagnostic tests". In: *Cancer* 3.1 (1950), pp. 32–35. DOI: https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3.