



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi 357

**PERBANDINGAN DI ANTARA ANALISIS POHON PENGELASAN DAN
REGRESI LOGISTIK: APLIKASI TERHADAP DATA BERAT
KELAHIRAN DI UNIT RAWATAN RAPI NEONATAL**

NOOR WAHIDA BT MD JUNUS



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

**DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS**

**FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI**

2011



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi



PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya



10 June 2011

NOOR WAHIDA BT MD JUNUS
P 50043





PENGHARGAAN

Syukur Alhamdulillah saya panjatkan kesyukuran saya kepada Allah SWT kerana dengan limpah rahmat dan kurniaan-Nya memberi saya kesihatan yang cukup, masa dan kematangan fikiran dalam menyiapkan projek penyelidikan ini sehingga ke tahap ini. Jutaan terima kasih yang rasanya tidak terbalas saya ucapkan kepada penyelia projek penyelidikan saya Prof. Dr Abdul Aziz b. Jemain yang banyak meluangkan waktu beliau untuk mengulas, memberi idea dan nasihat sepanjang projek penyelidikan ini dijalankan.

Tidak lupa juga kepada ketiga-tiga pensyarah statistik Prof Dr. Kamarulzaman b Ibrahim, Prof Madya Dr Ahmad Mahir b. Razali dan Prof Madya Najib Mahmood Rafee dengan kepakaran masing yang banyak memberi ilmu kepada saya dan rakan-rakan sepanjang pengajian kami di UKM ini.

Terima kasih yang tak terhingga juga saya ucapkan kepada pihak pengurusan pusat pengajian Sains dan Matematik (PPSM) dan sekretariat siswazah Fakulti Sains dan Teknologi atas keprihatinan anda kepada pelajar siswazah dan tidak lupa juga kepada Kementerian Pengajian Tinggi dan Universiti Pendidikan Sultan Idris di atas tajaan pengajian saya di bawah Skim Lanjutan Akademik Bumiputera yang banyak membantu dari segi kewangan sepanjang pengajian saya sehingga ke tahap ini.



Ucapan terima kasih juga saya ucapkan buat rakan-rakan seperjuangan yang selalu membantu menghulurkan semangat, dan memberi idea sepanjang projek penyelidikan ini dijalankan dan akhir sekali buat ibu dan kaum keluarga tersayang terima kasih atas doa kalian untuk melihat saya berjaya akhirnya dimakbulkan Allah. Alhamdulillah.

Sekian Terima Kasih



ABSTRAK

Kajian ini bermatlamat untuk membandingkan analisis regresi logistik dan pohon pengelasan dalam menentukan faktor bayi berat lahir rendah (BBLR) pada 7260 bayi yang ditempatkan di pusat rawatan rapi neonatal seperti yang telah didaftarkan dalam *Malaysian National Neonatal Registry* pada tahun 2006. Sampel yang dipilih adalah berdasarkan tiga etnik utama iaitu Melayu, Cina dan India. Kaedah analisis yang digunakan ialah dengan membandingkan keputusan yang diperolehi daripada kedua-dua kaedah yang digunakan iaitu analisis regresi logistik dan analisis pohon pengelasan. Analisis menunjukkan kedua-dua kaedah menghasilkan keputusan yang hampir sama di mana minggu gestasi adalah faktor utama kepada BBLR. Pohon pengelasan menunjukkan struktur biologi yang lebih baik berbanding regresi logistik. Walau bagaimanapun hasil dari analisis regresi logistik mempunyai kepekaan yang lebih tinggi. Secara kesimpulannya kedua-dua kaedah ini mempunyai kelebihan tersendiri. Gabungan kedua-dua kaedah mungkin dapat menghasilkan keputusan yang lebih baik.



THE COMPARISON OF CLASSIFICATION TREE AND LOGISTIC REGRESSION ANALYSIS : APPLICATION TO BIRTH WEIGHT DATA IN NEONATAL INTENSIVE CARE UNIT

ABSTRACT

This study aimed to compare the logistic regression and classification tree analysis in determine the factors of low birth weight (LBW) for 7260 baby in the neonatal intensive care unit (NICU) as registered in the Malaysian National Neonatal Registry in 2006. Samples are selected based on three major ethnic Malays, Chinese and Indians. The method of analysis was to compare the study outcome obtained from both methods of analysis which is logistic regression and classification tree analysis. Analysis showed that both methods produce similar results in which the gestational weeks is a main factor of LBW. Classification tree shows the structure of biology is better than logistic regression. However, the results of logistic regression analysis have higher sensitivity than classification tree. In conclusion, both methods have their own advantages. The combination of both methods may produce better results.



KANDUNGAN

	Halaman
PENGAKUAN	ii
PENGHARGAAN	iii
ABSTRAK	iv
ABSTRACT	v
KANDUNGAN	vi
SENARAI JADUAL	xi
SENARAI ILUSTRASI	x

BAB I PENDAHULUAN

1.1	Pengenalan	1
1.1.1	Analisis pohon pengelasan	1
1.1.2	Analisis regresi logistik	4
1.1.3	Berat kelahiran	6
1.2	Pernyataan Masalah	6
1.3	Objektif Kajian	7
1.4	Kekuatan dan Kepentingan Kajian	7

BAB II TINJAUAN KESUSATERAAAN

2.1	Pengenalan	8
2.2	Kaedah Analisis	8
2.2.1	Analisis pohon pengelasan	8
2.2.2	Perbandingan analisis pohon pengelasan dan regresi logistik	9
2.3	Faktor Risiko berat bayi lahir rendah	10
2.3.1	Umur ibu	10
2.3.2	Minggu gestasi kelahiran	10
2.3.3	Bangsa	11
2.3.4	Kandungan berganda	11
2.3.5	Jantina	11



2.4	Penutup	11
-----	---------	----

BAB III KAEDAH KAJIAN

3.1	Pengenalan	12
3.2	Reka bentuk Kajian	12
3.3	Populasi dan Pensampelan	12
3.4	Kaedah Analisis	13
3.5	Kaedah Analisis Pohon Pengelasan	14
3.5.1	Pemilihan faktor	14
3.5.2	Pemilihan rekursif	16
3.5.3	Penentuan terminal nods	17
3.5.4	Penentuan label kelas	17
3.5.5	Penentuan pohon terbaik	17
3.6	Analisis Regresi Logistik	21
3.6.1	Kriteria pemilihan pemboleh ubah peramal	23
3.6.2	Validasi statistik dan prosedur kajian	24
3.6.3	Nisbah odd	24
3.7	Penutup	25

BAB IV ANALISIS DAN KEPUTUSAN

4.1	Pengenalan	26
4.2	Analisis Deskriptif	26
4.3	Analisis Pohon Pengelasan	27
4.3.1	Ketepatan ramalan pengelasan	32
4.4	Analisis Regresi Logistik	33
4.4.1	Analisis univariat	33
4.4.2	Analisis multivariat	34
4.4.3	Anggaran model terbaik	35
4.4.4	Ketepatan ramalan pengelasan	36
4.5	Perbandingan di antara Pohon Pengelasan dan Model Regresi Logistik	37
4.6	Penutup	37



BAB V KESIMPULAN DAN CADANGAN

5.1	Pengenalan	38
5.2	Kesimpulan Kajian	38
5.3	Cadangan Kajian	39
5.3.1	Pengkaji	39
5.3.2	Pihak awam dan swasta	40
	RUJUKAN	41
	LAMPIRAN	44
A	Output R: Analisis Pohon Pengelasan	44
B	Output R: Analisis Regresi Logistik	49

SENARAI JADUAL

No. Jadual		Halaman
1.1	Istilah yang digunakan dalam pohon pengelasan	2
3.1	Senarai pemboleh ubah yang digunakan dalam model	13
3.2	Kemungkinan boleh pisah setiap pemboleh ubah peramal	15
3.3	Jadual 2 x 2 untuk pemisahan pemboleh ubah umur ibu	16
3.4	Interpretasi untuk nisbah odd	25
4.1	Deskriptif statistik faktor risiko bayi berat lahir rendah	26
4.2	Ukuran kos kompleksiti untuk semua pohon yang mungkin	27
4.3	Ramalan pengelasan untuk pohon pengelasan terbaik	32
4.4	Analisis univariat untuk faktor kepada BBLR	33
4.5	Analisis multivariat untuk faktor BBLR	34
4.6	Nisbah odd untuk faktor BBLR	36
4.7	Kadar salah pengelasan regresi logistik	36

SENARAI ILUSTRASI

No. Rajah		Halaman
1.1	Contoh pemisahan rekursif dalam pohon pengelasan	3
1.2	Lengkung model logistik	5
3.1	Kerangka ujian regresi logistik	22
4.1	Model analisis pohon pengelasan untuk faktor BBLR	29

BAB I

PENDAHULUAN

1.1 PENGENALAN

Pengelasan ialah satu kaedah mengumpulkan sesuatu objek ke dalam satu atau beberapa kumpulan yang mempunyai ciri-ciri yang sama. Antara kaedah analisis statistik yang digunakan untuk pengelasan ialah analisis diskriminan dan regresi logistik. Kedua-dua kaedah ini bersifat parametrik, di mana andaian atau *assumption* awal tentang taburan data perlu dipenuhi sebelum menggunakan analisis ini. Walau bagaimanapun satu lagi kaedah pengelasan yang lebih menarik dan tidak berparameter yang semakin popular ialah analisis pohon pengelasan. Analisis ini adalah satu analisis yang tidak berparameter dan tidak memerlukan andaian awal tentang taburan data. Kaedah ini juga sangat sesuai untuk data yang besar dan menghasilkan satu model yang lebih menarik dan mudah difahami. Oleh yang demikian kajian ini bertujuan membanding keputusan yang diperoleh melalui analisis pohon pengelasan dan regresi logistik dengan mengaplikasikan data berat kelahiran bayi di unit rawatan rapi neonatal.

1.1.1 Analisis Pohon Pengelasan

Analisis pohon pengelasan adalah salah satu kaedah statistik tidak berparameter dan tidak memerlukan sebarang andaian awal tentang taburan statistik sesuatu data atau pemboleh ubah yang digunakan. Analisis ini menghasilkan satu model ramalan yang menarik dan mudah difahami oleh semua golongan. Matlamat utama kaedah ini ialah untuk mengelaskan satu set data yang besar kepada kelas tertentu berdasarkan



persoalan pengkaji. Selain itu kaedah ini membolehkan seseorang meramalkan kelas untuk cerapan yang akan datang bagi data yang mempunyai profil sama.

Kaedah pohon pengelasan ialah melalui pemisahan rekursif langkah demi langkah yang mana satu pohon keputusan dibina dengan memisahkan atau tidak memisahkan setiap nod induk kepada nod anak. Kaedah pemisahan ini menggunakan soalan Boolean hierarki algoritma yang menarik dan mudah difahami. Contohnya adakah $X_i < \theta_j$ dimana θ_j salah satu nilai dalam X_i iaitu pembolehubah peramal.

Di dalam pohon pengelasan pemboleh ubah respons ialah pemboleh ubah binari, di mana dalam kajian ini pemboleh ubahnya ialah berat kelahiran bayi, sama ada berat lahir rendah (di kod sebagai 1) dan berat lahir normal (di kod sebagai 0). Pembolehubah peramal pula boleh sama ada ordinal atau nominal.

Terdapat beberapa terma yang digunakan dalam pohon pengelasan seperti dalam

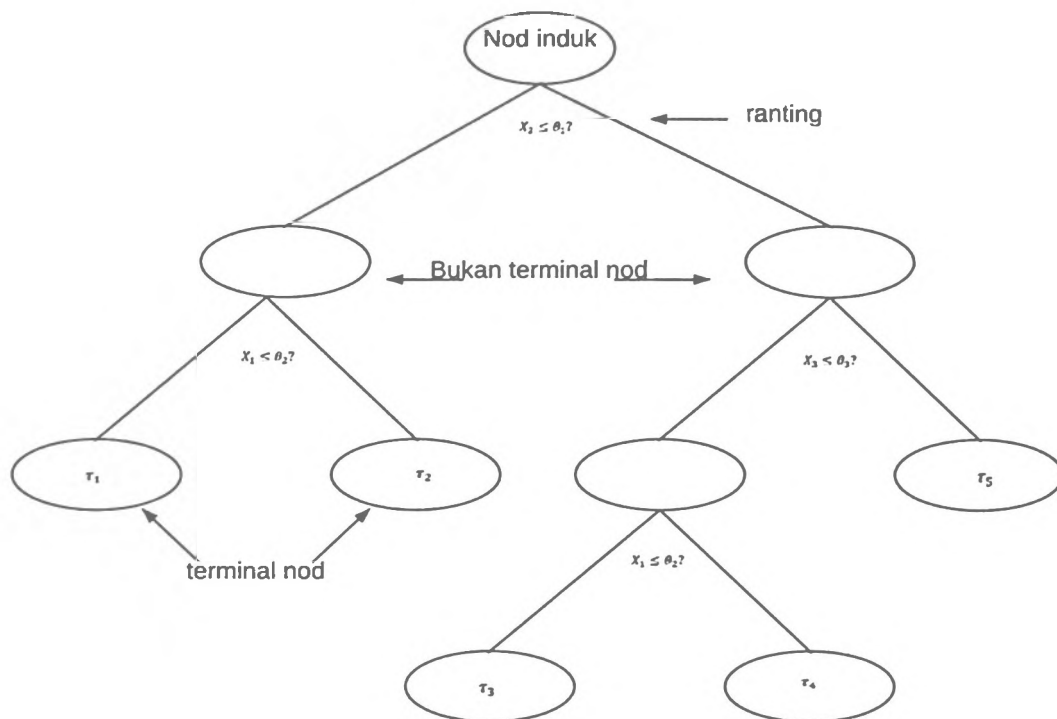
jadual 1.1



Jadual 1.1 Istilah yang digunakan dalam pohon pengelasan

Terma	Definisi
Nod	Sub set kepada pemboleh ubah
nod akar(root nod)	Pada nod ini semua cerapan $\mathcal{L}$ yang ingin dikaji dikumpul
Bukan terminal nod	Nod induk yang masih boleh dipisahkan kepada nod anak
Terminal nod	Nod anak yang sudah tidak boleh dipisahkan dan kelas nod ditentukan
Ranting	Mewakili pemboleh ubah mengikut soalan Boolean hierarki algoritma





Rajah 1.1 Contoh pemisahan rekursif dalam pohon pengelasan

Setiap pohon pengelasan bermula dengan nod akar, pada nod ini semua cerapan $\mathcal{L}$ yang ingin dikaji dikumpulkan. Nod boleh sebagai terminal atau bukan terminal. Nod bukan terminal ialah nod induk yang boleh di pisah kepada nod anak sama ada kiri atau kanan berdasarkan soalan Boolean hirarki algorithm. Nod terminal ialah nod yang tidak boleh dipisahkan lagi. Maka setiap cerapan $\mathcal{L}$ yang jatuh pada nod ini akan dikelaskan. Setiap cerapan $\mathcal{L}$ yang sampai pada mana-mana nod induk dan memenuhi sifat atau ciri pada nod tersebut akan diturunkan pada nod kiri. Manakala yang tidak mengikut akan diturunkan pada nod kanan.

Secara ringkasnya pohon pengelasan dapat diterangkan seperti rajah 1.1 di atas. Setiap cerapan $\mathcal{L}$ pada nod akar akan disoal siasat samada memenuhi ciri soalan Boolean algorithm, iaitu adakah $X_2 < \theta_1$? jika “ya” subset dari cerapan $\mathcal{L}$ akan diturunkan pada nod kiri, jika “tidak” akan diturunkan pada nod kanan. Seterusnya jika $X_2 < \theta_1$ ada lah “ya” soalan seterusnya dikemukakan adakah $X_2 < \theta_1$? Jika ya cerapan diturunkan ke kiri dan dikelaskan sebagai τ_1 dan jika “tidak” dikelaskan



sebagai τ_2 . Kedua- dua nod ini dinamakan sebagai terminal nod. Proses ini berulang pada setiap nod sehingga terminal nod ditemui.

Breiman et.al (1984) dan Steinberg & Cola (1995) menggariskan beberapa kekuatan dan kelemahan pohon pengelasan. Antaranya adalah seperti berikut:

- i. Pohon pengelasan tidak memerlukan sebarang andaian taburan untuk semua pemboleh ubahnya sama ada bersandar atau tidak bersandar. Tiada pemboleh ubah dalam Pohon Pengelasan diandaikan mengikut sebarang taburan statistik
- ii. Pemboleh ubah peramal atau penerang dalam pohon pengelasan boleh digabungkan di antara kategori atau selanjar
- iii. Mempunyai binaan algoritma untuk mengatasi masalah "missing value".
- iv. Tidak terpengaruh dengan masalah data terpencil, kolineariti, hetrokedasticity, atau sebarang ralat taburan yang mempengaruhi parameter model.
- v. Sangat sesuai untuk data set yang besar dan bilangan pembolehubah yang banyak
- vi. Mudah difahami oleh semua dalam semua disiplin
- vii. Tidak berdasarkan model berkebarangkalian
- viii. Tiada aras kebarangkalian yang ditetapkan atau aras keyakinan yang dikehendaki.

1.1.2 Analisis Regresi Logistik

Regresi logistik digunakan untuk meramal kebarangkalian berlakunya sesuatu peristiwa dengan menyesuaikan data kepada fungsi logit dengan menggunakan lengkung logistik. Model regresi logistik ialah satu model linear teritlak (GLM) yang menggunakan regresi binomial.

Seperti model regresi yang lain, ia memerlukan beberapa pemboleh ubah kovariat atau peramal yang mempunyai data numerik atau kategori. Model ini banyak digunakan dalam bidang kesihatan dan sosial sains.



Regresi logistik boleh diterangkan dengan fungsi logistik seperti berikut:

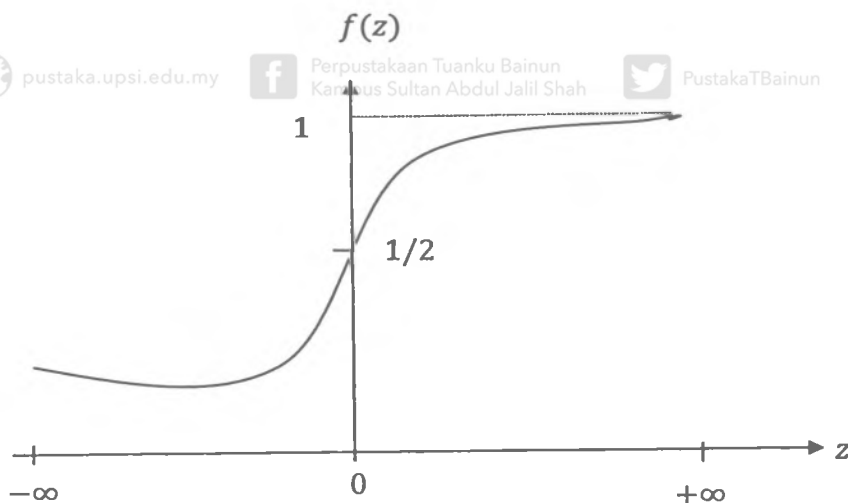
$$f(z) = \frac{1}{1 + e^{-z}} \quad (1.1)$$

Dengan z sebagai pembolehubah input dan $f(z)$ sebagai pembolehubah output. Nilai z boleh berada dalam selang $-\infty$ sehingga $+\infty$ tetapi nilai $f(z)$ hanya berada dalam julat 0 hingga 1.

z ialah ukuran jumlah sumbangan semua pembolehubah tidak bersandar yang digunakan dalam model dan dikenali sebagai logit, z boleh ditulis sebagai:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1.2)$$

Dengan β_0 sebagai pintasan graf, dan $\beta_1, \beta_2 \dots \beta_n$ sebagai pekali regresi untuk $x_1, x_2 \dots x_n$. Fungsi logistik ini boleh diwakili dengan lengkung logistik seperti dalam rajah 1.2 berikut.



Rajah 1.2 Lengkung Model Logistik

$f(z)$ dikenali juga sebagai pemboleh ubah pendedahan (*exposure*) seperti dalam kajian ini $f(z)$ ialah berat bayi dikelaskan 0 sebagai berat normal dan 1 sebagai berat lahir rendah (BBLR).

1.1.3 Berat Kelahiran

Menurut Valero et al.(2004) mendefinisikan berat kelahiran sebagai berat pertama fetus atau bayi yang mana ukuran berat ini diambil pada jam pertama kelahiran. Bayi ini dikategorikan sebagai bayi berat lahir rendah (BBLR) apabila beratnya kurang daripada 2500 gram dan mempunyai berat normal apabila mencapai lebih daripada 2500 gram.

Valero et al. (2004) juga menyatakan yang bayi berat lahir rendah dibahagikan kepada 3 kumpulan iaitu:

- 1) BBLR pra matang iaitu bayi yang lahir sebelum cukup 37 minggu gestasi atau kurang dari 259 hari gestasi dengan berat rendah
- 2) BBLR matang iaitu bayi yang lahir di antara minggu ke 37 dan 42 minggu gestasi atau di antara 259 dan 293 hari gestasi dengan berat kelahiran rendah.
- 3) BBLR terlalu matang (*post term*) iaitu bayi yang lahir selepas 42 minggu gestasi atau selepas 294 hari gestasi.

Kajian tentang BBLR ini sangat penting kerana pada abad ini kejadian BBLR adalah pembunuh kedua bayi selepas kelahiran bayi pramatang (Valero et al 2004).

1.2 PERNYATAAN MASALAH

Kajian tentang faktor kepada kelahiran bayi berat lahir rendah (BBLR) telah banyak dijalankan oleh ramai pengkaji di pelosok dunia. Kebanyakan pengkaji menggunakan analisis regresi logistik untuk menentukan faktor BBLR, antaranya kajian dari Yin dan Tzu (2005), dan Jafari et al (2010). Analisis regresi logistik biasa digunakan untuk menentukan faktor risiko dalam kajian perubatan, namun sejak beberapa tahun kebelakangan ini penggunaan pohon pengelasan semakin diminati kerana kaedah ini menggunakan perwakilan simbol yang mudah diinterpretasi dan difahami oleh semua golongan (Abu-Hanna & de Keizer 2003; Breiman et al. 1984; Fu 2004; Kline et al. 2003; Robnik-Sikonja, Cukjati, & Kononeko, 2003). Penggunaan kaedah ini dapat



memudahkan semua golongan memahami sesuatu kajian walaupun beliau bukan professional dalam bidang yang dikaji. Selain itu dengan kemajuan teknologi yang semakin pesat proses perhitungan yang dulu sukar kini lebih mudah dibuat dengan penggunaan perisian statistik yang pelbagai. Antaranya pakej *rpart* dalam perisian R memudahkan lagi pembinaan pohon pengelasan dibuat. Oleh yang demikian kajian ini bertujuan untuk melihat perbezaan di antara dua kaedah analisis iaitu analisis regresi logistik dan analisis pohon pengelasan dari segi keputusan yang diperoleh.

1.3 OBJEKTIF KAJIAN

- i. Membandingkan hasil analisis pohon regresi dengan regresi logistik
- ii. Menghasilkan satu model yang mudah difahami untuk rujukan semua golongan
- iii. Mengenal pasti faktor risiko bayi berat lahir rendah.

1.4 KEKUATAN DAN KEPENTINGAN KAJIAN



Kekuatan kajian ini ialah untuk melihat penggunaan model yang lebih ringkas tetapi mudah difahami oleh semua golongan. Di samping itu kajian ini juga dapat digunakan sebagai rujukan asas tentang faktor kelahiran BBLR.

