PERBANDINGAN VARIABEL DOMINAN FAKTOR RISIKO KEJADIAN BERAT BADAN LAHIR RENDAH ANTARA HASIL ANALISIS REGRESI LOGISTIK DAN POHON KLASIFIKASI

PIPIT FESTI W
Pembimbing : Dr.Hari Basuki Notobroto,dr,M.Kes.
LOGISTIC REGRESSION ANALYSIS
KKC KK TKM 05 / 11 Pit p

ABSTRACT

In health research that studies the influence of several determinants of an event used the regression method. One type of regression is a logistic regression model that is a mathematical approach that can be used to describe the relationship between the dichotomous dependent variable or polychotomus with dichotomous independent variables, polythomous and continuous. That Method has limitations on the processing of health data that can be addressed by the method of classification tree.
This research is a statistical study comparing the dominant variable risk factors event of LBW (low birth body weight) between the results of logistic regression analysis and classification tree. This research was applied on secondary data of Birth Weight at Health Department of Sumenep City and the cohort report of pregnant women at five health centers in Sumenep district. Dependent variable is low birth weight and the independent variable is maternal age, maternal education, maternal employment status, numbers of children, birth spacing, maternal hemoglobin, mothers LILA size (*Mother Upper Arm Circumference)*, increase in maternal weight and maternal height. The amount of data in this study was 337 data.
Logistic regression analysis with $\alpha = 0.05$ independent variables that affect the LBW were maternal education, maternal hemoglobin, LILA mother and Body Weight. Results of classification tree analysis on optimal tree were increase of maternal weight gain and maternal education. Analysis on maximum tree obtained the dominant variables such as increase maternal weight gain, maternal education, mothers LILA size, maternal Hemoglobin, numbers of children and total maternal weight. Result of classification accuracy showed that the logistics regression has a higher accuracy of the classification than classification tree method. The result of logistic regression classification accuracy is 80.7% higher than the optimum classification tree, 71.5% and 71.9% on maximum classification tree.

Keywords: logistic regression analysis, classification tree, dominant variable and classification accuracy

RINGKASAN

Dalam penelitian bidang kesehatan yang mempelajari pengaruh beberapa determinan terhadap suatu kejadian digunakan metode regresi. Metode regresi merupakan komponen penting dalam analisis data yang menggambarkan hubungan antara suatu variabel dependen dengan satu atau lebih variabel independen. Salah satu jenis regresi adalah regresi logistik. Regresi logistik adalah pendekatan model matematik yang dapat digunakan untuk menggambarkan hubungan antara variabel dependen yang dikotomus atau politomus dengan variabel independen yang dikotomus, politomus, atau kontinyu. Penggunaan analisis regresi logistik mempunyai beberapa keterbatasan antara lain adalah banyaknya variabel prediktor sehingga menyebabkan kesulitan dalam pemilihan variabel yang berpengaruh (terpenting), sering terdapat interaksi sesama variabel prediktor, dan model-model yang dihasilkan banyak mengalami kesulitan dalam penerapan

Untuk mengatasi keterbatasan tersebut di atas diperlukan metode yang tidak terlalu terikat dengan beberapa asumsi. *Classification And Regression Tree* (CART) merupakan sebuah metode inovatif untuk analisis data yang besar melalui prosedur pemilihan biner. Prosedur pemilihan dikenal sebagai *regression trees* (pohon regresi) bila variabel respon berupa data numerik, dan *classification trees* (pohon klasifikasi) jika variabel respon berupa data kategori

Penelitian ini merupakan kajian statistik dalam membandingkan variabel dominan faktor risiko kejadian berat badan lahir rendah antara hasil analisis regresi logistik dan pohon klasifikasi. Penelitian ini diaplikasikan pada data sekunder Berat Badan Lahir Dinas Kesehatan Kota Sumenep dan Laporan Kohort ibu hamil pada lima Puskesmas di Kabupaten Sumenep. Adapun variabel dependennya adalah Berat Badan Lahir Rendah dan variabel independennya adalah umur ibu, pendidikan ibu, status pekerjaan ibu, jumlah anak, jarak kelahiran, kadar HB ibu, ukuran LILA ibu, kenaikan berat badan ibu, dan tinggi badan ibu. Banyak data dalam penelitian ini adalah 337 data.

Berdasarkan hasil analisis regresi logistik, variabel independen yang mempengaruhi kejadian Berat Badan Lahir rendah adalah pendidikan ibu, haemoglobin ibu, LILA ibu dan penambahan berat badan. Karakteristik berat badan bayi di wilayah puskesmas di Kabupaten Sumenep dapat dikelompokkan sebagai berikut: 1. Sebagian besar ibu hamil adalah berpendidikan kurang atau sampai tingkat Sekolah Dasar sebanyak 232, sedangkan 119 atau 51.2 % ibu yang melahirkan bayi dengan berat badan kurang dari 2500 gram berpendidikan kurang atau sampai tingkat Sekolah Dasar.

2. Sebagian ibu yang melahirkan bayi dengan Berat Badan Bayi < 2500 gram adalah ibu dengan Hemoglobin kurang dari 11 g % adalah 66 bayi atau 74,15 % sedangkan untuk bayi dengan berat badan lebih atau sama dengan 2500 gram sebagian dilahirkan oleh ibu dengan HB ≥ 11 g% 186 ibu dengan persentasi 75 %.

3. Sebagian ibu yang melahirkan bayi dengan Berat badan ≥ 2500 gram memiliki ukuran LILA ≥23,5 cm sejumlah 98 ibu, sedangkan bayi dengan berat badan kurang dari 2500 gram sebagian ukuran LILA ibunya kurang dari 23,5 cm adalah 70 orang atau 71,4% .

4. Sebagian besar ibu dari 208 ibu mengalami penambahan berat badan kurang dari 9 kg. Bayi dengan berat badan kurang 2500 gram sebagian dari ibu yang mengalami peningkatan berat kurang dari 9 kg sebesar 120 orang atau 57,6 % Berdasarkan hasil analisis regresi logistik untuk menduga berat badan bayi adalah $P_{(BBLR)}()$kg) 9 BB(Penambahan 2.155+cm) 23,5LILA(1.842+gr%) 11 HB(1.224+ SD) (pendidikan 1.469+4,183-exp +11 <<<≤–= Dimana pendidikan bernilai 0 bila ≤ Sekolah Dasar dan bernilai 1 bila > Sekolah Dasar,

Penambahan berat badan bernilai 0 bila < 9 kg dan bernilai 1 bila ≥ 9 kg, LILA bernilai 0 bila < 23,5 cm dan bernilai 1 bila ≥ 23,5 cm, HB bernilai 1 bila ≥ 11 gr% dan bernilai 0 bila HB < 11 gr% Hasil Ketepatan Klasifikasi pada regresi logistik adalah 80,5% Adapun kelebihan metode pohon klasifikasi yaitu dapat mengetahui secara langsung pada variabel dominan yang terletak pada pemilah utama. Kelemahan pohon klasifikasi tidak dapat menentukan variabel mana dari data yang sangat berpengaruh didalam model, besar pengaruh tiap variabel independen tidak diketahui (p tidak ada). Pada pengolahan data dengan bantuan computer di dapatkan hasil pohon klasifikasi terbentuk 16 simpul dalam, 17 simpul terminal dan 10 kedalaman pohon klasifisikasi Interpretasi dari pohon klasifikasi didapatkan variabel dominan pada pohon Maksimal berturut turut adalah Penambahan Berat Badan dengan skor 100, Pendidikan dengan nilai 76.22, Lingkar Lengan Atas Ibu (LILA) skor 33.45, Haemoglobin ibu dengan skor 29.28, Jumlah anak skor 7.25 dan 5.84 untuk skor Tinggi Badan Ibu.Variabel dominan pada pohon optimal berturut turut adalah Penambahan Berat Badan ibu dengan skor 100 dan, Pendidikan ibu skor 63.49. Hasil pengelompokan ketepatan klasifikasi berat badan bayi lahir pada pohon klasifikasi optimal dan maksimal, yang tepat diklasifikasikan 71.5 % dan 71,9 % tidak terjadi perbedaan yang berarti

Kesimpulan dari penelitian ini adalah bahwa pengolahan menggunakan regresi logistik dan pohon klasifikasi hasilnya dapat dibandingkan dengan melihat ketepatan klasifikasi keduanya. Dari regresi logistik ketepatannya adalah 80,7% sedangkan dari pohon klasifikasi optimal hasil ketepatannya 71,5 % dan ketepatan pohon klasifikasi maksimal hasil ketepatannya adalah 71,9%. Hal ini menandakan bahwa regresi logistik lebih baik daripada Pohon klasifikasi dalam mengklasifikasikan Berat Badan Lahir Bayi di kabupaten Sumenep.

Model regresi logistik dan pohon klasifikasi keduanya menghasilkan variabel dominan yang hampir sama dalam mempengaruhi berat badan bayi. Untuk hasil pengolahan regresi logistik terdiri dari variabel penambahan berat badan, LILA ibu, pendidikan ibu, kemudian HB ibu. Sedangkan pada pohon klasifikasi optimal variabel dominanya adalah penambahan berat badan Ibu dan pendidikan ibu. Pohon klasifikasi maksimal menghasilkan variabel dominan penambahan berat badan , LILA ibu, HB ibu, jumlah anak dan tinggi badan ibu. Berdasarkan uraian diatas maka disarankan, jika ingin menduga kejadian berat badan bayi dengan menggunakan probabilitas dan ingin mengetahui faktor dominan serta ketepatan klasifikasi pada data dengan variabel independen skala data nominal/kontinue dan variabel dependen skala data nominal sebaiknya digunakan regresi logistik.

SUMMARY

In the field of health research that studies the influence of several determinants of an event used the regression method. Regression method is an important component in data analysis that shows the relationship between a dependent variable with one or more independent variables. The goal of regression analysis using this method obtained the best model (fit) and simple that can describe the relationship between outcome variables (dependent or response) with a set of independent variables. One type of regression is logistic regression. Logistic regression approach is a mathematical model that can be used to describe the relationship between the dichotomous dependent variable or polytomus with dichotomous independent variables, polytomus, or continuous. The use of logistic regression analysis has several limitations such as the number of predictor variables causing difficulties in the selection of the most affected variables (most important), most interactions among the predictor variables, and generated models had a lot of difficulties in implementation.

It is needed the methods that are not too attached to some assumptions to overcome the limitations mentioned above. Classification And Regression Tree (CART) is an innovative method for the analysis of large data, through a binary selection procedure. Selection procedure also known as regression trees if the response variable in the form of numerical data, and classification trees if the response variable in the form of data categories.

This research is a statistical study comparing dominant variable in the risk factor of low birth weight between the results of logistic regression analysis and *classification trees*. This research was applied on the secondary data of Birth Weight of Health Department of Sumenep City and cohort reports of pregnant women at five health centers in the District Sumenep. The dependent variable is the Low Birth Weight and the independent variable was maternal age, maternal education, maternal employment status, number of children, birth spacing, maternal hemoglobin, mothers LILA size, increase of maternal weight, and maternal height. The amount of this data is 337.

Based on the results of logistic regression analysis, independent variables that affect the incidence of low birth weight are maternal education, maternal hemoglobin, LILA mother and increase of body weight. Characteristics of weight infants at clinic centers in district Sumenep can be grouped as follows:

1. Most pregnant women are less educated or up to as many as 232 elementary school level, while 119 or 51.2% of mothers who had babies born weighing less than 2500 grams had less educated or up to primary school level.
2. Some mothers who had babies with weight <2500 grams were mothers with hemoglobin less than 11 g % were 66 infants, or 74.15%, while for infants with body weight more than or equal to 2500 g, had maternal hemoglobin ≥11 g with percentage of 75%.
3. Some mothers who had babies with weight ≥ 2500 grams had LILA size ≥ 23.5 cm of 98 women, whereas weight les than 2500 grams had LILA size less than 23,5 cm was 70 women or 71.4%.
4. Most of the mothers of 208 women experienced an increase of weight gain less than 9 kg. Babies weighing less than 2500 grams of a portion of mothers who experienced an increase of less than 9 kg weight of 120 persons or 57.6%.

Based on the output of logistic regression to predict the baby's weight is

$$P(y) = \frac{\exp\ (-4{,}182 + 1.463\,\text{education} + 1.221\,\text{HB} + 1.847\text{LILA} + 2.147\ \text{weigth addition})}{1 + \exp(-4{,}182 + 1.463\,\text{education} + 1.221\,\text{HB} + 1.847\,\text{LILA} + 2.147\ \text{weigth addition})}$$

It means that the educational value of 0 when < Elementary School and value of 1 when ≥ Elementary School, Weight Addition of 0 if < 9 kg, and of 1 if ≥ 9kg. LILA value 0 if < 23.5 cm, and of ≥ 23.5 cm. HB worth 1 if 1 if ≥ 11 g % and of 0 if HB <11 g %. Classification Accuracy Results of logistic regression is 80.5%

The advantages of classification trees method is that it can find out directly on the dominant variable that is located on the main filter. Weaknesses of classification trees can not determine which of these variables are very influential in the data model, the effect of each independent variable is unknown (p does not exist).

In the data processing with computer assistance in getting the results of classification tree nodes are formed 16, 17 terminal nodes and 10 tree depth classification.

Interpretation of the classification tree obtained the dominant variable on Maximum tree in turn is the addition of Weight Loss with a score of 100, Education with a value of 76.22, *Mother Upper Arm Circumference* (LILA) of 33.45, maternal Hemoglobin with a score of 29.28, number of children with score of 7.25 and 5.84 for maternal height.

Dominant variable on the optimum tree is the addition of Weight Loss with a score of 100 and maternal education with score of 63.49.

No differences of dominant variables on the maximum tree and the optimum tree. On the optimum tree pruning has been done on the basis of relative cost values to determine the value of the complexity of the data.

Results grouping of classification accuracy of infants weight at the optimum and maximum classification tree, the right classification is 71.5% and 71.9%, and there are no significant differences. It represents the advantages of the classification tree method that can find out directly on the dominant variable located on the main filter. Weaknesses of classification tree can not determine which variables from the data that is very influential in the model. The effect of each independent variable is unknown (p does not exist).

The conclusion of this study is that results treatment using logistic regression and classification tree can be compared by looking at the accuracy of both classifications. Logistic regression accuracy is 80.7% whereas the optimum classification tree is 71.5% and the accuracy of maximum classification tree is 71.9%. It indicates that logistic regression is better than the classification tree in classifying Birth Weight Infants in district Sumenep.

Both logistic regression model and classification tree produce a similar dominant variable in affecting of the infants' weight. The results of processing logistic regression consisted of the addition variable of weight, LILA mother, maternal education, and maternal HB, whereas on the optimum classification tree, the dominant variables are the maternal Weight Addition and maternal education. Maximum classification tree produces dominant variable of Weight Additions, LILA mother, maternal HB, number of children and maternal height.

Based on the description above it is recommended that if we want to guess baby's weight directly, we can use classification tree, whereas if we want to guess the probability of baby's weight we can use logistic regression.