Klasifikasi Naïve Bayes untuk Prediksi Kelahiran pada Data Ibu Hamil

Naïve Bayes Classification for Due Date Prediction of Pregnancy Data

Aris Nugroho dan Subanar
Jurusan Matematika FMIPA UGM

Abstrak

Dalam bidang kesehatan terkhusus Kesehatan Ibu dan Anak, memprediksi suatu kejadian resiko tinggi (resti) pada kehamilan ibu sehingga kemunculan resiko secara dini bisa ditanggulangi akan sangat mempengaruhi penurunan Angka Kematian Ibu (AKI) maupun Angka Kematian Bayi (AKB). Dengan Model pendekatan Bayesian berupa Klasifikasi Naïve Bayes dengan HMAP (Hipotesis Maksimum A Posteriori) dipakai memprediksi kelahiran yang akan dialami ibu hamil dengan karakteristik Usia ibu, Tinggi Badan, Jumlah Hb, Tekanan Darah, Riwayat Kehamilan lalu dan Penyakit bawaan. Semua data didiskritkan berdasar batasan yang dipakai Departemen Kesehatan dan hasil prediksi berupa probabilitas terjadinya resiko, bisa dipakai sebagai rujukan tempat melahirkan ataupun penilaian kinerja dari penyelenggara jasa persalinan. Dengan fungsi klasifNB dalam bahasa R, fase Training untuk estimasi maksimum likelihood dan sesuai dengan karakteristik ibu hamil, aplikasi menjadi dinamis melakukan prediksi sesuai wilayah dipilih.

Kata kunci: Naive Bayes, Maksimum A Posteriori, Prediksi, Klasifikasi, resiko, Angka Kematian Ibu (AKI), Angka Kematian Bayi (AKB).

Abstract

In the health sector particularly in view of Maternal and Child Health, predicts a high risk event (resti) the emergence of risk pregnancies that can be addressed at an early stage will greatly affect the decline in Maternal Mortality Rate (MMR) and Infant Mortality Rate (IMR). With models such as Bayesian classification approach with Naïve Bayes HMAP (Maximum A Posteriori hypothesis) which will be used to predict birth experienced by pregnant women with maternal age characteristics, Height, Total hemoglobin, blood pressure, and pregnancy history and congenital disease. All data used discritization based limits and the Department of Health in the form of results predicted probability of the risk, can be used as a reference to place of birth or the performance appraisal of delivery service providers. With klasifNB function in R through training phase for maximum likelihood estimation and according characteris-tics of pregnant women , a dynamic application to predict corresponding selected area.

Key words: Naïve Bayes, Maximum A Posteriori, predicts, classification, risk, Maternal Mortality Rate (MMR), Infant Mortality Rate (IMR).

1. Pendahuluan

Salah satu strategi untuk mempercepat penurunan AKI dan AKB adalah program "Making Pregnancy Safer (MPS)", yang ditingkat kecamatan dan kabupaten/kota ditindaklanjuti dengan Pedoman Manajemen Pelayanan Obstetri Emergensi Komprehensif (PONEK) 24 jam dengan langkah utama: peningkatan deteksi dini, pengelolaan ibu hamil resiko tinggi (resti) dan pemantapan kemampuan pengelolaan program di tingkat kota/kabupaten dalam perencanaan, penatalaksanaan, pemantauan juga penilaian kinerja upaya penurunan AKI dan AKB.

Suatu metode analisis yang bisa memfasilitasi adalah klasifikasi Naïve Bayes dengan memakai informasi awal yang dinyatakan sebagai distribusi prior dari masing-masing kelas kelahiran (resiko dan normal). Dari informasi sampel yang dinyatakan dengan fungsi likelihood karakteristik ibu didapat distribusi posterior yang berdasar HMAP (*Hypothesis Maximum A Posteriori*) dipakai untuk mencari probabilitas prediksi kelahiran yang akan dialami ibu hamil.

2. Landasan Teori

Teori yang mendasari adalah distribusi prior, bersama, MLE, pohon keputusan hingga posterior.

Distribusi Beta

Variabel random kontinu X dikatakan mempunyai distribusi Beta (α,β), bila Fungsi kepadatan probabilitas nya berbentuk:

$$f\left(x\mid\alpha,\beta\right)=\frac{1}{B\left(\alpha,\beta\right)}x^{\alpha-1}\left(1-x\right)^{\beta-1},x>0,\alpha>0,\beta>0$$

dimana $B\left(\alpha,\beta\right)=\frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}$

Distribusi Bersama

Misal vektor random $\underset{\sim}{X}=(X_1,X_2)^t$ dengan t menyatakan transpose vektor $\left(X_1,X_2\right)$, andaikan D menyatakan ruang jelajah yang bersesuaian dengan vektor random $\left(X_1,X_2\right)$ dan A menyatakan himpunan bagian dari D maka dapat menyatakan dengan tunggal Px_1,x_2 sebagai fungsi distribusi kumulatif (f.d.k);
$F_{x_1,x_2}(x_1,x_2)=P\left(\left\{X_1\le x_1\right\}\cap\left\{X_2\le x_2\right\}\right),\vee(x_1,x_2)\in R^2$

Distribusi bersama diskrit

Vektor random $\left(X_1,X_2\right)$ disebut vektor random diskret bila ruang penyokong $\left(X_1,X_2\right)$ berhingga atau terhitung. Akibatnya X_1 dan X_2 keduanya diskrit. Fungsi massa probabilitas (f.m.p.) gabungan dari $\left(X_1,X_2\right)$ didefinisikan sebagai $p_{x_1,x_2}(x_1,x_2)=P\left(\left\{X_1=x_1\right\}\cap\left\{X_2=x_2\right\}\right)$ untuk setiap $(x_1,x_2)\in D$.

Estimasi Maksimum Likelihood

Definisi 2.1 *Untuk setiap titik sampel $\underset{\sim}{x}$, misalkan $\hat{\theta}(\underset{\sim}{x})$ adalah harga parameter dimana $L\left(\underset{\sim}{\theta}\mid\underset{\sim}{x}\right)$ adalah fungsi $\underset{\sim}{\theta}$, dengan menganggap $\underset{\sim}{x}$ konstan mencapai maksimumnya, Estimator Maksimum Likelihood (MLE) dari parameter $\underset{\sim}{\theta}$ berdasar sampel $\underset{\sim}{X}$ adalah $\hat{\theta}(\underset{\sim}{x})$*

Peluang Bersyarat

Definisi 2.2. Probabilitas bersyarat

Jika kejadian B terjadi maka agar A terjadi, kejadian sebenarnya adalah titik-titik dalam A dan B yaitu harus berada dalam A∩B sehingga karena B telah terjadi maka B merupakan ruang sampel baru. Yang akibatnya probabilitas A∩B terjadi sama dengan probabilitas A∩B relative terhadap B, ditulis dengan:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}; P(B) > 0$$

Teorema Bayes

Misalkan Ω adalah ruang sampel dari suatu eksperimen dan A_1, A_2, … , A_n adalah peristiwa-peristiwa yang mungkin terjadi di Ω sehingga A_1, A_2, … , A_n saling asing dan $\sum_{i=1}^{n} A_i = \Omega$. Dikatakan bahwa A_1, A_2, … , A_n membentuk partisi dalam Ω. Jika n buah peristiwa A_1, A_2, … , A_n membentuk partisi dalam Ω maka peristiwa-peristiwa $A_1 \cap$B, $A_2 \cap$B, … , $A_n \cap$B membentuk partisi dalam B.

$$B = (A_1 \cap B) \cup (A_1 \cap B) \cup ... \cup (A_n \cap B) == \bigcup_{i=1}^{n} A_i \cap B$$

Karena peristiwa ruas kanan adalah saling asing maka

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) \qquad (1)$$

Jika $P(A_1)>0$ untuk i=1,2,…,n maka $P(A_1 \cap B)=P(A_1).P(B|A_1)$ sehingga didapat

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i) P(A_i \mid B) \qquad (2)$$

Teorema 2.1 Teorema Bayes

Misalkan peristiwa A_1, A_2, … , A_n membentuk partisi dalam Ω sedemikian hingga $P(A_i)>0$ untuk i=1, 2, … , n dan misalkan B sebarang peristiwa sedemikian hingga P(B)>0 maka untuk i=1, 2, … , n

$$P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j) P(B \mid A_j)}$$

Bukti:

Dari definisi 2.2 didapat $P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$, dan dari pers. (2) maka diperoleh

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{\sum_{j=1}^{n} P(A_j) P(B \mid A_j)} == \frac{P(A_i) P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j) P(B \mid A_j)}$$

Klasifikasi Naïve Bayes

Klasifikasi Naïve Bayes adalah klasifikasi berdasar teorema Bayes dan digunakan untuk menghitung probabilitas tiap klas dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung (independen). Jika y adalah suatu klas dan X_1, X_2,…,X_p adalah nilai observasi sejumlah p prediktor, maka nilai probabilitas masing-masing klasnya adalah:

$$P(Y = y \mid X_1, X_2, ..., X_p) = \frac{P(y) P(X_1, X_2, ..., X_p \mid y)}{P(X_1, X_2, ..., X_p)}$$

Dengan asumsi Independensi antara X_1, X_2,…,X_p didapat $P(X_1, X_2, ..., X_p \mid y) = P(X_1 \mid y) P(X_2 \mid y) P(X_3 \mid y) ... P(X_p \mid y)$ sehingga probabilitas masing-masing klas dalam Klasifikasi Naïve Bayes adalah

$$P(y \mid X_1, X_2, ..., X_p) = \frac{P(y)P(X_1 \mid y)P(X_2 \mid y)...P(X_p \mid y)}{P(X_1, X_2, ..., X_p)} = \frac{P(y)\prod_{i=1}^{p} P(X_i \mid y)}{P(X_1, X_2, ..., X_p)}$$

Dalam Naïve Bayes dinyatakan HMAP (*Hypothesis Maximum A Postreior*) dimana $H_{MAP} = \arg\max P(y \mid X_1, ..., X_p)$ adalah memaksimalkan nilai probabilitas masing-masing klas. Dengan teorema Bayes dan asumsi kondisi independen maka

$$H_{MAP} = \arg\max P(y \mid X_1, ..., X_p) = \arg\max \frac{P(y)\prod_{i=1}^{p} P(X_i \mid y)}{P(X_1, X_2, ..., X_p)} \propto \arg\max P(y)\prod_{i=1}^{p} P(X_i \mid y)$$

Tahapan Klasifikasi

Terdapat 2 proses dalam klasifikasi ini yaitu:

1. Tahap Training: dengan memakai data yang ada, membangun metode untuk mengestimasi parameter dari distribusi peluangnya dengan asumsi bahwa adanya independensi dari masing-masing kelas (data dengan karakteristik yang sama). Dalam tahapan ini dilakukan estimasi pada parameter $\boldsymbol{\theta}$ dengan Maximum Likelihood (ML).
2. Tahap Prediksi: Proses menggunakan model yang sudah dibangun tersebut untuk melakukan tes data untuk memperkirakan/mengukur akurasi dari aturan yang dibentuk dalam model dengan menghitung peluang posterior kemudian mengklasifikasi kedalam peluang posterior terbesar HMAP (*Hypothesis Maximum A Postreior*).

Kelompok Faktor Resiko Kehamilan

Berdasar kapan ditemukan (pada kehamilan muda atau kehamilan lanjut), cara pengenalan dan tingkat resikonya, ibu hamil dikelompokkan menjadi 3 kelompok FR(faktor resiko) I, II, III sebagai berikut:

- Kelompok FR I: Ada Potensi Gawat Obstetri (APGO) dengan 7 Terlalu dan 3 Pernah. Tujuh Terlalu adalah primi muda, primi tua, primi tua sekunder, umur ≥ 35 tahun, grande multi, anak terkecil < 2 tahun, tinggi badan rendah ≤ 145 cm; sedangkan Tiga Pernah adalah riwayat obstetric jelek, persalinan lalu mengalami pendarahan dan bekas operasi sesar.
- Kelompok FR II: Ada Gawat Obstetri (AGO) berupa penyakit ibu, preeklampsi ringan, hamil kembar, hidramnion, hamil serotinus, IUFD, letak sungsang dan letak lintang.
- Kelompok FR III: Ada Gawat Darurat Obstetri (AGDO) berupa pendarahan antepartum dan preeklampsi berat / eklampsia.

3. Klasifikasi Naïve Bayes

3.1 Model Naïve Bayes

Definisi 3.1 (**Model Naïve Bayes (NB)**). Model NB terdiri dari sejumlah k bilangan bulat kelas label yang mungkin dan sejumlah d bilangan bulat atribut selain dari parameter $P(y), y \in \{1, 2, ..., k\}$ sebagai peluang terjadinya kelas label y dimana $P(y) \geq 0, \sum_{y=1}^{k} P(y) = 1$ dan parameter $P_j(x \mid y), j \in \{1, 2, ..., d\}, x \in \{-1, +1\}, y \in \{1, 2, ..., k\}$ sebagai peluang terjadinya atribut j mendapatkan x dengan kondisi bersyarat pada suatu kelas label y tertentu dimana $P(x \mid y) \geq 0, \forall y, j, \sum_{x \in \{-1,+1\}} P(x \mid y) = 1$. Sehingga peluang untuk setiap $y, x_1, x_2, ..., x_d$ dapat dituliskan sebagai berikut: $p(y, x_1, x_2, ..., x_d) = P(y)\prod_{j=1}^{d} P_j(x_j \mid y)$

Variabel dalam studi kasus ini terdiri atas variabel diskrit yaitu resiko melahirkan yang terdiri dari k=2 kelas (katagorik) dan variabel independen yaitu Usia Ibu, Tinggi Badan, Tekanan Darah, Jumlah Hb, Riwayat Persalinan yang lalu dan Penyakit bawaan Ibu. Dengan membandingkan nilai $P(Y \mid \underset{\sim}{X}) = P(Y)\prod_{i=1}^{d} P(X_i \mid Y)$ masing-masing kelas (Y=0 untuk kelahiran Normal dan Y=1 untuk kelahiran Resiko) pada karakteristik ibu akan didapat:

Prediksi Kelahiran Resiko

Jika $P(Y = resti)\prod_{i=1}^{p} P(X_i \mid Y = resti) > P(Y = normal)\prod_{i=1}^{d} P(X_i \mid Y = normal)$

Prediksi Kelahiran Normal

Jika $P(Y = resti)\prod_{i=1}^{d} P(X_i \mid Y = resti) < P(Y = normal)\prod_{i=1}^{d} P(X_i \mid Y = normal)$

Estimasi Maksimum Likelihood Model Naïve Bayes

Diketahui data $\left(\underset{\sim}{x}^{(i)}, y^{(i)}\right), i = 1,2,...,n$ dan $x_j^{(i)}$ adalah karakteristik ke j ibu hamil dengan $x_j^{(i)} \in \{0,1\}$

Definisi 3.2 (**Estimas ML (MLE) untuk Model Naïve Bayes**). *Asumsikan suatu data training* $\left(\underset{\sim}{x}^{(i)}, y^{(i)}\right), i = 1,2,...,n$. *MLE adalah mengestimasi nilai parameter dalam Model NB yaitu parameter* $P(y), y \in \{1,2,...,k\}$ *dan* $P_j(x \mid y), j \in \{1,2,...,d\}, x \in \{0,+1\}, y \in \{1,2,...,k\}$ *yang dapat memaksimalkan* $L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \log P\left(y^{(i)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d} \log P_j\left(x_j^{(i)} \mid y^{(i)}\right)$ *dengan batasan (sifat):*

$P(y) \geq 0, \forall y \in \{1,2,...,k\}$ *dan* $\sum_{y=1}^{k} P(y) = 1$

$\forall y, j, x, P_j(x \mid y) \geq 0; \forall y \in \{1,2,...,k\}, \forall j \in \{1,2,...,d\}, \sum_{x \in \{0,+1\}} P_j(x \mid y) = 1$

Teorema 3.1 (**Estimas ML (MLE) untuk Model Naïve Bayes**). *Memakai definisi 3.2 maka MLE untuk Model NB dapat ditulis dengan* $P(y) = \frac{\sum_{i=1}^{n}\left[y^{(i)} = y\right]}{n} = \frac{count(y)}{n}$ *dan* $P(x \mid y) = \frac{\sum_{i=1}^{n}\left[\left(y^{(i)} = y\right) \wedge \left(x_j^{(i)} = x\right)\right]}{\sum_{i=1}^{n}\left[y^{(i)} = y\right]} = \frac{count_j(x \mid y)}{count(y)}$.

Bukti

Dari data yang ada, fungsi Log-Likelihood $L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \log P\left(y^{(i)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d} \log P_j\left(x_j^{(i)} \mid y^{(i)}\right)$ akan dihitung kembali sampai dengan jumlah kali parameter $P(y), y \in \{1,2,...,k\}$ dan $P_j(x \mid y), j \in \{1,2,...,d\}, x \in \{0,+1\}, y \in \{1,2,...,k\}$ muncul dalam fungsi Log-likelihood diatas sehingga didapat $L(\underset{\sim}{\theta}) = \sum_{y=1}^{k} count(y) \log P(y) + \sum_{j=1}^{d}\sum_{y \in Y}\sum_{x \in \{-1,+1\}} count_j(x \mid y) \log q_j(x \mid y)$ yang dimaksimalkan dengan memaksimalkan $\sum_{y=1}^{k} count(y) \log P(y)$ atau memaksimalkan

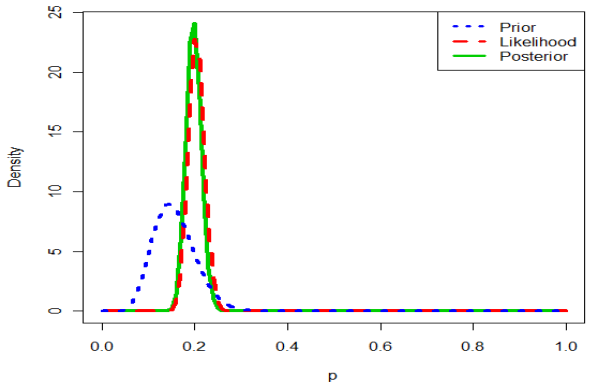
$\sum_{j=1}^{d}\sum_{y\in Y}\sum_{x\in\{0,+1\}} count_j(x\mid y)\log P_j(x\mid y)$. MLE dari $P(Y=y), y\in\{0,1\}$ adalah $P(Y=y)=\frac{\sum_{i=1}^{n}(y^{(i)}=y)}{n}=\frac{count(y)}{n}$ sedang estimasi $P(X_i=x\mid Y=y), \forall x,y\in\{0,1\}$ adalah

$$P(X_i=x\mid Y=y)=\frac{\sum_{i=1}^{n}(y^{(i)}=y\_dan\_x_j^{(i)}=x)}{\sum_{i=1}^{n}(y^{(i)}=y)}=\frac{count_j(x\mid y)}{count(y)}$$

Model Satu Parameter

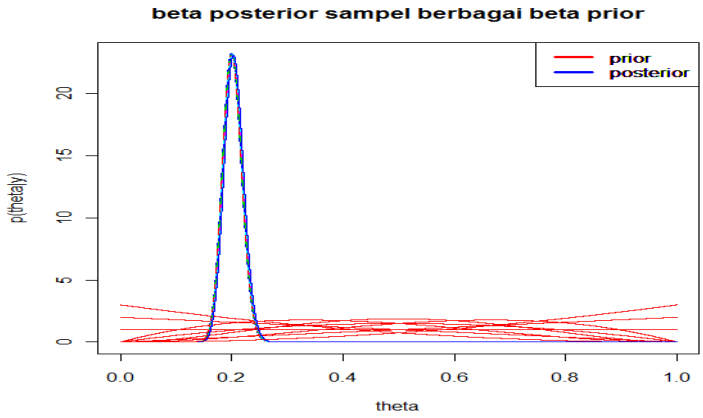
Misal $X_1, X_2, X_3, ..., X_n$ *conditionally independent and identically distributed*(i.i.d) Bernoulli(θ) dan diasumsikan distribusi prior θ adalah $\pi(\theta)\sim Beta(\alpha,\beta)$. Karena $Y=\sum_{i=1}^{n}x_i\sim binomial(n,\theta)$ maka distribusi bersama (joint distribution) dari y dan θ adalah $f(y,\theta)=f(y\mid\theta)\pi(\theta)$ dan distribusi posteriornya $\pi(\theta\mid y)\sim Beta(y+\alpha, n-y+\beta)$.

Misalkan $\alpha=10, \beta=54, n=538, y=109$ maka akan didapat grafik seperti dibawah.



Gambar 1. prior. Likelihood, posterior

Meskipun distribusi beta prior yang dipilih memiliki nilai a dan b yang berbeda-beda namun didapat bentuk grafik dari beta posteriornya tidak terlalu berbeda seperti gambar dibawah:



Gambar 2. Keajegan posterior

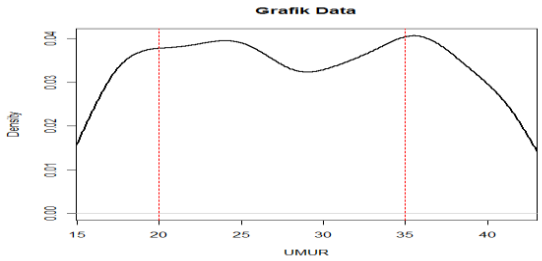
3.2 Diskritisasi

Diskritisasi adalah suatu cara untuk melakukan pengelompokan dari suatu nilai berdistirbusi kontinu kedalam suatu interval sehingga menjadi berdistibusi diskrit. Misal

untuk variabel umur yang berdistirbusi kontinu maka dengan memakai acuan bahwa umur yang tidak beresiko untuk melahirkan adalah antara 20 s.d 35 tahun maka proses diskritisasinya adalah mengubah sesuai fungsi berikut:

$$C_{UMUR} = \begin{cases} c_1\left(\min \le X_i < batas_1 = 20\right) \\ c_2\left(batas_1 = 20 \le X_i \le batas_2 = 35\right) \\ c_3\left(batas_2 = 35 < X_i \le \max\right) \end{cases}$$

Sehingga didapat hasil seperti gambar berikut:



Gambar 3. Data Umur dengan batas atas dan batas bawah

3.3 Joint Diskrit Distribution, Marginal Distribusi dan Konditional distribusi

Pandang suatu variable random diskrit (X_1, X_2) dimana keduanya berdistribusi Binomial sehingga vector random (X_1, X_2) disebut vektor tandom diskrit. Fungsi massa probabilitas (fmp) gabungan dari (X_1, X_2) didefinisikan sebagai $p_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$. Dari fmp gabungan tersebut didapat fmp marginal $X_i, i = 1, 2$ adalah:

$$p_{X_1}(x_1) = P(X_1 = x_1) = \sum_{x_2 \in X_2} P(X_1 = x_1, X_2 = x_2) = \sum_{x_2 \in X_2} p_{X_1,X_2}(x_1, x_2)$$

$$p_{X_2}(x_2) = P(X_2 = x_2) = \sum_{x_1 \in X_1} P(X_1 = x_1, X_2 = x_2) = \sum_{x_1 \in X_1} p_{X_1,X_2}(x_1, x_2)$$

Dari joint dan marginal didapat conditional density adalah $p_{X_2|X_1}(x_2 \mid x_1) = \dfrac{p_{X_1,X_2}(x_1, x_2)}{P(X_1 = x_1)}$

Dipilih variable kelahiran dan tinggi badan sesuai tabel berikut:

Tabel 1. Crosstabulasi data Tinggi Badan berdasar Kelahiran

kelahiran	**tinggi badan**		**jumlah**
	1=resiko	0=normal	
1=resiko	10	100	110
0=normal	45	383	428
jumlah	55	483	538

Tabel 2. fmp gabungan dan fmp marginal data Tinggi Badan dan Kelahiran

kelahiran	**tinggi badan**		**fmp marginal kelahiran**
	1=resiko	0=normal	
1=resiko	0.018587	0.185874	0.204460967
0=normal	0.083643	0.711896	0.795539033
fmp marginal tb	0.10223	0.89777	1

Tabel 3. P(tinggi badan|kelahiran)

kelahiran	**tinggi badan**		**condtional**
	1=resiko	0=normal	
1=resiko	0.090909	0.909091	1
0=normal	0.10514	0.89486	1

3.4 Pohon Keputusan

Dalam membangun pohon keputusan dipergunakan informasi Gain dengan rumus $Gain(X_i) = Info(D) - Info_{X_i}(D)$ dimana $Info(D) = -\sum_{i=1}^{2} \frac{n_{(X_i = x_i | Y = y_i)}}{n_{(X_i)}} \log\left(\frac{n_{(X_i = x_i | Y = y_i)}}{n_{(X_i)}}\right)$ adalah prob. prior terbesar dalam node dan $Info_{X_i}(D) = -\sum_{i=1}^{2} \frac{|D_i|}{|D|} Info(D_i)$ adalah gain X_i node.

Algoritma pembuatan pohon keputusan adalah sebagai berikut:

(1). Membuat Node 1 dengan probabilitas prior terbesar.
(2). Mencari nilai maksimal Gain masing-masing kovariate dalam Node-nya.
(3). Membuat koneksi kelas dari kovariate dengan nilai gain terbesar.
(4). Memberi label (nama) koneksi node sesuai dengan kelasnya.
(5). Membuat node baru dan memberi label sesuai kelas probabilitas prior
(6). Kembali ke-langkah (2) dengan tidak mengikutsertakan kovariate yang sudah terpilih.
(7). Dilakukan perulangan sampai keadaan berhenti tercapai.
(8). Berhenti bila memenuhi salah satu dari keadaan:
   a) Salah satu prob. prior = 0 (nol)
   b) Tidak ada nilai Gain terbesar.
   c) Jika node di bawahnya memiliki label (nama) yang sudah sama.
   d) Semua kovariate sudah terpilih.
(9). Untuk kasus dikotomus (dalam tesis ini) maka tiap node selalu memiliki 2 koneksi node dengan node dibawahnya (normal dan resiko) dan label koneksi hanya resiko dan tidak.

Fase Training

Pada fase training, berdasar data yang ada akan didapat nilai estimasi dari $P(Y = y), y \in \{0,1\}$ sebagai prior dari masing-masing kelasnya dan nilai estimasi likelihood $P(X_j \mid Y = y), y \in \{0,1\}, j = 1,2,...,6$. Untuk kasus 538 data kehamilan maka nilai Maximum Likelihood Estimates (MLE) untuk sejumlah 6 vector kovariate adalah:

$$P(X_j \mid Y = 0) = \frac{\sum_{i=1}^{538} 1\{X_j = x_{ij} \,\&\, Y_i = 0\}}{\sum_{i=1}^{538} 1\{Y_i = 0\}}, j = 1,2,...,6 \quad P(X_j \mid Y = 1) = \frac{\sum_{i=1}^{538} 1\{X_j = x_{ij} \,\&\, Y_i = 1\}}{\sum_{i=1}^{538} 1\{Y_i = 1\}}, j = 1,2,...,6$$

Distribusi priornya adalah:

$$P(Y = 0) = \frac{\sum_{i=1}^{538} 1\{Y_i = 0\}}{538}, \text{ dan } P(Y = 1) = \frac{\sum_{i=1}^{538} 1\{Y_i = 1\}}{538}$$

Jika terdapat kelas dengan nilai peluangnya adalah 0 (nol) maka dipakai koreksi Laplace, nilai MLE untuk $P(\underset{\sim}{X} \mid Y = 0), P(\underset{\sim}{X} \mid Y = 1)$ menjadi:

$$P(X_j \mid Y = 0) = \frac{\sum_{i=1}^{538} 1\{X_j = x_{ij} \,\&\, Y_i = 0\} + 1}{\sum_{i=1}^{538}\sum_{i=1}^{538} 1\{Y_i = 0\} + 2}, j = 1,2,...,6 \quad P(X_j \mid Y = 1) = \frac{\sum_{i=1}^{538} 1\{X_j = x_{ij} \,\&\, Y_i = 1\} + 1}{\sum_{i=1}^{538} 1\{Y_i = 1\} + 2}, j = 1,2,...,6$$

Fase Prediksi

Dalam fase ini, dari seluruh model yang ada dilakukan penghitungan peluang posterior terbesar dengan membandingkan nilai probabilitas posterior kedua kelas yang ada untuk

suatu karakteristik yang ada, misalkan hasil prediksi menyatakan bahwa klas yang didapat adalah Y=1, maka hanya bisa dihasilkan apabila $P(Y=1|\underset{\sim}{X}) > P(Y=0|\underset{\sim}{X})$ berarti:

$$P(Y=1)\prod_{i=1}^{p} P(X_i|Y=1) > P(Y=0)\prod_{i=1}^{p} P(X_i|Y=0)$$

4. Aplikasi Studi Kasus

Data Percobaan

Contoh studi kasus menggunakan data yang memuat karakteristik ibu hamil seperti Usia ibu, Tinggi Badan, Tekanan Darah, Jumlah Hb, riwayat persalinan dan Penyakit bawaan ibu. Data dari 3 wilayah yang diolah secara per-wilayah dan juga secara keseluruhan.

Konstruksi Variabel

Variabel- variabel dalam penelitian:

(1) Persalinan Ibu, dengan kelas pada variabel persalinan adalah dikotomus:
Persalinan berresiko (kode 1)
Persalinan Normal (kode 0)

(2) Umur ibu dengan indikator:
Umur berresiko (kode 1) jika umur < 20 tahun atau > 35 tahun
Umur tidak berresiko (kode 0) jika 20 tahun ≤ umur ibu ≤ 35 tahun.

(3) Tinggi badan ibu dengan indikator:
Tinggi badan berresiko (kode 1) jika TB ≤145 cm
Tinggi badan normal (kode 1) jika TB >145 cm.

(4) Tekanan darah dengan indikator:
Tekanan darah berresiko (kode 1) jika diatas 125 mmHg.
Tekanan darah normal bila dibawah 125 mmHg.

(5) Jumlah Hb dengan indikator:
Berresiko (kode 1) jika dibawah 11.
Normal bila ≥ 11.

(6) Riwayat persalinan ibu hamil, dengan indikator:
Riwayat persalinan berresiko (kode 1) bila:
Kehamilan kedua tetapi kehamilan pertama mengalami keguguran, lahir belum cukup bulan, lahir mati dan lahir hidup namun mati ketika berumur ≤ 7 hari,
Kehamilan kedua atau lebih dimana kehamilan terakhir janin mati dalam kandungan.
Kehamilan ketiga atau lebih tetapi kehamilan lalu mengalami keguguran minimal 2 kali.
Riwayat persalinan baik atau normal (kode 0)

(7) Penyakit bawaan (yang diderita ibu) pada saat melahirkan adalah
Penyakit bawaan berresiko (kode 1) bila ibu menderita Malaria, TBC, Payah jantung, Kencing Manis, HIV/AIDS, Toksoplasmosis, dll.
Tidak ada penyakit bawaan (kode 0)

Fungsi utama dalam Naïve Bayes

```
Fungsi Training, Algoritma1: algoritma untuk menghitung probabilitas bersyarat
Input: Function() dengan memanggil fungsi diskritisasi
Output: Matrik probabilitas bersyarat kelas respon
Deklarasi jumlah tiap kolom data
Panggil fungsi Baca Data Baru
Class← jumlah kolom respon
```

```
Hitung baris dan kolom Data
Prior←class/jml baris
For i←2 sampai jumlah kolom Data
For j←1 sampai jumlah baris Data
Hitung probabilitas bersyarat masing-masing kovariat terhadap kelas respon
Next j, Next i
Menampilkan hasil dalam bentuk matrik

Fungsi Prediksi, Algoritma2: algoritma untuk menghitung peluang terbesar
Input: Function() dengan masukan karakteristik ibu
Output: berupa pohon keputusan dan peluang yang terjadi
Input wilayah(domisili) dan karakteristik ibu yang akan diprediksi kelahirannya
Menampilkan pohon keputusan berdasar wilayah(domisili)
Menghitung posterior masing-masing kelas dan mencari maksimalnya
Menampilkan hasil prediksi berdasar maksimal posteriornya.
```

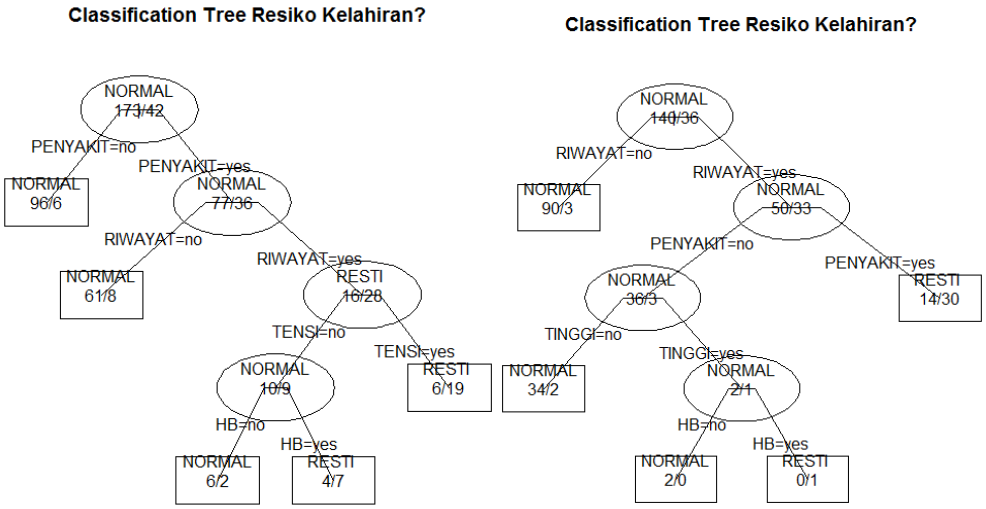
(a) Estimasi Parameter

Nilai estimasi hasil fase training untuk parameter $P(Y=y), y \in \{0,1\}$ dan parameter $P(X_i \mid Y)$ untuk $P(X_i \mid Y=y), y \in \{0,1\}$ dari wilayah 2 da 3 adalah sebagai berikut:

```
PRIOR PROBABILITAS
RESIKO=ya    sebesar  0.1953488

RESIKO=tidak sebesar  0.8046512


CONDITION PROBABILITY  UMUR
                          yes       no
resti_yes 0.1953488 0.4523810 0.5476190
resti_no  0.8046512 0.4104046 0.5895954

CONDITION PROBABILITY  TINGGI
                           yes        no
resti_yes 0.1953488 0.02380952 0.9761905
resti_no  0.8046512 0.13294798 0.8670520

CONDITION PROBABILITY  HB
                          yes       no
resti_yes 0.1953488 0.4285714 0.5714286
resti_no  0.8046512 0.4219653 0.5780347

CONDITION PROBABILITY  TENSI
                          yes       no
resti_yes 0.1953488 0.7619048 0.2380952
resti_no  0.8046512 0.5491329 0.4508671

CONDITION PROBABILITY  RIWAYAT
                          yes       no
resti_yes 0.1953488 0.7619048 0.2380952
resti_no  0.8046512 0.3872832 0.6127168

CONDITION PROBABILITY  PENYAKIT
                          yes       no
resti_yes 0.1953488 0.8571429 0.1428571
resti_no  0.8046512 0.4450867 0.5549133
```

```
PRIOR PROBABILITAS
RESIKO=ya    sebesar  0.2045455

RESIKO=tidak sebesar  0.7954545


CONDITION PROBABILITY  UMUR
                          yes       no
resti_yes 0.2045455 0.3611111 0.6388889
resti_no  0.7954545 0.3714286 0.6285714

CONDITION PROBABILITY  TINGGI
                           yes        no
resti_yes 0.2045455 0.13888889 0.8611111
resti_no  0.7954545 0.09285714 0.9071429

CONDITION PROBABILITY  HB
                          yes       no
resti_yes 0.2045455 0.4444444 0.5555556
resti_no  0.7954545 0.4571429 0.5428571

CONDITION PROBABILITY  TENSI
                          yes       no
resti_yes 0.2045455 0.6666667 0.3333333
resti_no  0.7954545 0.5214286 0.4785714

CONDITION PROBABILITY  RIWAYAT
                          yes        no
resti_yes 0.2045455 0.9166667 0.08333333
resti_no  0.7954545 0.3571429 0.64285714

CONDITION PROBABILITY  PENYAKIT
                          yes       no
resti_yes 0.2045455 0.8333333 0.1666667
resti_no  0.7954545 0.4214286 0.5785714
```

Gambar 1. Hasil fase training wilayah 2 dan 3 fungsi klasifNB()



Gambar 2. Pohon keputusan wilayah 2 dan 3 fase training fungsi klasifNB()

(b) Estimasi MAP (Maximal A Posterior)

Dimisalkan seorang ibu (Ade) yang berumur 32 tahun, memiliki tinggi badan 165 cm, jumlah Hb 12, tekanan darah 135, riwayat persalinan lalu melakukan operasi Cesar dan tidak memiliki penyakit bawaan. Dari karakteristik diatas dengan Klasifikasi Naïve Bayes didapat hasil prediksi dari masing-masing wilayah sebagai diperlihatkan pada Gambar 3 atau Tabel 8 berikut.

```
DATA IBU HAMIL WILAYAH 3 YANG AKAN DIPREDIKSI DENGAN NAIVE BAYES

NAMA IBU                          = Ade
UMUR IBU                          = 32
TINGGI IBU(dlm cm)                = 165
JUMLAH HB                         = 12
TEKANAN DARAH IBU                 = 135
ADA RIWAYAT KEHAMILAN RESIKO(y/t) = y
ADAKAH RIWAYAT PENYAKIT RESIKO(y/t) = t

NILAI PREDIKSI RESIKO    =  0.006367527        NILAI PREDIKSI NORMAL    =  0.0265292$

PELUANG KELAHIRAN RESIKO =  0.1935608   PELUANG KELAHIRAN NORMAL =  0.8064392

*********** KESIMPULAN AKHIR ***********
IBU BERNAMA :  Ade  TIDAK AKAN BERESIKO KETIKA MELAHIRKAN

    MAU LAGI (y/t)?

DATA IBU HAMIL WILAYAH 2 YANG AKAN DIPREDIKSI DENGAN NAIVE BAYES

NAMA IBU                          = Ade
UMUR IBU                          = 32
TINGGI IBU(dlm cm)                = 165
JUMLAH HB                         = 12
TEKANAN DARAH IBU                 = 135
ADA RIWAYAT KEHAMILAN RESIKO(y/t) = y
ADAKAH RIWAYAT PENYAKIT RESIKO(y/t) = t

NILAI PREDIKSI RESIKO    =  0.004948678        NILAI PREDIKSI NORMAL    =  0.02806029

PELUANG KELAHIRAN RESIKO =  0.1499192   PELUANG KELAHIRAN NORMAL =  0.8500808

*********** KESIMPULAN AKHIR ***********
IBU BERNAMA :  Ade  TIDAK AKAN BERESIKO KETIKA MELAHIRKAN

    MAU LAGI (y/t)?
```

Gambar 3. Hasil fase prediksi fungsi klasifNB()

Table 4. Hasil Akhir prediksi kelahiran dari karakteristik Ibu

Kelahiran	Probabilitas	Wilayah
Normal	0.8835409	1
Normal	0.8500808	2
Normal	0.8064392	3
Normal	0.8442455	Kabupaten

5. Kesimpulan dan Saran

Kesimpulan

Dari data dalam kasus ini didapat peluang terbesar ada di wilayah 1 (i) Adanya perbedaan dalam pohon keputusan berkenaan dengan node awal dari masing-masing wilayah sebagai faktor penyebab pertama prediksi kelahiran; (ii) Adanya perbedaan prioritas program menurunkan AKI dan AKB dari masing-masing wilayah mengacu pada node awal

Saran

(i) Memakai metode tertentu pada proses diskritisasi yang dilakukan untuk menentukan batas kelas terutama untuk data dengan skala bukan nominal; (ii) Melakukan kombinasi yang lain dalam memprediksi bukan lagi dengan decision tree dan Naïve Bayes namun bisa dengan kombinasi-kombinasi yang lain.

Sedang saran bagi pembuat keputusan dilingkungan departemen kesehatan adalah: Ketika akan dibuat suatu kebijakan berdasar prediksi dengan Naïve Bayes perlu diperlengkapi dengan pohon keputusan karena dimungkinkan adanya kebijakan yang spesifik di wilayah karena perbedaan hasil dari masing-masing wilayah.

Daftar Pustaka

Albert, J., 2009, *Bayesian Computation with R,* Springer NY.

Bain, Lee J dan Engelhardt, M., 1992, *Introduction to Probability and Mathematical Statistics. California:* Duxbury.

Berger, J.O., 1985, *Statistical Decision Theory and Bayesian Analysis 2nd with 23 illustrations,* Springer NY.

Bolstad, W.M., 2004, *Introduction to Bayesian Statistics,* John Wiley & Sons.

Collin, M, nd., *The Naïve Bayes Model, Maximum-Likelihood Estimation, and the EM Algorithm*.

Han. J and Kamber. M, 2006, *Data Mining 2 nd: Concepts and Techniques,* Morgan Kaufmann Publisers.

Hoff. P.D , 2009, *A First Cource in Bayesian Statistical Methods,* Springer NY.

Poedji Rochjati**,** 2003, *Skrining Antenatal pada Ibu Hamil*. Pusat Safe Motherhood-Lab/SMF ObGin RSU Dr. Soetomo/Fak. Kedokteran UNAIR Surabaya.

Rokach, Lior and Oded Maimon**,** 2008, *Data Mining With Decision Trees: Theory and Application* ,World Scientific Publishing.

Ross, M.Sheldon, 2010, *Introduction to Probability Models, 10th Edition,* Academic Press.

Subanar**,** 2006, *Inferensi Bayesian.* Penerbit Universitas Terbuka.

Subanar**,** 2013,*Statistika Matematika.* Graha Ilmu, 2013.

Subanar**,** 2013, *Statistika Matematika; Probabilitas, Distribusi dan Asimtotis dalam Statistika.* Graha Ilmu.

Turban, E dkk, 2005, *Decision Support Systems and Intelegent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas).* Penerbit ANDI Yogyakarta.

Wahyono, T**.** nd, *Penerapan Metode Naïve Bayes Untuk Pengembangan Sistem Peringatan Dini Serangan Organisme Pengangu Tanaman (OPT) Padi dengan Auto SMS.* Tesis, Yogyakarta: Jurusan FMIPA UGM.