Jurnal Matematika UNAND
Vol. **IX** No. **1** Hal. 23 – 29
Edisi Januari 2020
ISSN : 2303–291X

KLASIFIKASI DAERAH TERTINGGAL DI INDONESIA MENGGUNAKAN METODE *NAIVE BAYEA CLASSIFIER*

WINDA LIDYA, HAZMIRA YOZZA*, FERRA YANUAR

Program Studi S1 Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia.
email : windalidya27@gmail.com, hazmirayozza@sci.unand.ac.id, ferrayanuar@sci.unand.ac.id

Diterima 29 November 2019 Direvisi 3 Desember 2019 Dipublikasikan 12 Januari 2020

Abstrak. Status daerah dapat diprediksi berdasarkan klasifikasi dengan metode *Naive Bayes*. *Naive* Bayes merupakan teknik prediksi berbasis peluang sederhana yang berdasarkan penerapan teorema Bayes dengan tingkat akurasi cukup tinggi. Klasifikasi status daerah ditentukan berdasarkan indikator yang terkait dalam penentuan status daerah. Menghitung klasifikasi *Naive* Bayes untuk indikator kontinu menggunakan sebaran normal. Pengukuran Kinerja Klasifikasi ditentukan dengan menggunakan matriks konfusi, diperoleh nilai akurasi sebesar 0,905 yang artinya nilai akurasi yang diperoleh cukup baik dalam klasifikasi status daerah.

Kata Kunci: Status daerah, klasifikasi, *Naive* Bayes, sebaran normal, matriks konfusi, akurasi.

1. Pendahuluan

Daerah tertinggal adalah suatu daerah kabupaten/kota yang masyarakat dan wilayahnya relatif kurang berkembang dibandingkan daerah lain. Menurut Peraturan Presiden Nomor 131 Tahun 2015 tentang penetapan daerah tertinggal tahun 2015-2019, ditetapkan 122 daerah kabupaten/kota sebagai daerah tertinggal [4]. Status daerah dapat ditentukan dengan menggunakan suatu indeks komposit yang dihitung dengan suatu metode tertentu dari banyak variabel. Sebahagian diantara variabel tersebut tidak dipublikasikan, sehingga menjadi tidak mudah untuk memperkirakan apakah suatu daerah merupakan daerah tertinggal atau bukan daerah tertinggal. Pada penelitian ini, metode Klasifikasi *Naive* Bayes digunakan untuk memprediksi apakah suatu kabupaten/kota dikategorikan sebagai daerah tertinggal atau tidak dengan menggunakan variabel yang telah dipublikasikan. *Naive* Bayes merupakan teknik prediksi berbasis peluang sederhana yang berdasarkan pada penerapan teorema Bayes dengan tingkat akurasi yang lebih baik dibandingkan dengan metode klasifikasi lainnya [3].

*penulis korespondensi

2. Landasan Teori

2.1. *Sebaran Normal*

Definisi 2.1. *[5] Bila X adalah suatu peubah acak normal dengan nilai tengah μ dan ragam σ^2, maka persamaan kurva normalnya adalah*

$$n(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2\left(\frac{x-\mu}{2}\right)^2}, \ \textit{untuk} \ -\infty < x < \infty. \tag{2.1}$$

2.2. *Metode Naive Bayes*

Metode Klasifikasi *Naive* Bayes merupakan gabungan dari pemodelan peluang dengan suatu aturan keputusan [2].

2.2.1. *Model Peluang Naive Bayes*

Misal Y adalah peubah acak yang menyatakan kelas suatu pengamatan dan $\mathbf{X}= (X_1, X_2, \cdots, X_k)$ adalah vektor peubah acak yang menyatakan nilai indikator yang terkait dengan Y. Lebih lanjut nilai y adalah label dari kelas tertentu, dan $\mathbf{x}= (x_1, x_2, \cdots, x_k)$ adalah nilai dari $\mathbf{X}$ [7]. Pada metode ini dengan aturan Bayes akan diformulasikan fungsi peluang untuk Y bersyarat X sebagai berikut

$$P(Y = y|\mathbf{X}=\mathbf{x}) = \frac{\prod_{i=1}^{k} P(X_i = x_i|Y = y)P(Y = y)}{\prod_{i=1}^{k} P(X_i = x_i)}, \tag{2.2}$$

dengan $P(X_i = x_i)$ merupakan nilai yang selalu tetap. Aturan umum yang digunakan adalah mengelompokkan suatu pengamatan ke kategori yang paling mungkin, yang dikenal dengan MAP *(Maximum A Posterior)* yang didefinisikan sebagai:

$$h(MAP) = arg\left(\max P(Y = y)\prod_{i=1}^{n} P(X_i = x_i|Y = y)\right).$$

Artinya suatu pengamatan akan dikategorikan ke dalam kelas $\mathbf{Y}=\mathbf{y}$ jika $P(\mathbf{Y}=\mathbf{y})\Pi P(P(X_i = x_i|\mathbf{Y}=\mathbf{y})$ maksimum untuk semua $y = 1, 2$.

2.2.2. *Penduga Parameter dan Model Kejadian*

Untuk menduga peluang posterior pada Persamaan (2.2), perlu diduga terlebih dahulu peluang prior kejadian $P(Y = y)$, dan $P(X_i = x_i|Y = y)$. Peluang prior kejadian Y diduga dari peluang kelas tersebut yaitu:

$$P(Y = y) = \frac{\text{Banyak pengamatan pada kelas } y}{\text{Banyak pengamatan}}. \tag{2.3}$$

Klasifikasi *Naive* Bayes untuk indikator kontinu dapat dilakukan dengan menggunakan sebaran normal yang dapat dinyatakan sebagai berikut:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}(\frac{x_i-\mu_i}{\sigma_i})^2}. \tag{2.4}$$

Pendekatan sebaran normal ini dilakukan untuk setiap kelas $Y = y$. Nilai μ_i dan σ_i dari sebaran ini diduga langsung dari data, biasanya data *training*.

2.3. *Pengukuran Kinerja Klasifikasi*

Pengukuran kinerja klasifikasi digunakan untuk menentukan nilai akurasi dari suatu klasifikasi dengan menggunakan metriks konfusi. Matriks konfusi merupakan tabel pencatat hasil penghitungan klasifikasi yang terdiri dari dua kelas yaitu kelas 0 dan 1 [3]. Berdasarkan matriks konfusi akan dtentukan akurasi dan eror klasifikasi sebagai berikut.

$$\text{Akurasi} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}, \quad \text{error} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

3. Metodologi Penelitian

3.1. *Sumber Data*

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Survei Sosial Ekonomi Nasional (SUSENAS) yang dilaksanakan oleh Badan Pusat Statistik (BPS) pada tahun 2018. Data lengkap dari SUSENAS tersebut disajikan dalam buku Data dan Informasi Kemisikinan Kabupaten/Kota tahun 2018. Objek pengamatan pada penelitian adalah 514 kabupaten/kota di Indonesia. Dari seluruh kabupaten/kota tersebut diambil sampel secara *purposive* sekitar 40% dari 122 daerah tertinggal, yaitu sebanyak 49 kabupaten/kota pada daerah tertinggal dan sekitar 40% dari 392 untuk daerah tidak tertinggal, yaitu sekitar 159 kabupaten/kota pada daerah tidak tertinggal. Dengan demikian banyak sampel pada data penelitian ini adalah sebanyak 208 kabupaten/kota.

3.2. *Variabel Penelitian*

Variabel yang digunakan pada penelitian ini terdiri dari variabel tak bebas yaitu status daerah kabupaten/kota yang dikategorikan menjadi dua yaitu:

$$\begin{aligned} Y = 1 &\quad \text{untuk daerah tertinggal,} \\ Y = 0 &\quad \text{untuk daerah tidak tertinggal.} \end{aligned}$$

Variabel bebas yang digunakan pada penelitian ini, yaitu:

(a) Persentase Penduduk Miskin (X_1).
(b) Angka Melek Huruf (X_2).
(c) Angka Harapan Hidup (X_3).
(d) Jumlah Puskesmas (X_4).

3.3. *Metode Analisis Data*

Langkah-langkah yang dilakukan untuk menganalisis data penelitian adalah:

(1) Melakukan analisis deskriptif terhadap data yang digunakan dalam penelitian ini. Pada tahap ini akan dibandingkan nilai variabel X_1, X_2, X_3, X_4 untuk

daerah tertinggal ($Y = 1$) dengan bukan daerah tertinggal ($Y = 0$). Analisis deskriptif dilakukan dengan menggunakan statistik deskriptif (berupa nilai minimum, nilai maksimum, rata-rata dan ragam), dan boxplot.

(2) Menguji kesamaan matriks ragam peragam keempat variabel untuk daerah tertinggal dan bukan daerah tertinggal dengan menggunakan uji Bartlett. Hipotesis yang diuji adalah

$$H_0 : \sum_0 = \sum_1,$$
$$H_1 : \sum_0 \neq \sum_1.$$

(3) Melakukan uji kesamaan vektor nilai tengah kedua daerah dengan uji T^2-Hotteling dengan hipotesis:

$$H_o : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Statistik uji yang digunakan tergantung dari hasil uji Bartlett.

(4) Membagi data asli menjadi data *training* dan data *testing* menggunakan metode *Holdout*. Pada penelitian ini banyak data *training* adalah 166 data (80%) dan data *testing* sebanyak 42 data (20%).

(5) Menggunakan metode *Naive* Bayes pada data *training* untuk mendapatkan model untuk klasifikasi daerah tertinggal dan bukan daerah tertinggal. Langkah-langkah yang dilakukan adalah

- (•) Menduga probabilitas prior $P(Y = y)$; $y = 0, 1$
- (•) Menghitung nilai tengah dan ragam dari keempat variabel secara terpisah untuk kedua status daerah. Nilai rata-rata dan ragam tersebut digunakan sebagai penduga parameter dari sebaran normal.
- (•) Menentukan $f_{(X_i|Y)}(x_i|y)$ dengan pendekatan sebaran normal.
- (•) Menentukan model peluang *Naive* Bayes.

(6) Menduga klasifikasi setiap pengamatan yang terdapat pada data *testing*

- (•) Dengan menggunakan model peluang yang diperoleh pada langkah lima dihitung $P(Y = 0|\mathbf{X}=\mathbf{x})$ dan $P(Y = 1|\mathbf{X}=\mathbf{x})$
- (•) Jika $P(Y = 0|\mathbf{X}=\mathbf{x}) > P(Y = 1|\mathbf{X}=\mathbf{x})$ maka kabupaten/kota dikategorikan bukan daerah tertinggal dan sebaliknya.

(7) Menghitung akurasi model

- (•) Membentuk matriks konfusi.
- (•) Menghitung pengukuran kinerja klasifikasi.

4. Hasil dan Pembahasan

4.1. *Prediksi Status Daerah*

Pada proses klasifikasi *Naive Bayes*, pertama data akan dibagi menjadi data *training* dan data *testing* dengan metode *holdout*. Diperoleh data *training* sebanyak 166

data dan data *testing* sebanyak 42 data. Berdasarkan dari hasil pembagian data tersebut, data pada data *training* terdapat 39 daerah yang dikategorikan sebagi daerah tertinggal dan 127 daerah sebagai daerah tidak tertinggal. Selanjutnya akan diduga nilai probabilitas prior $P(Y = y)$ pada masing-masing daerah pada data *training*.

Probabilitas prior bukan daerah tertinggal, $P(Y = 0) = \frac{127}{166} = 0,766$.

Probabilitas prior daerah tertinggal $P(Y = 1) = \frac{39}{166} = 0,234$.

Berikut akan diilustrasikan pendugaan status daerah dari Kabupaten Sumba Barat Daya, dengan persentase penduduk miskin sebesar $28,88$, angka melek huruf sebesar $79,24$, angka harapan hidup $67,76$ dan jumlah puskesmas sebanyak 13. Akan diduga peluang Kabupaten Sumbar Barat Daya masuk ke dalam kabupaten daerah tertinggal atau bukan. Berikut akan diduga klasifikasi peluang pada data *testing* berdasarkan model yang diperoleh dari data *training*.

$$f_{X_1|Y}(x_1 = 28,88|0) = \frac{1}{\sqrt{2(3,14)(18,6)}}\epsilon^{-}\frac{1}{2}(\frac{28,88 - 8,72}{4,31})^2 = 1,68 \times 10^{-6}$$

$$f_{X_2|Y}(x_2 = 79,24|0) = \frac{1}{\sqrt{2(3,14)(69,2)}}\epsilon^{-}\frac{1}{2}(\frac{79,24 - 96,66}{8,31})^2 = 0,00534$$

$$f_{X_3|Y}(x_3 = 67,76|0) = \frac{1}{\sqrt{2(3,14)(8,71)}}\epsilon^{-}\frac{1}{2}(\frac{67,76 - 69,88}{2,96})^2 = 0,10423$$

$$f_{X_4|Y}(x_4 = 13|0) = \frac{1}{\sqrt{2(3,14)(158,24)}}\epsilon^{-}\frac{1}{2}(\frac{13 - 19,79}{12,57})^2 = 0,0274$$

Peluang untuk daerah tidak tertinggal

$$\begin{aligned} P(Y = 0|X = x) &= \prod_{i=1}^{4} f_{X_i|Y}(x_i|0)p(y = 0) \\ &= ((1,68 \times 10^{-}6)(0,00534)(0,10423)(0,02741))(0,766) \\ &= 1,96319 \times 10^{-11}. \end{aligned}$$

$$f_{X_1|Y}(x_1 = 28,88|1) = \frac{1}{\sqrt{2(3,14)(73,51)}}\epsilon^{-}\frac{1}{2}(\frac{28,88 - 18,26}{8,57})^2 = 0,02161$$

$$f_{X_2|Y}(x_2 = 79,24|1) = \frac{1}{\sqrt{2(74,88)}}\epsilon^{-}\frac{1}{2}(\frac{79,24 - 92,13}{8,65})^2 = 0,01520$$

$$f_{X_3|Y}(x_3 = 67,76|1) = \frac{1}{\sqrt{2(3,14)(8,5)}}\epsilon^{-}\frac{1}{2}(\frac{67,76 - 66,04}{2,91})^2 = 0,11503$$

$$f_{X_4|Y}(x_4 = 13|1) = \frac{1}{\sqrt{2(3,14)(86,55)}}\epsilon^{-}\frac{1}{2}(\frac{13-17,38}{9,3})^2 = 0,03838$$

Peluang untuk daerah tertinggal adalah

$$\begin{aligned} P(Y = 1|X = x) &= \prod_{i=1}^{4} f_{X_i|Y}(x_i|0)p(y = 1) \\ &= ((0,02161)(0,01520)(0,11503)(0,03838))(0,234) \\ &= 3,40957 \times 10^{-7}. \end{aligned}$$

Karena $\prod_{i=1}^{4} f_{X_i|Y}(x_i|1)p(y = 1) > \prod_{i=1}^{4} f_{X_i|Y}(x_i|0)p(y = 0)$ maka dengan metode *Naive* Bayes, Kabupaten Sumba Barat Daya dengan persentase penduduk miskin sebesar $28,88$, angka melek huruf sebesar $79,24$, angka harapan hidup $67,76$ dan jumlah puskesmas sebanyak 13 diprediksi masuk ke dalam kelompok daerah tertinggal. Hal ini berarti bahwa metode *Naive* Bayes mengklasifikasikan Kabupaten Sumba Barat Daya secara benar.

4.2. *Pengukuran Kinerja Klasifikasi*

Berdasarkan klasifikasi peluang pada data *testing* akan dihitung akurasi ki-nerja klasifikasi dengan matriks konfusi. Diperoleh nilai akurasi sebagai berikut.

$$\begin{aligned} Akurasi &= \frac{6+32}{6+4+0+32} = 0,905, \\ error &= \frac{4+0}{6+4+0+32} = 0,095. \end{aligned}$$

5. Kesimpulan

Pendugaan status daerah yang terdiri dari 208 kabupaten/kota, diperoleh nilai akurasi sebesar $90,5\%$ dan error sebesar $9,5\%$. Metode pengklasifikasian dengan metode *Naive* Bayes pada kajian ini cukup baik dalam mengklasifikasi status daerah berdasarkan variabel yang dipublikasikan pada tahun 2018.

6. Ucapan Terima kasih

Penulis mengucapkan terima kasih kepada bapak Dr. Dodi Devianto, bapak Yudiantri Asdi, M.Sc dan bapak Dr. Jenizon yang telah memberikan masukan dan saran sehingga makalah ini dapat diselesaikan dengan baik.

Daftar Pustaka

[1] Han, J., Kamber, M. and Pei, J. 2011. *Data Mining Concepts and Techniques.* Morgan Kaufmann, British.

[2] John, G H. 1995. *Estimating Countinuos Distributions in Bayesian Classifier.* Morgan Kaufman, California.

[3] Prasetyo, E. 2014. *Data Mining Mengolah Data Menjadi Informasi menggunakan Matlab.* Andi Offset, Yogyakarta.

[4] Republik Indonesia. 2015. *Peraturan Presiden Republik Indonesia No 131 Tahun 2015 Tentang Penetapan Daerah Tertinggal Tahun 2015-2019*. Sekretariat Negara. Jakarta.
[5] Walpole, R. E. 1988. *Pengantar Statistika Edisi Tiga*. PT Gramedia, Jakarta.