ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 6, Nomor 2, Tahun 2017, Halaman 193-200
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

KLASIFIKASI CALON PENDONOR DARAH MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* (Studi Kasus : Calon Pendonor Darah di Kota Semarang)

Dhimas Bayususetyo[1], Rukun Santoso[2], Tarno[3]
[1]Mahasiswa Departemen Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Departemen Statistika FSM Universitas Diponegoro

ABSTRACT

Classification is the process of finding a model or function that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. There are some methods that are included in the classification methods, one of them is *Naïve Bayes*. *Naïve Bayes* is a prediction technique that based simple probabilistic are based on the application of Bayes theorem with strong independence assumption. On this study carried out correction to the *Naïve Bayes* method in calculating the conditional probability of each feature using two approaches, normal density function and cumulative distribution function approaches. These two approaches are used to classify prospective blood donors in Semarang City. The predictor variables used are hemoglobin level, upper blood pressure, lower blood pressure, and weight. The result of this study shows that both approaches have the same *Matthews Correlation Coefficient* (MCC) values, 0.8985841 or close to +1. It means that both approaches equally well doing classification.

Keywords: Classification, Naïve Bayes, Normal Density Function, Cumulative Distribution Function, Blood Donors, *Matthews Correlation Coefficient* (MCC).

1. PENDAHULUAN

Donor darah adalah proses menyalurkan darah atau unsur-unsur darah dari satu orang ke sistem peredaraan orang lainnya. Banyak orang yang tidak tahu tentang manfaat donor darah bagi kesehatan. Padahal dengan melakukan donor darah, maka sel-sel darah di dalam tubuh menjadi lebih cepat terganti dengan yang baru. Dengan meningkatnya permintaan suplai darah di masyarakat, persediaaan darah yang mencukupi sangat dibutuhkan. Meskipun demikian, pendonor harus terlebih dahulu menjalani pemeriksaan kesehatan, baik pengukuran tekanan darah, golongan darah, kadar hemoglobin (Hb) maupun konsultasi medis (Depkes RI, 2009).

Pada penelitian ini digunakan metode *Naïve Bayes* dalam mengklasifikasi dan memprediksi seseorang apakah bisa mendonorkan darahnya atau tidak, berdasarkan kadar hemoglobin, tensi atas, tensi bawah, berat badan, dan usia yang dimilikinya sebagai variabel pendukung. Menurut Han dan Kamber (2006), klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat dipergunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Ada beberapa metode yang termasuk dalam metode klasifikasi, salah satunya adalah *Naïve Bayes*. Menurut Prasetyo (2013), *Naïve Bayes* merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema *Bayes* dengan asumsi independensi yang kuat. *Naïve Bayes* dapat diterapkan pada data fitur kategorik maupun kontinu. Pada data fitur yang bersifat kontinu, *Naïve Bayes* mengasumsikan data kontinu ke dalam distribusi tertentu dan memperkirakan parameter

distribusi dengan data latih. Biasanya digunakan distribusi Gaussian untuk menghitung probabilitas bersyarat dari fitur kontinu pada sebuah kelas. Parameter distribusi Gaussian adalah mean dan standar deviasi (Han dan Kamber, 2006).

Pada penelitian ini dalam menghitung nilai probabilitas bersyarat dari fitur kontinu pada sebuah kelas digunakan selisih peluang kumulatif dengan interval tertentu. Setelah itu hasil klasifikasi menggunakan pendekatan fungsi densitas normal dibandingkan dengan hasil klasifikasi menggunakan pendekatan selisih peluang kumulatif. Selain itu penelitian mengenai klasifikasi pendonor darah juga sebelumnya belum pernah dilakukan, sehingga membuat peneliti tertarik menggunakan studi kasus ini. Dalam penelitian ini difokuskan pada perbandingan nilai ketepatan klasifikasi dari metode *Naïve Bayes* yang didekati dengan fungsi densitas dan selisih peluang kumulatif.

2. TINJAUAN PUSTAKA

2.1 Klasifikasi

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat dipergunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Model itu sendiri diperoleh berdasarkan analsis dari data yang sudah diketahui label kelasnya (Han and Kamber, 2006).

Dalam pengklasifikasian terdapat dua tahap di dalamnya, yaitu tahap pengamatan dan tahap pengujian. Tahap pengamatan merupakan tahap ketika algoritma membangun model klasifikasi dari data latih yang sudah diketahui label kelasnya. Sedangkan tahap pengujian merupakan langkah untuk menerapkan model tersebut pada data uji sehingga kelas yang sesungguhnya dari data uji dapat diketahui (Han and Kamber, 2006).

2.2 Probabilitas

2.2.1 Kejadian Saling Asing dan Kejadian Saling Bebas

Menurut Rumsey (2008), dua kejadian, A dan B, dikatakan saling bebas (*independent*) jika $\mathrm{P(A)} = \mathrm{P(A|B)}$. Bisa juga dikatakan bahwa kejadian A tidak dipengaruhi oleh kejadian B atau sebaliknya kejadian B tidak dipengaruhi oleh kejadian A. Sementara itu, menurut Vidakovic, dua kejadian, A dan B dikatakan saling asing (*mutually exclusive*) jika kedua kejadian tersebut tidak memiliki elemen yang sama. Dengan kata lain, mustahil bagi kedua kejadian tersebut terjadi dalam percobaan tunggal atau pada saat yang bersamaan. Oleh karena itu dua kejadian yang saling asing (*mutually exclusive*) tidak memiliki irisan, $\mathrm{P(A \cap B)} = \mathrm{P}(\emptyset) = 0$.

Menurut Kemp (2014), jika A dan B adalah dua kejadian yang saling bebas serta kejadian saling asing, maka berlakuseperti yang tercantum pada Tabel 1.

Tabel 1. Perbandingan Kejadian Saling Asing dan Saling Bebas

Kejadian Saling Asing	Kejadian Saling Bebas
$\mathbf{P(A \cup B) = P(A) + P(B)}$	$\mathrm{P(A \cup B) = P(A) + P(B) - (P(A) \times P(B))}$
$\mathbf{P(A \cap B) = P}(\emptyset) \mathbf{= 0}$	$\mathrm{P(A|B) = P(A)}$
$\mathbf{P(A|B) = 0}$	$\mathrm{P(A|B^c) = P(A)}$
$\mathbf{P(A|B^c) = P(A)/(1 - P(B))}$	$\mathrm{P(A \cap B) = P(A) \times P(B)}$

2.2.2 Teorema Bayes

Menurut Subanar (2013),jika X dan Y adalah suatu kejadian, maka probabilitas bersyarat X terjadi bila diketahui kejadian Y sudah terjadi adalah

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}; P(Y) > 0$$

Misalkan $Y_1, Y_2, \ldots, Y_n$ adalah kejadian-kejadian yang saling asing sedemikian hingga $\bigcup_{i=1}^{n} Y_i = \Omega$, dan X kejadian di dalam Ω, maka

$$X = (X \cap Y_1) \cup (X \cap Y_2) \ldots \cup (X \cap Y_n)$$

$$= \bigcup_{i=1}^{n} (X \cap Y_i)$$

Dengan menggunakan kenyataan kejadian-kejadian $X \cap Y_i$, $i = 1,2,..,n$ saling asing, maka

$$P(X) = \sum_{i=1}^{n} P(X \cap Y_i)$$

$$= \sum_{i=1}^{n} P(X|Y_i)P(Y_i)$$

Misalkan X telah terjadi, untuk menentukan mana salah satu dari Y_j yang terjadi, maka

$$P(Y_j|X) = \frac{P(X \cap Y_j)}{P(X)}$$

$$= \frac{P(X|Y_j)P(Y_j)}{\sum_{i=1}^{n} P(X|Y_i)P(Y_i)}$$

2.3 Fungsi Distribusi Normal

Bila X sebagai sebuah variabel random kontinu yang berdistribusi normal, maka distribusi peluang dari X adalah sebuah fungsi yang berada di antara dua batas, misalnya a dan b dengan $b \geq a$. Berikut persamaan fungsi peluangnya

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Menurut Devore (2008), nilai peluang dari X yang berada di interval [a, b] adalah luas wilayah yang berada di atas sumbu absis dan di bawah kurva normal. Oleh karena itu nilai peluang suatu titik dari variabel random kontinu akan selalu benilai nol karena luas wilayah di bawah kurva densitas dari suatu titik nilainya nol. Dibuktikan dengan persamaan berikut

$$P(X = c) = \int_c^c f(x)dx = 0$$

Akan tetapi sering kali dalam beberapa kasus, dalam mencari nilai peluang suatu titik didekati dengan fungsi densitas normal.

$$P(X = x) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

.

Menurut Devore (2008), fungsi distribusi kumulatif (cdf) $F(x)$ diperoleh dari mengintegralkan fungsi densitas peluang $f(y)$ dengan interval limit $-\infty$ sampai x. Fungsi distribusi kumulatif $F(x)$ dari sebuah variabel random kontinu didefinisikan untuk setiap nilai x oleh

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Untuk setiap x, nilai dari fungsi distribusi kumulatif $F(x)$ adalah luas wilayah di bawah kurva normal yang ditarik dari sebelah kiri atau dari nilai $-\infty$ sampai x.

2.4 Klasifikasi Naïve Bayes

Naïve Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Klasifikasi *Naïve Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar (Han dan Kamber, 2006).

Pada dasarnya maksud dari klasifikasi *Naïve Bayes* adalah mencari peluang bersyarat (posterior) dari dua kejadian, misalnya X dan Y, yang dinotasikan dengan P(Y|X). Jika X adalah vektor masukan yang berisi fitur dan Y adalah label kelas, maka notasi P(Y|X) berarti peluang label kelas Y didapatkan setelah fitur-fitur X diamati. Menurut Prasetyo (2013), formula *Naïve Bayes* untuk klasifikasi adalah

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

Dengan P(X) adalah probabilitas X yang nilainya selalu tetap.

P(Y|X) adalah probabilitas data dengan vektor X pada kelas Y.

P(Y) adalah probabilitas awal kelas Y.

$\prod_{i=1}^{q} P(X_i|Y)$ adalah probabilitas independen kelas Y dari semua fitur dalam vektor X.

Dalam menentukan kelas hasil prediksi yang dibutuhkan hanya nilai maksimum dari $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ karena P(X) bernilai konstan, yang dikenal dengan sebutan *Maximum A Posterior* (MAP) dimana MAP ini dapat dinotasikan dengan

$$hMAP = \arg(\max P(Y) \prod_{i=1}^{q} P(X_i \mid Y))$$

Menurut Prasetyo (2013), perhitungan peluang bersyarat pada klasifikasi Naïve Bayes dapat dilakukan pada fitur kategorik maupun kontinu. Untuk fitur kontinu, fitur diasumsikan kedalam suatu distribusi, biasanya distribusi normal. Oleh karena itu dalam menghitung P(Xi|Y) didekati dengan

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

Parameter μ_{ij} bisa didapat dari mean sampel x_i ($\bar{X}$) dari semua data latih yang menjadi milik kelas y_j, sedangkan σ^2_{ij} dapat diperkirakan dari varian sampel (s^2) dari data latih.

Pada penelitian ini akan dibandingkan metode Naïve Bayes yang didekati dengan fungsi densitas maupun selisih peluang kumulatif dalam menghitung peluang dari setiap fitur X bersyarat label kelasnya, $P(X_i \mid Y)$, untuk di cari pendekatan mana yang lebih baik dalam melakukan proses klasifikasi.

Fungsi densitas: $P(X_i \mid Y) = P(X = x) = f(x) \times$ nilai koreksi selebar 1

Selisih peluang kumulatif: $P(X_i \mid Y) = P(a \leq X \leq b) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx$

2.5 Pemilihan Fitur Berbasis Statistik

Memilih fitur adalah mangamati setiap fitur yang dibangkitkan secara independen dan menguji kemampuan diskriminasinya pada setiap kelas. Dalam hal ini digunakan uji t untuk memilih fitur yang signifikan kemampuan diskriminasinya pada setiap kelas. Sebelumnya dilakukan uji homogenitas terlebih dahulu untuk menentukan uji t seperti apa yang digunakan. Pengujian hipotesis uji t melakukan pengujian fitur secara individu dan memeriksa ada atau tidaknya informasi diskriminasi data terhadap kelas. Jika tidak ada, maka fitur tersebut akan dibuang. Ide dasarnya adalah menguji apakah nilai rata-rata fitur yang dipunyai berbeda dalam dua kelas secara signifikan.

2.6 Metode Holdout

Menurut Tan *et al* (2006) dalam Prasetyo (2014), dalam metode Holdout, data mentah yang sudah diketahui label kelasnya dibgi menjadi dua bagian terpisah, yaitu set data latih dan set data uji. Proporsi tersebut bebas sesuai pertimbangan peneliti.

2.7 *Matthews Correlation Coefficient* (MCC)

Matthews Correlation Coefficient (MCC) adalah ukuran kualitas klasifikasi biner (2 kelas). Metode ini menggunakan nilai-nilai yang ada pada matriks konfusi sebagai dasar perhitungannya. MCC menjadi solusi untuk klasifikasi yang memiliki jumlah kelas yang berbeda jauh. Pada dasarnya MCC merupakan koefisien korelasi antara data latih dan data uji pada klasifikasi biner dan menghasilkan nilai antara -1 sampai +1.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data primer, yaitu data calon pendonor darah di Kota Semarang terhitung dari tanggal 16 – 27 September 2016. Data tersebut diperoleh dari mobile unit UDD PMI Kota Semarang yang mengadakan kegiatan donor darah di Bank Danamon Kota Semarang, SMA 2 Semarang, R.S Telogorejo, dan Kampus Polines. Variabel yang digunakan dalam penelitian ini terdiri dari status calon pendonor sebagai variabel dependen dan berat badan, usia, tensi atas/sistole, tensi bawah/diastole, dan kadar hemoglobin sebagai variabel independen.

Langkah-langkah yang dilakukan untuk menganalisis data penelitian adalah:

1. Mengumpulkan data calon pendonor yang akan digunakan dalam penelitian.
2. Menentukan varian kedua kelas dari sebuah fitur sama atau berbeda menggunakan uji homogenitas.
3. Memilih fitur yang sesuai dengan metode yang akan digunakan menggunakan uji t atau *Welch's t-test.*
4. Membagi data tersebut menjadi data observasi dan data uji menggunakan metode Holdout dengan mempertimbangkan kosistensi nilai ketepatan model klasifikasi.
5. Menghitung probabilitas prior (P(Y)) dari data uji berdasarkan data observasi.
6. Menghitung probabilitas atribut terhadap masing-masing kelas ($P(X_i|Y)$) pada data uji berdasarkan data observasi menggunakan pendekatan fungsi densitas normal.
7. Menghitung perkalian probabilitas prior dengan semua probabilitas atribut pada masing-masing kelas ($P(Y)\prod_{i=1}^{k} P(Xi|Y)$).
8. Mencari nilai maksimal dari $\frac{P(Y)\prod_{i=1}^{k} P(X_i|Y)}{P(X)}$ pada kedua kelas. Nilai terbesar dari perhitungan tersebut merupakan hasil prediksi.
9. Menghitung probabilitas atribut terhadap masing-masing kelas ($P(X_i|Y)$) pada data uji berdasarkan data observasi menggunakan pendekatan selisih peluang kumulatif.
10. Menghitung perkalian probabilitas prior dengan probabilitas atribut pada masing-masing kelas ($P(Y)\prod_{i=1}^{k} P(Xi|Y)$).
11. Mencari nilai maksimal dari $\frac{P(Y)\prod_{i=1}^{k} P(X_i|Y)}{P(X)}$ pada kedua kelas. Nilai terbesar dari perhitungan tersebut merupakan hasil prediksi.
12. Pengukuran kinerja klasifikasi dengan menghitung nilai ketepatan model klasifikasidari masing-masing pendekatan.
13. Membandingkan nilai ketepatan model klasifikasi dari masing-masing pendekatan.

4. HASIL DAN PEMBAHASAN

4.1 Deskripsi Data

Deskripsi data digunakan untuk mengetahui gambaran umum tentang data yang digunakan. Adapun gambaran umum status calon pendonor darah dapat dilihat pada Tabel 2 berikut ini

Tabel 2. Proporsi Label Kelas Calon Pendonor

Status Calon Pendonor	Frekuensi	Persentase
Pendonor	221	71,06%
Bukan Pendonor	90	28,94%
Total	311	100%

Berikut gambaran umum fitur-fitur calon pendonor darah dapat dilihat pada Tabel 3 berikut ini

Tabel 3. Fitur-Fitur Calon Pendonor Darah

Fitur	Mean	Std. Deviasi	Minimum	Maximum
Hemoglobin	13,87	1,43	9,1	16,9
Tensi Atas	119,132	13,54	90	180
Tensi Bawah	77,51	8,5	60	110
Berat Badan	64,66	13,18	41	105
Usia	28,5	11,55	17	61

4.2 Pemilihan Fitur Berbasis Statistika

4.2.1 Uji Homogenitas Variansi dari Setiap Fitur

Sebelum dilakukan uji t untuk menentukan fitur mana saja yang signifikan mempengaruhi label kelas, dilakukan uji homogenitas terlebih dahulu untuk mengetahui apakah varian dari setiap kelas di suatu fitur sama atau tidak. Berikut hasil uji homogenitas dari setiap fitur yang bisa dilihat pada Tabel 4.

Tabel 4. Hasil Uji Homogenitas

Fitur	F Hitung	F Tabel	Keputusan	Kesimpulan
Hemoglobin	2,9426	1,3270	H_0 ditolak	Varian hemoglobin pada kelas pendonor dan bukan pendonor berbeda.
Tensi Atas	4,9419	1,3270	H_0 ditolak	Varian tensi atas pada kelas pendonor dan bukan pendonor berbeda.
Tensi Bawah	2,4735	1,3270	H_0 ditolak	Varian tensi bawah pada kelas pendonor dan bukan pendonor berbeda.
Berat Badan	1,4713	1,3270	H_0 ditolak	Varian berat badan pada kelas pendonor dan bukan pendonor berbeda.
Usia	1,0600	1,3270	H_0 diterima	Varian usia pada kelas pendonor dan bukan pendonor sama.

4.2.2 Uji Beda Rata-Rata Fitur (Uji t)

Setelah dilakukan uji homogenitas pada setiap fiturnya, kemudian dilakukan uji t berdasarkan hasil dari uji homogenitas. Berikut hasil uji t dai setiap fitur yang bisa dilihat pada Tabel 5.

Tabel 5. Hasil Uji t

Fitur	t Hitung	t Tabel	Keputusan	Kesimpulan
Hemoglobin	8,8994	1,3270	H_0 ditolak	Hemoglobin memiliki rata-rata yang berbeda signifikan antara pendonor dan bukan pendonor.
Tensi Atas	2,4381	1,3270	H_0 ditolak	Tensi atas memiliki rata-rata yang berbeda signifikan antara pendonor dan bukan pendonor.
Tensi Bawah	5,4320	1,3270	H_0 ditolak	Tensi bawah memiliki rata-rata yang berbeda signifikan antara pendonor dan bukan pendonor.
Berat Badan	5,0998	1,3270	H_0 ditolak	Berat badan memiliki rata-rata yang berbeda signifikan antara pendonor dan bukan pendonor.
Usia	-0,57948	1,3270	H_0 diterima	Rata-rata usia pada pendonor dan bukan pendonor tidak berbeda secara signifikan.

4.3. Ketepatan Klasifikasi

Setelah dilakukan proses klasifikasi terhadap data uji menggunakan dua pendekatan yaitu pendekatan fungsi densitas normal dan pendekatan selisih peluang kumulatif dengan proporsi 80% data latih dan 20% data uji diperoleh nilai *Matthews Correlation Coefficient* (MCC). MCC untuk pendekatan fungsi densitas yaitu sebesar 0,8985841 atau mendekati +1, yang berarti klasifikasi berjalan baik. Pendekatan selisih peluang kumulatif juga menghasilkan nilai MCC yaitu sebesar 0,8985841 atau mendekati +1, yang berarti klasifikasi berjalan baik.

5. KESIMPULAN

Klasifikasi data calon pendonor darah di Kota Semarang menggunakan metode *Naïve Bayes* yang didekati dengan fungsi densitas maupun selisih peluang kumulatif, menghasilkan nilai *Matthews Correlation Coefficient* (MCC) yang sama yaitu sebesar 0,8985841 atau mendekati +1. Sehingga dapat disimpulkan bahwa kedua pendekatan tersebut, baik pendekatan fungsi densitas ataupun selisih peluang kumulatif memiliki kemampuan yang cukup baik dan sama baiknya dalam mengklasifikasi data calon pendonor darah.

6. DAFTAR PUSTAKA

Depkes RI. 2009. *Donor Darah, Hidup Sehat Sambil Beramal*. Jakarta. www.health.detik.com

Devore, J. L. 2008. *Probability and Statistics for Engineering and the Sciences.* Enhanced Review Edition.

Han, J., and Kamber, M. 2006. *Data Mining Concepts and Techniques.* Second Edition. California: Morgan Kaufman.

Prasetyo, E. 2012a. *Data Mining Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

Prasetyo, E. 2014. *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

Rumsey, D. J. *Letting Go to Grow: Independent vs Mutually Exclusive.* Jounal of Statistics Education. Vol. 16. Number 3, 2008. Diambil dari http://ww2.amstat.org/publications/jse/v16n3/rumsey.html. (28 Februari 2017)

Subanar. 2006. *Pengantar Teori Ukuran dan Probabilitas*.Yogyakarta: FMIPA UGM.

Vidakovic, B. *Probability, Conditional Probability, and Bayes Formula.* http://-www2.isye.gatech.edu/~brani/isyebayes/bank/handout1.pdf. Diakses pada 28 Februari 2017.