JSA
Jurnal Statistika dan Aplikasinya

Volume 3 Issue 1,
June 2019

e-ISSN: 2620-8369

Received: 9 Maret 2019
Revised: 5 April 2019
Accepted: 17 April 2019
Published: 28 Juni 2019

# METODE *NAIVE BAYES* DALAM MENDETEKSI SEL KANKER PAYUDARA

Riska Agustin[1, a)], Vera Maya Santi[2, b)], Bagus Sumargo[2, c)]

[1]*Program Studi Matematika, Fakultas Matematika dan Ilmu Pengatahuan Alam, Universitas Negeri Jakarta*
*Jl. Rawamangun Muka, Rawamangun, Jakarta Timur, DKI Jakarta.*
[2]*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengatahuan Alam, Universitas Negeri Jakarta*
*Jl. Rawamangun Muka, Rawamangun, Jakarta Timur, DKI Jakarta.*

Email: [a)]riskaagustin.ra@gmail.com, [b)]vera.indr4@gmail.com, [c)]bagusumargo63@gmail.com.

**Abstract**

Breast cancer is the most common cancer and occurs in women around the world. In Indonesia, breast cancer cases occupy the first position of cancer cases that cause the most deaths in women. The high mortality rate due to breast cancer in Indonesia needs to be observed with preventive measures and early detection. Cancer that is found at an early stage and get a fast and appropriate treatment will provide a greater chance of recovery. There are several factors that can be used to predict the presence of breast cancer in a person's body, including age, glucose, insulin, HOMA (Homeostatic Model Assessment), leptin, adiponectin, resistin, MCP-1 (Monocyte Chemoattractant Protein) -1), as well as BMI (Body Mass Index). This research will discuss the possibility of a person suffering from breast cancer based on the driving factors observed using the Naive Bayes method. The accuracy obtained is 80%, from the 35 tests data observed there are 28 data that is classified correctly and 7 data that has been misclassified.

**Keywords**: breast cancer, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1, BMI, Naive Bayes.

**Abstrak**

Kanker payudara merupakan kanker paling umum dan paling banyak terjadi pada wanita di seluruh dunia. Di Indonesia sendiri kasus kanker payudara menempati posisi pertama kanker yang paling banyak menyebabkan kematian pada wanita. Tingginya angka kematian karena kanker payudara di Indonesia perlu dicermati dengan tindakan pencegahan dan deteksi dini. Kasus kanker yang ditemukan pada stadium dini serta mendapat pengobatan yang cepat dan tepat akan memberikan kesembuhan dan harapan hidup lebih lama. Ada beberapa faktor yang dapat digunakan untuk memprediksi keberadaan kanker payudara pada tubuh seseorang, diantaranya adalah usia, glukosa, insulin, HOMA (*Homeostatic Model Assessment*), leptin, adiponektin, resistin, MCP-1 (*Monocyte Chemoattractant Protein-1*), serta IMT (Indeks Massa Tubuh). Pada penelitian ini akan dibahas mengenai kemungkinan seseorang menderita kanker payudara berdasarkan faktor-faktor pendorong yang diamati menggunakan metode *Naive Bayes*.

Hasil akurasi yang diperoleh adalah sebesar 80% dimana dari 35 data uji yang diamati terdapat 28 data yang di klasifikasikan dengan tepat dan 7 data yang mengalami kesalahan klasifikasi.

**Kata-kata kunci**: kanker payudara, glukosa, insulin, HOMA, leptin, adiponektin, resistin, MCP-1, BMI, *Naive Bayes*.

## PENDAHULUAN

Berdasarkan data WHO September 2018, sudah terdapat 2.088.849 wanita yang terdiagnosa kanker payudara dan diprediksi jumlahnya akan meningkat tiap tahunnya. Kasus kanker payudara terbanyak dapat ditemukan di benua Asia dengan 43.3% kasus dari seluruh kasus kanker payudara di dunia serta tingkat kematian 49,4% dari seluruh kematian yang terjadi.

Tingginya angka kematian karena kanker payudara perlu dicermati dengan tindakan pencegahan dan deteksi dini. Kasus kanker yang ditemukan pada stadium dini serta mendapat pengobatan yang cepat dan tepat akan memberikan kesembuhan dan harapan hidup lebih lama (Sun *et al*, 2017).

Terdapat beberapa faktor yang dapat digunakan untuk memprediksi kanker payudara pada tubuh pasien. Faktor-faktor tersebut diperoleh dari data konsultasi rutin dan pemeriksaan darah berupa usia, glukosa, insulin, HOMA (*Homeostatic Model Assessment*), leptin, adiponektin, resistin, MCP-1 (*Monocyte Chemoattractant Protein-1*), serta IMT (Indeks Massa Tubuh) (Patricio *et al*, 2018). Faktor-faktor tersebut dapat digunakan untuk memprediksi berapa persen kemungkinan pasien menderita kanker payudara. Salah satu metode yang dapat digunakan untuk mengetahui peluang tersebut adalah dengan menggunakan metode *Naive Bayes*.

*Naive Bayes* merupakan sebuah metode klasifikasi menggunakan metode probabilitas dan statistik. Algoritma *Naive Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Metode lain yang biasa digunakan untuk proses klasifikasi adalah metode regresi logistik, namun analisis regresi logistik memiliki asumsi-asumsi yang harus terpenuhi antara lain adalah kebebasan antar pilihan variabel respon, terdapat hubungan yang linier antara variabel penjelas yang kontinu dan transformasi fungsi logit dari variabel respon, dan kelompok dari variabel respon tidak terklasfikasikan secara sempurna oleh variabel-variabel penjelas (Santi, 2018). Schoot dan Depaoli (2014) dalam penelitiannya menjabarkan empat alasan utama mengapa metode *Bayesian* sering digunakan: (1) Model yang rumit terkadang tidak bisa diselesaikan dengan metode konvensional, (2) Sebagian besar peneliti lebih memilih menggunakan definisi probabilitas, (3) Latar belakang pengetahuan dapat dimasukkan kedalam analisis, dan (4) Metode *bayesian* tidak memerlukan sampel yang besar.

Oleh karena itu, pada penelitian ini pengklasifikasian data kanker payudara dilakukan dengan menggunakan metode *Naive Bayes*.

## METODE

### Teorema Bayes

Analisis Bayesian menggali dua buah sumber informasi tentang parameter suatu model statistik. Sumber informasi pertama berasal dari sampel dan disebut dengan informasi sampel (*sample information*) dimana informasi sampel digali dari suatu fungsi *likelihood* (*likelihood function*). Sumber informasi kedua berasal dari opini yang bersifat subjektif dan disebut dengan informasi prior (*prior information*). Gabungan dua buah sumber informasi ini akan membentuk informasi posterior (*posterior information*). Penggabungan kedua sumber informasi ini dicapai melalui teorema Bayes (*Bayes theorem*).

Secara umum aturan Bayes yang terbentuk adalah (Bolstad, 2007),

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

**Fungsi *Likelihood***

**Deftnisi 1.** (Bain dan Engelhardt,1992)

Fungsi likelihood adalah fungsi densitas bersama dari n variabel acak $X_1$, $X_2$, ..., $X_n$ dan dinyatakan dalam bentuk $f(x_1, x_2, ..., x_n; \theta)$, jika $x_1, x_2, ..., x_n$ ditetapkan maka fungsi likelihood adalah fungsi dari parameter $\theta$ dan dinotasikan dengan $L(\theta)$. Jika $X_1, X_2, ..., X_n$ menyatakan suatu sampel acak dari $f(x, \theta)$ maka,

$$L(\theta) = f(x_1;\theta) f(x_2;\theta)... f(x_n;\theta)$$
$$= \prod_{i=1}^{n} f(x_i;\theta)$$

***Naive Bayes***

Klasifikasi *Naive Bayes* adalah metode klasifikasi yang didasarkan pada penerapan teorama Bayes. Klasifikasi *Naive Bayes* dapat digunakan untuk memprediksi probabilitas dimasa depan berdasarkan pengalaman masa sebelumnya, dengan asumsi bahwa setiap variabel *X* yang mempengaruhi variabel respon bersifat saling bebas atau tidak memiliki hubungan antar satu dengan lainnya. Asumsi saling bebas pada metode *Naive Bayes* tentunya akan mempermudah perhitungan untuk permasalahan yang kompleks.

Secara umum dengan menerapkan teorema Bayes, klasifikasi *Naive Bayes* dapat dinyatakan dengan rumus sebagai berikut,

$$P(y \mid x_1,...,x_n) = \frac{P(y)P(x_1,...,x_n \mid y)}{P(x_1,...,x_n)} \quad (2)$$

$P(y, x_1, ...x_n)$ merupakan "*joint probability*" pada persamaan Bayes yang ekuivalen dengan pembilang dalam persamaan (2), dengan menggunakan aturan rantai maka diperoleh (Bolstad, 2007),

$$\begin{aligned} P(y, x_1, ...x_n) &= P(x_1, ...x_n, y) \\ &= P(x_1|x_2, ..., x_n, y)P(x_2, ..., x_n, y) \\ &= P(x_1|x_2, ..., x_n, y)P(x_2|x_3..., x_n, y)P(x_3, ..., x_n, y)... \\ &= P(x_1|x_2, ..., x_n, y)P(x_2|x_3..., x_n, y)...P(x_{n-1}|x_n, y)P(y) \\ &= P(x_i|x_{i+1}, ..., x_n, y)P(y) \end{aligned}$$

Dengan asumsi tersebut, maka berlaku suatu persamaan sebagai berikut:

$$P(x_i|x_{i+1}, ..., x_n, y) = P(x_i|y) \qquad i = 1, ..., n$$

**Deftnisi 2.** (Bain dan Engelhardt,1992)

*K* kejadian $A_1, A_2, ..., A_k$ dikatakan *independent* atau *mutually independent* jika untuk setiap $j = 2, 3, ..., k$ dan setiap subset pada indeks yang berbeda $i_1, i_2, ..., i_j$

$$P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_j}) = P(A_{i_1})P(A_{i_2})...P(A_{i_j}) \quad (3)$$

Berdasarkan Definisi (2) untuk semua *i* hubungan ini disederhanakan menjadi,

$$P(y \mid x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1,...,x_n)} \quad (4)$$

Karena $P$ ($x_1$, ..., $x_n$) bernilai konstan pada setiap kelasnya maka persamaan (4) dapat ditulis menjadi,

$$P(y \mid x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

Simbol ($\propto$) pada persamaan menyatakan bahwa $P$ ($y|x_1$, ..., $x_n$) proporsional atau sebanding dengan $P(y)\prod_{i=1}^{n} P(x_i \mid y)$ dengan menggunakan terminologi probabilitas Bayesian, persamaan di atas dapat ditulis sebagai,

$$Posterior \propto prior \times likelihood$$

### *Confusion Matrix*

*Confusion Matrix* merangkum kinerja dari suatu metode klasifikasi dengan memperhatikan beberapa data uji. *Confusion Matrix* merupakan sebuah tabel dengan dua baris dan dua kolom yang menampilkan jumlah *True Positive*, F*alse Positve*, *False Negative*, dan *True Negative*. Hal ini memungkinkan analisis yang lebih rinci dari sekedar proporsi klasifikasi yang benar (akurasi) (Visa *et al*, 2011).

**TABEL 1.** *Confusion Matrix*

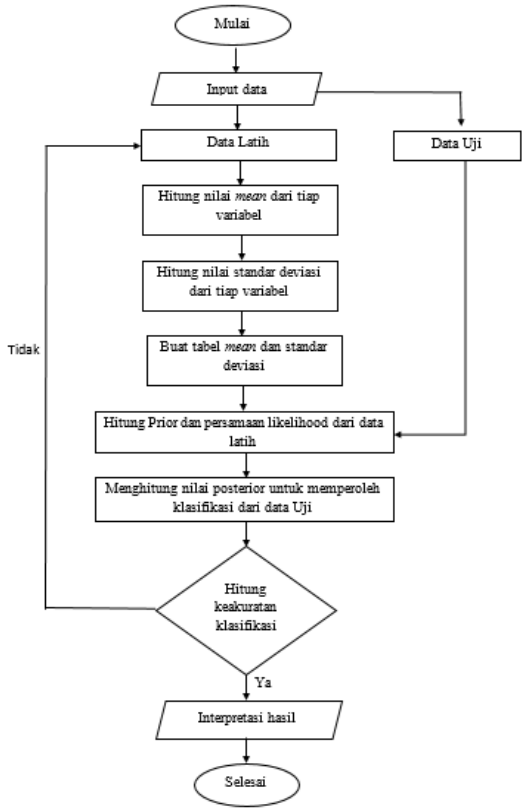
| **Aktual** | **Prediksi** | | **Total** |
|---|---|---|---|
| | **P** | **N** | |
| **P** | TP | FN | TP+FN |
| **N** | FP | TN | FP+TN |
| **Total** | TP+FP | FN+TN | |

Berdasarkan Tabel 1 nilai akurasi dari suatu klasifikasi diperoleh dengan membagi total data yang diklasifikasikan dengan tepat dengan seluruh data yang diamati, atau dapat dinyatakan dengan,

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

Hasil yang diperoleh dari persamaan (5) menggambarkan ketepatan dari proses klasifikasi. Jadi semakin tinggi hasil akurasi maka semakin tepat proses klasifikasinya.

**Desain Model**




**Gambar 1.** Desain Model

## HASIL DAN PEMBAHASAN

### Deskripsi Data

Data yang digunakan dalam penelitian adalah dataset kanker payudara Coimbra tahun 2018 yang diperoleh dari UCI (*University of California*, Irvine) *Machine Learning Repository*. Data terdiri dari 116 responden yang dibagi menjadi data latih dan data uji, dimana 70% dari total data yaitu sebanyak 81 data dijadikan data latih dan 30% dari total data yaitu sebanyak 35 data dijadikan data uji (Phim *et al*, 2016). Variabel respon yang diamati terdiri dari dua kategori yaitu positif mengidap kanker payudara atau tidak mengidap kanker payudara. Deskripsi data amatan ditampilkan pada Tabel 2.

**Tabel 2.** Dekripsi Dataset

| Variabel | Keterangan | Simbol |
|---|---|---|
| Y(respon) | 1: Positif Kanker Payudara<br>0: Negatif Kanker Payudara | $y_i$, dengan $i = 0, 1$ |
| Usia | Tahun | $x_1$ |
| IMT (Index Massa Tubuh) | $kg/m^2$ | $x_2$ |
| Glukosa | *mmol/L* | $x_3$ |
| Insulin | *µU/mL* | $x_4$ |
| HOMA (*Homeostatic Model Assessment*) | Glukosa × Insulin | $x_5$ |
| Leptin | *ng/mL* | $x_6$ |
| Adiponektin | *µg/mL* | $x_7$ |
| Resistin | *ng/mL* | $x_8$ |
| MCP.1 (*Monocyte Chemoattractant Protein-1*) | *pg/dL* | $x_9$ |

**Distribusi Prior**

Jika pada suatu penelitian tidak diketahui secara pasti tentang sebaran dari suatu parameter yang diamati, maka cara yang paling umum adalah dengan menganggap seluruh anggota parameter memiliki peluang yang sama atau dengan kata lain menggunakan distribusi seragam (Stone, 2013).

**Definisi 3.** (Walpole dan Myers, 1995)

Bila peubah acak X mendapat nilai $x_1$, $x_2$, ..., $x_k$ dengan peluang yang sama, maka distribusi seragam diskrit diberikan oleh,

$$f(x;k) = \frac{1}{k}, x = x_1, x_2, ..., x_k \quad (6)$$

Lambang $f(x; k)$ telah dipakai sebagai pengganti $f(x)$ untuk menunjukkan bahwa distribusi seragam tersebut bergantung pada parameter $k$.

**Teorema 4.2.1.** (Walpole dan Myers, 1995)
Rataan distribusi seragam diskrit $f(x; k)$ adalah,

$$\mu = \frac{\sum_{i=1}^{k} x_i}{k} \quad (7)$$

Bukti. Menurut definisi $\mu = E(X)$, maka,

$$\mu = E(X) = \sum_{i=1}^{k} x_i f(x;k) = \sum_{i=1}^{k} \frac{x_i}{k} = \frac{\sum_{i=1}^{k} x_i}{k}$$

Dengan menggunakan persamaan (7) maka informasi awal atau (*prior information*) yang diperoleh adalah, P ($y_1$) = 0, 51 yaitu terdapat 51% sampel pada data latih yang terdiagnosis dengan kanker payudara, dan P ($y_0$) = 0, 49 yaitu terdapat 49% sampel pada data latih yang tidak terdiagnosis dengan kanker payudara.

**Persamaan Likelihood**

Variabel X pada dataset merupakan data numerik sehingga tipe *Naive Bayes* yang digunakan adalah *Gaussian Naive Bayes*. Dalam *Gaussian Naive Bayes* distribusi data yang digunakan adalah distribusi normal.

**Deftnisi 4.** (Walpole dan Myers, *1995*)

Fungsi kepadatan peubah acak normal $X$, dengan rataan $\mu$ dan variansi $\sigma^2$ adalah,

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty \quad (8)$$

dengan $\pi$ = 3, 14159.. dan $e$ = 2, 71828..

Misal diambil salah satu peluang bersyarat dari pengamatan tunggal $f(x_k|\mu)$, karena $f(x_k|\mu)$ dianggap menyebar normal maka $f(x_k|\mu)$ mengikuti fungsi dari distribusi normal,

$$f(x_k \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_k-\mu)^2} \quad (9)$$

Bagian yang tidak bergantung pada parameter $\mu$ memiliki nilai yang sama untuk semua parameter sehingga dapat dianggap sebagai konstanta. Persamaan dari peluang gabungan yang terbentuk adalah,

$$f(x_k \mid \mu) \propto e^{-\frac{1}{2\sigma^2}(x_k-\mu)^2} \tag{10}$$

**Distribusi Posterior**

Dibandingkan penelitian tunggal, biasanya suatu penelitian dibangun dari beberapa sampel acak misal $x_1$, ..., $x_n$. Posterior selalu proposional terhadap *prior* × *likelihood*. Semua sampel acak pada penelitian saling bebas satu sama lain, jadi persamaan *likelihood* gabungan dari sampel adalah perkalian dari *likelihood* pada penelitian tunggal, yang ditunjukkan pada persamaan (11).

Untuk sampel acak $x_1, x_2, ..., x_n$ akan terbentuk persamaan yang akan dijabarkan sebagai berikut.

$$\begin{aligned} f(x_1, x_2, ..., x_n \mid \mu) &= \prod_{i=1}^{n} f(x_i \mid \mu) \\ &= f(x_1 \mid \mu) \times f(x_2 \mid \mu) \times ... \times f(x_n \mid \mu) \end{aligned} \tag{11}$$

Persamaan posterior yang terbentuk dengan prior yang diskrit adalah,

$$\begin{aligned} g(\mu \mid x_1, ..., x_n) &\propto g(\mu) \times f(x_1 \mid \mu) \times f(x_2 \mid \mu) \times ... \times f(x_n \mid \mu) \\ &\propto g(\mu) \times e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \times ... \times e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} \end{aligned} \tag{12}$$

***Confusion Matrix***

Uji tingkat kebaikan model dapat dilihat dari nilai akurasi, *precision*, *recall*, *f-measure*, serta *ROC area* yang diperoleh dengan menggunakan *Confusion Matrix*.

**Tabel 3.** *Confusion Matrix*

| **Aktual** | **Prediksi** | | **Total** |
|---|---|---|---|
| | Terdeteksi | Tidak Terdeteksi | |
| Terdeteksi | 10 | 2 | 12 |
| Tidak Terdeteksi | 5 | 18 | 23 |
| **Total** | 15 | 20 | 35 |

$$\begin{aligned} Akurasi &= \frac{TP + TN}{TP + FP + FN + TN} \\ &= \frac{10 + 18}{10 + 2 + 5 + 18} \\ &= \frac{28}{35} \\ &= 0.8 \end{aligned}$$

Hasil pengklasifikasian data penderita kanker payudara menghasilkan tingkat akurasi sebesar 80%. Lalu dengan menggunakan bantuan *software* Weka diproleh nilai *precision*, *recall*, *f-measure*, serta *ROC area* yang ditunjukan pada Tabel 4.

**TABEL 4.** *Output* Weka

| | ***Precision*** | ***Recall*** | ***F-Measure*** | ***ROC Area*** |
|---|---|---|---|---|
| Terdeteksi | 0.900 | 0.783 | 0.837 | 0.906 |
| Tidak | 0.667 | 0.833 | 0.741 | 0.906 |

| | | | | |
|---|---|---|---|---|
| Terdeteksi | | | | |
| Rata-Rata | 0.7835 | 0.808 | 0.789 | 0.906 |

Rata-rata tingkat presisi, *recall*, *f-measure* berada di atas 75%. Serta nilai ROC area 9.06, nilai dari area ROC hanya terdiri dari 0 sampai 1. Semakin nilai dari area ROC mendekati 1 maka akan semakin baik hasil klasifikasinya. Karena nilai ROC 0.906 maka hasil klasifikasi sel kanker payudara menggunakan metode *Naive Bayes* merupakan klasifikasi yang sangat baik.

## KESIMPULAN DAN SARAN

### Kesimpulan

Setelah melalui tahap perhitungan tingkat akurasi metode *Naive Bayes* dalam mengklasifikasikan data kanker payudara adalah 80%. Dari 35 data uji yang diamati hanya terdapat 7 data yang mengalami kesalahan klasifikasi. Jadi dapat disimpulkan bahwa metode *Naive Bayes* dapat mengklasifikasikan data kanker payudara dengan baik.

### Saran

Asumsi saling bebas yang dimiliki metode *Naive Bayes* dapat memberikan keuntungan karena mempermudah perhitungan namun dapat memberikan tingkat akurasi yang cukup baik. Pada penelitian real sulit ditemukannya variabel yang saling bebas satu sama lain. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan metode klasifikasi yang tidak mengabaikan keterkaitan antar variabelnya. Selain itu, disarankan untuk menggunakan metode klasifikasi lain sebagai pembanding dengan metode *Naive Bayes*.

## UCAPAN TERIMA KASIH

Terima kasih kepada Ibu Vera Maya Santi, M.Si. dan Bapak Dr.Ir.Bagus Sumargo, M.Si. atas saran dan ketersiadaan waktu dalam membimbing penulis, sehingga tulisan ini berhasil selesai dengan baik.

## REFERENSI

Bain, Lee J., and Max Engelhardt. (1992) *Introduction to Probability and Mathematical Statistics Second Edition*, California: Duxbury Thomson Learning.

Bolstad, W.M. (2007) *Introduction to Bayesian Statistics Second Edition.* A John Wiley & Sons.Inc; America.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. and Caramelo, F. (2018) 'Using Resistin, glucose, age and BMI to predict the presence of breast cancer'. *BMC cancer*, 18(1), p.29.

Pham, B.T., Bui, D., Prakash, I. and Dholakia, M. (2016) 'Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS'. *Journal of Geomatics*, 10(1), pp.71-79.

Santi, V.M. (2018) 'Pengembangan Model Regresi Logistik Multinomial untuk Klasifikasi Politik pada Pemilihan Umum', *Jurnal Statistika dan Aplikasinya*, *2*(1), pp.37-43.

van de Schoot, R. and Depaoli, S. (2014) 'Bayesian analyses: Where to start and what to report. *The European Health Psychologist'*, 16(2), pp.75-84.

Stone, James V. (2013) *Bayes Rule: A Tutorial Introduction to Bayesian Analysis*. England: Sebtel Press.

Visa, Sofia, Anca Ralescu, Brian Ramsay, Esther Van Der Knaap (2014) *Confusion Matrix-Based Feature Selection*. proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, Ohio, USA.

Walpole, R.E. dan Myers, R. H. (1995) *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan Edisi ke - 4*. Alih bahasa oleh Sembiring, R.K. Penerbit ITB: Bandung.