Ovarian Cancer Classification using Bayesian Logistic Regression

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052049 doi:10.1088/1757-899X/546/5/052049

Ovarian Cancer Classification using Bayesian Logistic Regression

Theresia Lidya Octaviani[1], Zuherman Rustam[1*], Titin Siswantining[1]

[1]Department of Mathematics, FMIPA Universitas Indonesia, Depok 16424, Indonesia

Corresponding author's email: rustam@sci.ui.ac.id

Abstract. Cancer is one of the most common cause of death. One of the diseases that can be threaten women all over the world is ovarian cancer. Ovarian cancer is the eighth type of cancer that most women suffer from. Estimated that around 225.000 new cases are detected every year and around 140.000 people die each year from ovarian cancer. Based on WHO data, published in 2014, in Indonesia 7,6% of all cancer deaths are caused by ovarian cancer. So far there is no effective screening method for ovarian cancer. Current screening applications for high-risk women are still very controversial. There are many classification techniques has been applied for ovarian cancer prediction, for example deep learning, neuro fuzzy, neural network, and so many more. In this paper, we propose Bayesian logistic regression for ovarian cancer classification. We use data of patients suffer from ovarian cancer from RS Al-Islam Bandung to demonstrate the method. The accuracy expectation in this paper around 70%.

1. Introduction

43.5% of world's population are women. Cancer is one of the highest cause of death in high and middle income country [1]. The fact that more than 70% women with ovarian cancer diagnosed at an advance stage is one of the reasons of the cause high mortality rate. Survival rates for women with an advance stage are around 20% - 30% [2]. This is because lack of awareness of prevention and early effective detection. If the patient with ovarian cancer can be detected at an early stage, then the survival rates can become around 80% - 90%. Furthermore, the survival rate is due to the fact that ovarian cancer is a virulent disease. Most women with ovarian cancer live with fear [3].

There is no specific symptoms in patients with ovarian cancer. Symptoms that may occur are changing in bowel habits, significant weight loss, until massive abdominal swelling, so early detection is hard to do because of these non-specific complaints [4].

According to American Cancer Society (2015); ovarian cancer starts from the ovary which is female reproductive gland. Ovary produce ovum, then ovum proceed through fallopian tubes to uterus where ovum is fertilized and will thrive into a fetus. Ovary also produce the hormone estrogen and progesterone in women. It can be seen from the image below [5]:

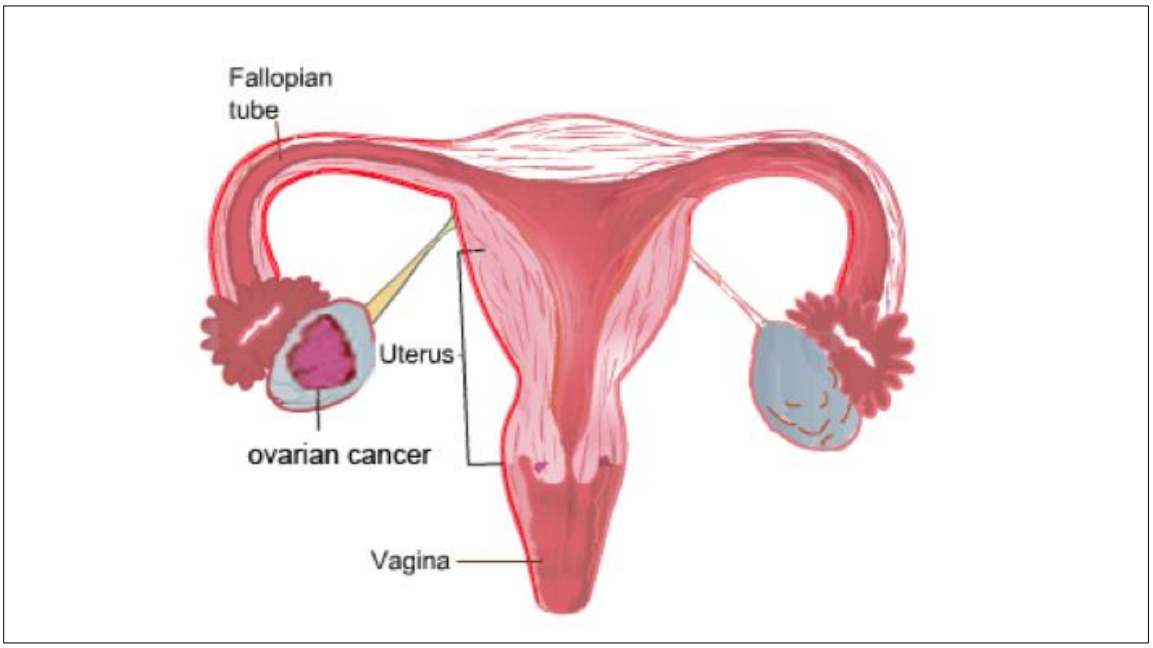

Figure 1. Ovarian cancer (http://www.ovarydisease.com/p/ovarian-cancer.html)

Some factors caused people suffer from ovarian cancer [6]:

- Age
  The older the women, the chances of getting ovarian cancer are getting bigger. Most of the women with ovarian cancer are women over 63 years old.
- Obesity
  Women with BMI (Body Mass Index) more than 30 have higher risk of suffering from ovarian cancer.
- Estrogen and Hormone Therapy
- Family History of Ovarian Cancer

There are lots of benefit introducing machine learning into medical field, that are more accurate in diagnosis, minimize costs, and reduce human resource [7]. Classification is identifying or grouping to which category the observation belongs to [8].

Dirk Timmerman, et. al. use Logistic Regression for distinguish benign and malignant of ovarian cancer. The results are sensitivity 93% sensitivity and 76% specificity [9].

N. Nunes, et. al. use IOTA Logistic Regression Model LR2 for diagnose ovarian cancer. The results are 97% sensitivity and 69% specificity [10].

We use Bayesian Logistic Regression and will be used to classify ovarian cancer dataset from RS Al Islam Bandung.

2. Method

2.1. Logistic Regression

Logistic regression model originally developed for survival analysis that usually has output (y) in form 0 or 1 (binary) [11]. Logistic regression model for a binary dependent variable is

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)} \qquad (1)$$

where

$$y = \begin{cases} 1 \text{ ; category A} \\ 0 \text{ ; category B} \end{cases}$$

$x_1, x_2, \dots, x_i$ are some i predictors.

The model above can be expressed in terms as follows:

$$E(y) = P(y = 1) = \pi \qquad (2)$$

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)} \qquad (2)$$

Equation (2) can be expressed in log-odds terms:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \qquad (3)$$

Given some random samples $(Y_j, X_{1j}, \dots, X_{ij})$, where $j = 1,2, \dots, k$ and Y_j is a result of Bernoulli experiment with probability of success as we can see in equation (3); coefficient β_j from the model is a constant that we don't know the value and it will be estimated from the data.

2.2. Likelihood Function

Likelihood function from the sample is [12]

$$L(\boldsymbol{\beta}; Y) = \prod_{j=1}^{k} \pi_j^{Y_j} \left(1 - \pi_j\right)^{1-Y_j} \qquad (4)$$

Or we can write it in another form,

$$L(\boldsymbol{\beta}; Y) = \prod_{j=1}^{k} \left[\left(\frac{e^{\beta_0 + \beta_2 x_2 + \cdots + \beta_i x_{ij}}}{1 + e^{\beta_0 + \beta_2 x_2 + \cdots + \beta_i x_{ij}}} \right)^{y_j} \left(1 - \frac{e^{\beta_0 + \beta_2 x_2 + \cdots + \beta_i x_{ij}}}{1 + e^{\beta_0 + \beta_2 x_2 + \cdots + \beta_i x_{ij}}} \right)^{1-y_j} \right] \qquad (5)$$

In logistic regression, model parameters can be determined by Maximum Likelihood Estimation (MLE) method [12],

$$\sum_{j=1}^{k} [Y_j \log(\pi_j) + (1 - Y_j) \log(1 - \pi_j)] \qquad (6)$$

But in this paper we use Bayesian to estimate the model parameters.

2.3. Bayes Theorem

The main foundation of the Bayesian method is the Bayes theorem. Bayes theorem can be stated as follows [13],

$$P(\theta|y) = \frac{P(y|\theta)\, P(\theta)}{P(y)} \qquad (7)$$

Where $P(\theta|y)$ is posterior distribution
$P(\theta)$ is prior distribution
$P(y|\theta)$ is sampling distribution or we known as likelihood function
$P(y)$ is marginal likelihood

According to [13], there are 3 types of Bayes theorem equation

$$P(\theta|y) = \frac{P(y|\theta)\, P(\theta)}{P(y)} \qquad (8)$$

$$P(\theta|y) \propto P(\theta) P(y|\theta) \qquad (9)$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood} \qquad (10)$$

2.4. Prior Distribution

Prior distribution is a distribution that gives information about the parameters. There are several types of prior distribution [14],

a. Non-informative Prior Distribution
   For the selection of the prior distribution is not based on existing data.
b. Informative Prior Distribution
   This prior distribution is based on parameter value from the prior distribution that has selected either conjugate prior or not. Parameter value from the prior distribution will affect prior distribution form which will obtained in data information that we has obtained.

2.5. Posterior Distribution

The conditional sample likelihood in (3) is combined with joint prior distribution of the parameters with bayes theorem. Recall equation (9), so the joint posterior distribution of the model parameters is [15],

$$\text{posterior} \propto \text{prior} \times \text{likelihood} \tag{11}$$

$$\text{Posterior} \propto \prod_{p=1}^{i} \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left\{-\frac{1}{2}\left(\frac{\beta_p-\mu_p}{\sigma_p}\right)^2\right\}$$

$$\times \prod_{j=1}^{k}\left[\left(\frac{e^{\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}}}{1+e^{\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}}}\right)^{y_j}\left(1-\frac{e^{\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}}}{1+e^{\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}}}\right)^{1-y_j}\right] \tag{12}$$

Where the prior is the pdf of normal distribution.

The marginal posterior distribution can be computed from the joint posterior distribution. The means of these distributions are the parameter estimates.

3. Experiments

We use Ovarian Cancer data from RS Al-Islam Bandung. It contains 203 observations, each observation consists of 5 attributes. The attributes are:

1. CA125 (U/ml)
2. Haemoglobin (g/dl)
3. Leukocytes $(10^3/\mu l)$
4. Haematocrit (%)
5. Platelets $(10^3/\mu l)$

From 203 observations, 130 observations have possibility of suffering from ovarian cancer, and the rest have not.

In this experiment, we use personal computer with i-5 processor 4 GB RAM, and software RStudio Version: 1.1.453.

4. Result and discussion

We can see the correlations among the predictors in **Fig. 2**. Correlation implies a relationship between two variables. To interpret **Fig. 2**, we can see these following values of the correlation's predictor is closest to:

- -1 indicates a strong negative correlation
- 0 indicates there is no association between the two variables
- 1 indicates a strong positive correlation

And the formula for the correlations is [16]

$$r = \frac{\mathrm{SS_{xy}}}{\sqrt{\mathrm{SS_{xx}} SS_{yy}}} \tag{13}$$

Where r is the correlations among two predictors

$\mathrm{SS}_{\mathrm{xy}} = \sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})$

$\mathrm{SS}_{\mathrm{xx}} = \sum_{j=1}^{k}(x_i - \bar{x})^2$

$\mathrm{SS}_{\mathrm{yy}} = \sum_{i=1}^{k}(y_i - \bar{y})^2$

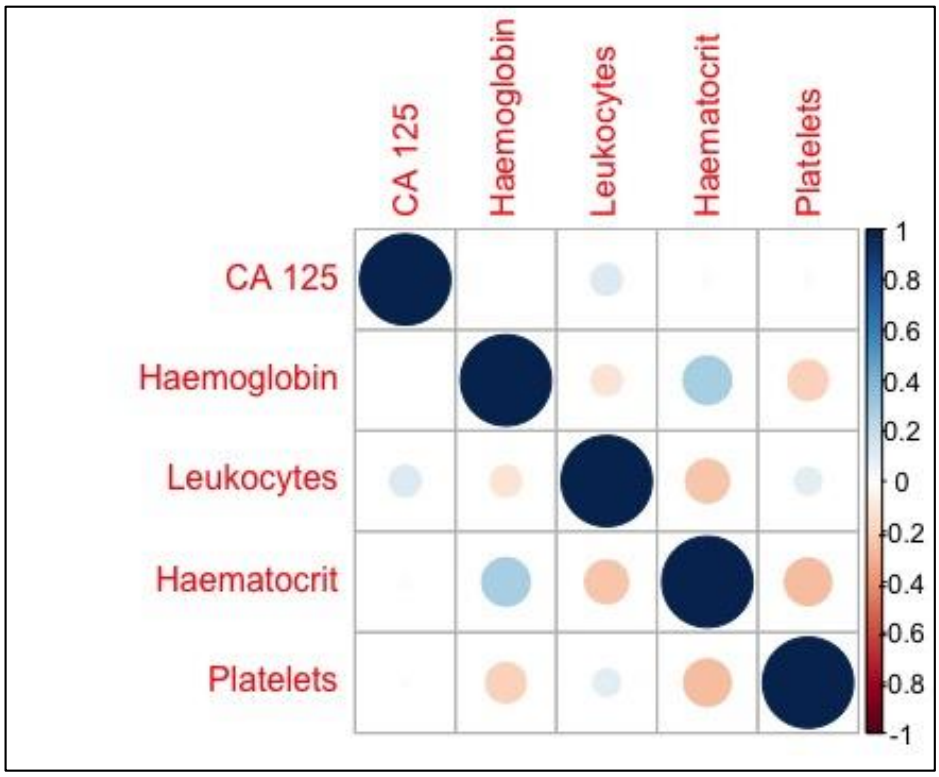

Figure 2. Correlations among the predictors

Next, we compute the model parameters estimate from the posterior interval. In this experiment we use $\mathrm{prob} = 90\%$, then we get the lower and upper posterior interval, which is 5% $\left(100\frac{\alpha}{2}\%\right)$ and 95% $\left(100\left(1-\frac{\alpha}{2}\right)\%\right)$ where $\alpha = 1 - \mathrm{prob}$, in **Table 1**. After that, we compute the posterior median estimates from the upper and lower intervals limit, in **Table 2**. The posterior median estimates is the model parameters estimates.

Table 1 Posterior interval

Variable	**5%**	**95%**
Intercept	1.57	2.77
CA 125 (U/ml)	6.97	13.53
Haemoglobin (g/dl)	-1.15	0.02
Leukocytes (10^3/µl)	-0.38	0.22
Haematocrit (%)	-0.64	0.11
Platelets (10^3/µl)	-0.09	0.51

Table 2. Model parameters estimates

Variable	Coefficient Posterior Median Estimates
Intercept	2.11
CA 125 (U/ml)	10.13
Haemoglobin (g/dl)	-0.37
Leukocytes (10^3/µl)	-0.07
Haematocrit (%)	-0.28
Platelets (10^3/µl)	0.20

There are several ways to measure performance of the method. In this paper we use accuracy, precision, recall, and F1 to measure the performance of our method [17].

Table 3. Confusion matrix

Prediction	Actual: N	Actual: P
N	TN	FN
P	FP	TP

- Accuracy
  the level of closeness between the predicted value and the actual value.
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$
- Precision
  The level of exactness between the information requested by the user and the answer given by the system.
$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$
- Recall (Sensitivity)
  The success rate of the system in rediscovering information.
$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$
- F1 score
  F1 score is one of the evaluation calculations in the retrieval information that combines recall and precision.
$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The performance result from this experiment we can see in **Table 4**.

Table 4. Confusion matrix result

Prediction	**Actual**	
	0	1
0	42	15
1	31	115

Where 0 for negative and 1 for positive.
And the result for the accuracy, precision, recall, and F1 we can see in **Table 5.**

Table 5. Result accuracy, precision, recall and F1

	Result
Accuracy	77.33%
Precision	78.76%
Recall	88.46%
F1	83.33%

From the results in **Table 4**, we can say that the accuracy of this method is 77.33%. Our method is good if we see from the precision result, because the level of exactness between the information requested by the user and the answer given by the system is 78.76%. The success rate of the system in rediscovering information is quiet good, we can see from the recall result is 88.46%. If we see the F1 score result for our method is 83.33%, it means it was good enough.

5. Conclusion

This method for the classification of ovarian cancer data can be one of the reference to help doctor in their final decision. But it's need more modification to get a better accuracy. For next, it can be used in another fields, for example in economic field.

Acknowledgement
This research was financially supported by University of Indonesia, with PITTA B 2019 research grant scheme (ID number NKB-0688/UN2.R3.1/HKP.05.00/2019).

References

[1] Lindsey A., Farhad I., et. al., "Global Cancer in Women: Burden and Trends", Cancer Epidemiol Biomarkers Prev, Vol. 26, 2017, pp. 444-457.
[2] Jose A., Thomas C., et. al., "Ovarian Cancer Screening and Early Detection in the General Population", Reviews in Obstetrics & Gynecology, Vol. 4 No. 1, 2011, pp 16-21.

[3] Francesmary M., Robert P., "Ovarian Cancer: Prevention, Detection and Treatment of the Disease and Its Recurrence. Molecular Mechanisms and Personalized Medicine Meeting Report", Int J Gynecol Cancer, Vol. 22, 2012, pp. 45-57.
[4] Alexander B., Barbara S., "Ovarian Cancer Diagnosis and Treatment", Deutsches Ärzteblatt International, Vol. 108, 2011, pp. 635-641.
[5] Vedika H. D., "Application of Machine Learning in Predicting Ovarian Cancer Survivability", Master of Science Thesis, Oklahoma State University, 2015.
[6] Liang, C., Peng, L., "An automated diagnosis system of liver disease using artificial immune and genetic algorithms", Journal of medical systems, Vol. 37(2), 2013, pp. 1-10.
[7] Mitchell T., "Machine Learning", Machine Learning. New York: McGraw- Hill, 1997.
[8] Jiliang T., Salem A., et. al., "Feature Selection for Classification: A Review"
[9] Dirk T., Antonia C. T., et. al., "Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group", Journal of Clinical Oncology, Vol. 23 (34), 2005, pp. 8794 – 8801.
[10] Natalie N., Joseph Y., et. al, "Prospective evaluation of the IOTA logistic regression model LR2 for the diagnosis of ovarian cancer", Ultrasound Obstet Gynecol, Vol. 40, 2012, pp. 355-359.
[11] William .M, Terry, .S, *A Second Course in Statistis: Regression Analysis*, 2012, pp. 494 – 505.
[12] Douglas C., Elizabeth A., et. al., *Introduction to Linear Regression Analysis,* 2012, pp. 424.
[13] http://www.hep.upenn.edu/~johnda/Papers/Bayes.pdf (accessed on 23rd February 2019)
[14] Andrew G., "Prior Distribution", Encyclopedia of Environmetrics, Vol. 3, 2002, pp. 1634 – 1637.
[15] http://web.stanford.edu/class/stats200/Lecture20.pdf (accessed on 25t February 2019)
[16] William .M, Terry, .S, *A Second Course in Statistis: Regression Analysis*, 2012, pp. 116.
[17] Alireza .B, Mostafa H., et. al., "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity, and Specifity", Emergency, Vol. 3, 2015, pp. 48-49.