JAMS (Jurnal Aplikasi Matematika dan Statistik)
Vol.1 Issue.1 December 2022, pp: 27-32
ISSN:
DOI: https://doi.org/10.53625/Jams.v1i1.3948

Logistic Regression Modeling with A Bayesian Approach to The Risk Factors of Colorectal Cancer Patients

Rinda Nariswari[1], Thomas Edisson Runkat[2], I.G.A. Anom Yudistira[3]
[1,2,3]Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: [1]rinda.nariswari@binus.ac.id

Article Info

Article history:
Received Oct 11, 2022
Revised Nov 19, 2022
Accepted Dec 20, 2022

Keywords:
Colorectal Cancer
Probability
Bayesian
Logistic Regression

ABSTRACT

This research aims to create a model, determine the factors that affect the probability of someone suffering from colorectal cancer also design a web application to do prediction using the probability logit function which has been made. There are 4 independent variables and 1 dependent variable which are used in this research. The independent variables used in this research are heredity-pedigree value, age, Adenomatous Polyposis Coli, and MutS Homolog 2. And the dependent variable used in this research is someone's status of suffering from colorectal cancer symbolized with Y. The stages of designing a web application include making use case diagrams, use case descriptions, activity diagrams, sequence diagrams, class diagrams, entity relationship diagrams, and designing the application interface. The results of this research will be a logit probability model which can be used to calculate the probability value of someone suffering from colorectal cancer. Model constructed using significant 3 independent variables which are heredity-pedigree, age, and Adenomatous *Polyposis Coli.*

Corresponding Author:
Rinda Nariswari
Statistics Department, School of Computer Science, Bina Nusantara University
Email: rinda.nariswari@binus.ac.id

1. INTRODUCTION

Cancer is a disease caused by abnormal growth of body tissue cells (growing very quickly and uncontrollably) and suppressing body tissues so that it affects body organs (Akmal, Indahaan, Sari, & Widhawati, 2010). Cancer is a disease that causes a lot of misery and death in humans. Colorectal cancer refers to malignant tumors found in the colon and rectum. The colon and rectum are part of the large intestine in the digestive system where their function is to produce energy for the body and get rid of substances that are not useful (Siregar, 2008). The number of new cases of colorectal cancer in the United States in 2021 is around 104,270 colon cancer patients and 45,230 rectal cancer patients (The American Cancer Society, 2021). The high cases of colorectal cancer are caused by almost half of the patients being diagnosed at an advanced stage, so treatment is difficult. The number of colorectal cancer patients diagnosed at an advanced stage is due to the fact that in the early stages, symptoms usually do not appear in patients.

In addition, the symptoms of colorectal cancer are often non-specific (Rahmawati, 2016). Medical research on colorectal cancer has also been done before (Sayuti & Nouva, 2019). Until now is still concluded that the etiology of colorectal cancer is still unknown. The results of the latest research conducted show that genetic factors have the greatest correlation with colorectal cancer. Research that aims to determine the factors that influence a person to suffer from certain diseases has been done before using the Logistics Regression method with the Bayesian approach that has been applied by (Octaviani, 2019). In this study, logistic regression analysis was carried out with a Bayesian approach to classify patients who have a high potential for ovarian cancer and patients who have a low potential for ovarian cancer. The study was conducted on 203 patients using 5 predictors and using the normal distribution as the prior distribution. From the results of this research, it is concluded that the Bayesian method can be an alternative to estimate parameters and still produce a good model for certain cases. From the background described above, the researcher will create a web-based application that aims to calculate the probability value of a person suffering from colorectal cancer using the opportunity model that has been generated. This application is expected to help predict a person's chances of getting colorectal cancer and make it easier for medical personnel or the wider community to be able to find out the chances of someone suffering from colorectal cancer, in order to take the necessary preventive steps. With this application, it is hoped that the number of patients with colorectal cancer can decrease in the future.

2. RESEARCH METHOD

The study was conducted from September 2020 to August 2021 and for the data collection was carried out at X Tarakan Hospital and Y Hospital Tangerang. The population used in this study were all patients registered at Hospital X Tarakan and Hospital Y Tangerang. The sampling technique used in this study is purposive sampling because the characteristics of the sample have been determined by the researcher. The criteria for the sample were patients who had no history of cancer except colorectal cancer, patients who had never undergone chemotherapy or radiotherapy, did not have inflammatory bowel disease, and patients who were willing to participate. The number of samples used in this study amounted to 22 sample data. The source of data used in this study is secondary data. The data used by the researcher was obtained from a specialist in Gastroenterology Hepatology (Sp. PD-KGEH) and is a consultant for Hepatology Gastroenterology. The data used are patient data from Hospital X Tarakan and Hospital Y South Tangerang. This study uses one dependent variable and four independent variables. The dependent variable in this study was colorectal cancer suffered by all patients at Hospital X and Hospital Y. The independent variables in this study included hereditary-pedigree value (HV), age, adenomatous polyposis coli (APC) value, and MutS value. Homolog 2 (MSH2) from all patients with colorectal cancer or patients with a family history of colorectal cancer at Hospital X and Hospital Y. The analytical technique used in this study was binary logistic regression and Bayesian for parameter estimation techniques.

3. RESULTS AND ANALYSIS

In this study, the data used were 22 patients diagnosed with colorectal cancer or patients who came from families who had colorectal cancer. The data used by the researcher was obtained from Hospital X Tarakan and Hospital Y Tangerang. In this study, variable Y is declared as the dependent or dependent variable. For more complete information, data Y has two categories, namely '1' to represent that the patient is diagnosed positively with colorectal cancer and '0' to represent that the patient is negatively indicated to have colorectal cancer. Hereditary-Pedigree Value (HV), Age, Adenomatous polyposis coli (APC) variables, and MutS homolog 2 (MSH2) were independent variables. The logistic regression used in this study is binary logistic regression because the dependent variable in this study has two possibilities, namely 1 (positive indication) and 0 (negative indication). The parameter estimation method used is the Bayes method. Where the basis of the Bayes method is a conditional probability, to estimate it requires initial information of the parameters which is also called the prior distribution. The prior distribution used in this study is the conjugate prior distribution because this study refers to model analysis in the formation of the likelihood function so that in determining the conjugate prior, we always think about determining the prior distribution pattern which has a conjugate form with the likelihood density function. In this study, several prior distributions will be used to construct the posterior distribution. The prior distributions that will be used in this study are prior normal, prior hierarchical shrinkage, prior LASSO (least absolute shrinkage and selection operator), prior product normal, and prior student t. From the prior distribution to be used, the researcher will also determine the hyperparameter values of these distributions. The hyperparameter values that will be used in this study can be seen in the following table.

Table 1 Prior Distribution and Hyperparameter Value

Prior Distribution	Prior Hyperparameter
Prior Normal	$\hat{\mu} = 0$ $\hat{\sigma} = 2.5$
Prior Hierarchical Shrinkage	$v = 4$
Prior LASSO	$v = 4$ $\hat{\mu} = 2$ $\hat{\sigma} = 0.25$
Prior Product Normal	$v = 4$ $\hat{\mu} = 0$ $\hat{\sigma} = 1$
Prior Student T	$v = 4$ $\hat{\mu} = 2$ $\hat{\sigma} = 0.25$

To estimate the regression parameters, it is necessary to form a new distribution, namely the posterior distribution. The posterior distribution can be formed from the prior distribution that has been determined and combined with the sample likelihood function. After forming the posterior distribution of the prior distribution and the likelihood function, a simulation of the posterior distribution that is formed will be carried out. The simulation method used is the Markov Chain Monte Carlo algorithm. The data simulation is carried out by generating various types of data conditions involving parameter values and sample size, namely n = 22, then it is repeated 4000 times for each combination of parameters and n. The resulting data will be used to estimate the parameters. In this study, the confidence interval used was 85%. The 85% confidence interval was calculated with the lower limit being the 7.5% quantile and the upper limit being the 92.5% quantile. Furthermore, the estimated value of each parameter is calculated by taking the average value of the resulting posterior interval. The following is the estimated parameter value generated from the posterior average using several kinds of prior distributions, which can be seen in the following table.

Table 2 Parameter Estimation Value

Prior Distribution	**Variable**	**Mean**	**Quantile**		**Significance**	**Conclusion**
			7.5%	92.5%		
Prior Normal	Y	-11.549	-27.758	2.693		
	HV	-0.972	-11.993	9.990	No	Not Affect
	Age	0.094	-0.021	0.237	No	Not Affect
	APC	0.769	-0.104	1.695	No	Not Affect
	MSH2	-0.015	-1.180	1.142	No	Not Affect
Prior Hierarchical Shrinkage	Y	-9.463	-25.039	2.858		
	HV	-1.213	-11.803	8.793	No	Not Affect
	Age	0.081	-0.019	0.208	No	Not Affect
	APC	0.596	-0.158	1.575	No	Not Affect
	MSH2	0	-1.049	1.074	No	Not Affect
Prior LASSO	Y	-18.929	-35.980	-4.246		
	HV	3.546	-9.837	14.761	No	Not Affect
	Age	0.157	0.025	0.309	Yes	Affect
	APC	0.591	-0.281	1.635	No	Not Affect
	MSH2	0.394	-0.962	1.725	No	Not Affect
Prior Product Normal	Y	-4.854	-16.585	4.579		
	HV	-2.131	-10.448	5.914	No	Not Affect
	Age	0.045	-0.025	0.127	No	Not Affect
	APC	0.333	-0.165	0.924	No	Not Affect
	MSH2	0.017	-0.524	0.631	No	Not Affect
Prior Student T	Y	-34.498	-40.451	-26.664		
	HV	10.577	0.521	20.262	Yes	Affect
	Age	0.324	0.241	0.366	Yes	Affect
	APC	-0.277	-0.535	0.095	No	Not Affect
	MSH2	1.859	1.281	2.186	Yes	Affect

Hypothesis testing on the significance level of the regression parameters was carried out with an 85% confidence interval approach for each parameter. The 85% confidence interval was calculated with the lower limit being the 7.5% quantile and the upper limit being the 92.5% quantile. The parameter is declared significant if the 85% confidence interval of the parameter does not contain a zero value. Significant parameters indicate that the independent variable has no effect on the response and insignificant parameters indicate that the independent variable has no effect on the response. Based on Table 2, the researcher can conclude that the decision-making hypothesis is as follows.

Table 3 Hypothesis Decision Making

Prior Distribution	Predictor Variable	Decision
Prior Normal	HV	Fail to Reject H_0
	Age	Fail to Reject H_0
	APC	Fail to Reject H_0
	MSH2	Fail to Reject H_0
Prior Hierarchical Shrinkage	HV	Fail to Reject H_0
	Age	Fail to Reject H_0
	APC	Fail to Reject H_0
	MSH2	Fail to Reject H_0
Prior LASSO	HV	Fail to Reject H_0
	Age	Reject H_0
	APC	Fail to Reject H_0
	MSH2	Fail to Reject H_0
Prior Product Normal	HV	Fail to Reject H_0
	Age	Fail to Reject H_0
	APC	Fail to Reject H_0
	MSH2	Fail to Reject H_0
Prior Student T	HV	Reject H_0
	Age	Reject H_0
	APC	Fail to Reject H_0
	MSH2	Reject H_0

From Table 3 it can be seen that the posterior model formed using the prior Student T distribution has three variables with significant parameter estimates. The three variables are HV, age, and MSH2. Researchers can conclude that in the posterior model formed using the prior Student T distribution, the variables HV, age, and MSH2 together have a significant effect on the status of colorectal cancer in a person. Based on Table 2, the researcher can determine the logit probability model by entering the estimated parameter values possessed into the logit probability equation which is written in the following equation.

$$\pi(x_i) = \frac{e^{(-34.498+10.577x_1+0.324x_2+1.859x_4)}}{1+e^{(-34.498+10.577x_1+0.324x_2+1.859x_4)}}$$

After writing the equation for the logit probability, the next step is to enter it into the logit transformation form in the equation below.

$$\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = -34.498 + 10.577HV + 0.324Age + 1.859\text{MSH2}$$

After forming the logit probability model, for example, the researcher will perform calculations from the logit probability model formed by entering the values of several observations in this study. The following is an example of the results of calculations carried out based on the 9th and 18th observations.

$$\pi(x_9) = \frac{e^{(-34.498+10.577\times0.5+0.324\times43+1.859\times13.295)}}{1+e^{(-34.498+10.577\times0.5+0.324\times43+1.859\times13.295)}} = 0.999923$$

The results of the calculation in equation (4.3) can be interpreted that the patient data in the 9th observation who has an HV value of 0.5, is 43 years old, and has an MSH2 value of 13,295 is at risk of suffering from colorectal cancer with a high probability value of 0.999923.

$$\pi(x_{18}) = \frac{e^{(-34.498+10.577\times0.5+0.324\times49+1.859\times7.016)}}{1+e^{(-34.498+10.577\times0.5+0.324\times49+1.859\times7.016)}} = 0.436109$$

The results of the calculation in equation (4.4) can be interpreted that the patient data in the 18th observation who has an HV value of 0.5, is 49 years old, and has an MSH2 value of 7,016 is at risk of suffering from colorectal cancer with a low probability value of 0.436109.

4. CONCLUSION

Based on the results of the research and discussion conducted, it is concluded that the logit probability model formed by the Logistics Regression method with the Bayesian approach is

$$\pi(x_i) = \frac{e^{(-34.498+10.577x_1+0.324x_2+1.859x_4)}}{1+e^{(-34.498+10.577x_1+0.324x_2+1.859x_4)}}$$

The factors that influence colorectal cancer are a person's hereditary-pedigree score (X_1), a person's age (X_2), and a person's MSH2 value (X_4). The application of the logit probability model into a web-based application that can predict a person's chances of developing colorectal cancer can be seen from the prediction results based on factors determined by the user. For further research, it can still be developed by adding other factors, adding observations, and comparing with other methods.

REFERENCES

[1] M. Akmal, Z. Indahaan, S. Sari and Widhawati, Ensiklopedi Kesehatan Untuk, Ar-ruzz Media, 2010.

[2] M. Sayuti and Nouva, "KANKER KOLOREKTAL," Fakultas Kedokteran, Universitas Malikussaleh, 2019.

[3] T. L. Octaviani, "Ovarian Cancer Classification using Bayesian," 2019.

[4] G. A. Siregar, "Deteksi Dini dan Penatalaksanaan Kanker Usus Besar," USU e-Repository, p. 39, 2008.

[5] The American Cancer Society, "American Cancer Society," 12 January 2021. [Online]. Available: https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html#:~:text=Excluding%20skin%20cancers%2C%20colorectal%20cancer,new%20cases%20of%20rectal%20cancer.

[6] S. Hamilton and L. Aaltonen, Pathology and Genetics of Tumours of the Digestive System, Lyon: IARC Press, 2000.

[7] F. Quayle and J. Lowney, "Colorectal Lymphoma. Clinics in Colon and Rectal Surgery," p. 53, 2006.

[8] R. Setianingrum, "Klasifikasi Stadium Kanker Kolorektal Menggunakan Model Recurrent Neural Network," Universitas Neggeri Yogyakarta, 2014.

[9] Y. Khosama, "Faktor Risiko Kanker Kolorektal," p. 832, 2015.

[10] D. Casciato, "Manual of Clinical Oncology. 5th Edition," Lippincott Williams and Wilkins, 2004.

[11] F. Haggard and R. Boushey, "Colorectal Cancer Epidemiology: Incidence,," Clinics in Colon and Rectal Surgery, 2009.

[12] R. L. Scheaffer, W. Mendenhall, R. L. Ott and K. G. Gerow, Elementary Survey Sampling, 7th ed., 2011.

[13] D. G. Kleinbaum and M. Klein, Logistic Regression: A Self-Learning Text, 2006.

[14] R. Hogg, E. Tanis and D. Zimmerman, Probability and Statistical Inference Ninth Edition, Pearson Education, Inc, 2015.

[15] S. Utomo, "Model Regresi Logistik untuk Menunjukkan Pengaruh Pendapatan Per Kapita, Tingkat Pendidikan, dan Status Pekerjaan Terhadap Status Gizi Masyarakat Kota Surakarta," Universitas Sebelas Maret Surakarta, 2009.

[16] A. Varamita, "Analisis Regresi Logistik dan Aplikasinya pada Penyakit Anemia untuk Ibu Hamil di Rskd Ibu dan Anak Siti Fatimah Makassar," Universitas Negeri Makassar., 2017.

[17] F. Menezes, G. Liska, M. Cirillo and M. Vivanco, "Data Classification with Binary Response Through The Boosting Algorithm and Logistic Regression," Expert Systems With Application, p. Data Classification with Binary Response Through The Boosting Algorithm and, 2017.

[18] S. Nirwana, "Regresi Logistik Multinomial dan Penerapannya dalam Menentukan Faktor yang Berpengaruh pada Pemilihan Program Studi di Jurusan Matematika UNM," Universitas Negeri Makassar, 2015.

[19] D. W. Hosmer, S. Lemesho and R. X. Sturdivant, Applied Logistic Regression Third Edition, 2013, pp. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. USA: John Wiley and Sons..

[20] A. Agresti, An Introduction to Categorical Data Analysis Third Edition, 2018.

[21] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif, dan R&D, Bandung: CV, 2013.

[22] J. W. Satzinger, R. B. Jackson and S. D. Burd, Systems Analysis and Design in a Changing World 7th Edition, Cengage Learning, 2015.

[23] Z. Rahmawati, "Klasifikasi Stadium Kanker Kolorektal Menggunakan Model Fuzzy Neural Network.," Universitas Negeri Yogyakarta, 2016.

[24] S. Krishna, P. Sheela and A. Regina, "RISK FACTORS AND ITS AWARENESS AMONGPATIENTS WITH COLORECTAL CANCER," The official journal of trained nurse' association of India, 2016.

[25] D. Collet, Modelling Survival Data in Medical Research Third Edition, 2003.

[26] Z. Soejoeti and Soebanar, "Inferensi Bayesian," Krunika Universitas Terbuka, 1988.

[27] G. Box and G.C.Tiao, Bayesian Inference in Statistical Analysis, 1973.
[28] F. Galindo-Garre and J. Vermunt, "Bayesian Posterior Estimation of Logit Parameters with Small Samples," 2004.
[29] I. Ntzoufras, Bayesian Modeling Using WinBUGS, 2008.
[30] A. F. K. Sibero, Web Programming Power Pack, 2013.
[31] Haviluddin, "Memahami Penggunaan UML (Unified Modelling Language)," Jurnal Informatika Mulawarman Vol 6 No. 1, pp. 1-15, 2011.
[32] S. Y. Sugiarti, Analisis dan Perancangan UML (Unified Modelling Language) Generated VB 6, 2013.
[33] T. M. Connolly and C. E. Begg, Database Systems: A Practical Approach to Design, Implementation, and Management Third Edition, 2002.
[34] B. Shneiderman and C. Plaisant, Designing The User Interface : Strategies for Effective Human-Computer Interaction Fifth Edition, 2010.
[35] R. S. Pressman and B. R. Maxim, Software Engineering: A Practitioner's Approach 8th Edition, 2014.
[36] W. Laurie, " Testing Overview and Black-Box Testing Techniques," 2006.
[37] F. Masitha, "DETEKSI KANKER KOLOREKTAL MENGGUNAKAN METODE GRAY LEVEL COOCCURENCE MATRIX DAN K-NEAREST NEIGHBOR BERBASIS PENGOLAHAN CITRA," Universitas Telkom Bandung, 2017.
[38] L. Torgo, Data Mining with R, 2011.
[39] A. Poudel, "A Comparative Study of Project Management System Web Applications Built on ASP. Net Core and Laravel MVC Frameworks," 2018.
[40] L. V. A. C. E. V. O. D. Y. P. C. L. Fabrice Denis, "Detection of lung cancer relapse using self-reported symptoms transmitted via an Internet Web-application: pilot study of the sentinel follow-up," 2014.