Prediction Model For Goiter Verbal Screening On Children Using Logistic Regression As Machine Learning Technique (Study in Brebes)

Maula Ismail M[*], Besral[**], Milla Herdayati[***]

[*] Medical Record and Health Information Study Program, Tasikmalaya Health Polytechnic Ministry of Health , Jl. Pemuda No 38 Cirebon, Indonesia
[**] Biostatistic and Population Studies Department Faculty of Public Health Universitas Indonesia, A Building 2nd Floor Kampus Baru UI Depok 16424, Indonesia
[***] Biostatistic and Population Studies Department Faculty of Public Health Universitas Indonesia, A Building 2nd Floor Kampus Baru UI Depok 16424, Indonesia

DOI: 10.29322/IJSRP.9.06.2019.p9080
http://dx.doi.org/10.29322/IJSRP.9.06.2019.p9080

Abstract- Goiter is a swelling of the neck due to enlargement of the thyroid gland. Beside as a body image disorders, thyroid gland disorder can result including cardiovascular disease, hypertension, stunting and impaired fertility in women. Another impact of goiter was students which has goiter experiences has lower average grade than normal students. Bulakamba, Brebes District was a region with a severe goiter categories. This study aim to make prediction model with Logistic Regression as a Machine Learning Technique that can be used to do verbal screening for Goiter in children influenced pesticide using evaluation parameters, namely Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value. Data was obtained from Dr. dr. Rasipin M.Kes research on 2011 that was located in the Bulakamba, Brebes District. Data that will be used is 53 positive and 48 negative Goiter on children. From these data, the Logistic Regression Algorithm is used by using the WEKA version 3.8.2, performance will be evaluated using the evaluation parameters. The resulting Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value respectively was 0.774, 0.708, 0.745 and 0.739. From the evaluation parameters it was found that the model can be used to do verbal screening. This model can recognize patients who are positive and negative goiter above 70%.

Index Terms- Goiter, Logistic Regression, Machine Learning, Weka

I. INTRODUCTION

Goiter (Goitre) is a swelling of the neck due to enlargement of the thyroid gland. The gland enlarges as compensation for increasing or decreasing the output of thyroid hormone. This swelling of the thyroid gland can be associated[1] with thyroid gland abnormalities called Hypothyroidism (if there is a decrease in thyroid hormone levels) or Hyperthyroidism (if there is an increase in thyroid hormone levels).

Goiter can no longer be considered a cosmetic disease, surgery is often done to remove goitre because it interferes with body image. Patient perception of it's body image is important, ashame, self-awareness and social discomfort often accompanies this perception[2]. Some avoidance behaviors are often used to suppress negative emotions and thoughts, such as avoiding visual contact, ignoring self-care needs. In the end this negative reaction can contribute to increasing social isolation. Another impact of goiter was children with goiters have an average lower value in his study than normal student, this was mentioned by Apoina Kartini in Budiyono et al. 2015[3].

Diseases due to Iodine deficiency suffer from 541 million people in the Association of South East Asian Nations (ASEAN)[4]. RISKEDAS 2013, the prevalence of hyperthiroid at the age above 15 (fifteen) years based on a doctor's diagnosis for Central Java had a prevalence of 0.5% this is above the national prevalence which is 0.4%. Goitre on Kluwut Health Center, Brebes still above 5% and there has been an increase in the discovery of goiters in 2012, 2013 and 2014, respectively 32.17%, 48.97% and 50.46%.

Brebes still using Universal Salt Iodization to prevent Goiter occurrence. But research conducted by Rasipin in Bulakamba Sub-district shows that almost certainly Goiter not because lack of iodine on child body[5]. The major risk can came from the pesticides that majorly used in Brebes. We can use this to take precautionary measures on children. So because the use of Universal Salt Iodization, for universal ways to prevent goiter and the use of palpation for IDD surveys in children in Brebes, which this method has been introduced since 1974 and still in use today opened a new space for the development of a screening method for predicting the incidence of goitre in children with a case study on the District Health Center Bulakamba Kluwut Brebes. This study aim to make good prediction model with Logistic Regression as a Machine Learning Technique that later be used to do verbal screening for Goiter in children.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

This study uses secondary data. Data is obtained from Rasipin[5] research that was conducted on same location. In this study the population is children that influenced by pesticide. The amount of data that will be used is 53 positive goiter children and 48 goiter negative children.

Variable that will be used in this study, namely the age of the child, gender, exposure to pesticides (is a composite from playing in the agricultural area, playing / coming at the agricultural drug store, engaging in agricultural activities, storing crops in the house , storing pesticides at home, spraying pesticides into crops, working parents as farm laborers or farmers, exposure to cigarette smoke), exposure to insect repellent, Body Mass Index (BMI), habit of using plastic as a food container, habit of consuming vegetables / vegetables without washing, habit of not washing hands after playing / playing from the agricultural area. In this study, we tried to compare prediction result with full variable and the reduced variable(from Chi Square). Calculation of the relationship between each variable was carried out with the dependent using Chi Square. A significant factor will be used as a variable that will be used. variable which has a p-value <0.005 will be considered as variables that will determine the outcome prediction.

This study uses the Waikato Environment for Knowledge Analysis (WEKA) version 3.2.8 for machine learning technique. Named after a flightless New Zealand bird, WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. WEKA GUI Chooser is appears on he first screen on the WEKA tool. The GUI Chooser window encompasses Simple CLI, Explorer, Experimenter, Knowledge Flow methods.

In a study conducted by Revathi[6], the study compared several Machine Learning algorithms for the heart disease data set. The study assessed the level of accuracy produced by the three algorithms by using WEKA as a research base. In another study, for predicting a disease[7], WEKA was used to see the performance of the algorithms used in Machine Learning.

The Variables selected will be included in the WEKA Application to do the training and testing process using the Logic Regression algorithm. 10-Fold Cross Validation is done on this data. Which means that the data will be separated into 10 parts and from every 10 parts taken 1 part will be used for the testing process.

Performance parameters used for measurements are Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV) and Negative Predictive Value (NPV). This study uses only one System namely Fedora Spin XFCE with Core i5-4310 specifications and 8 GB RAM.

III. WRITE DOWN YOUR STUDIES AND FINDINGS

The characteristics of each variable are shown in Table 1. In the existing data, the study was conducted on elementary school children in grade IV to VI. So that the age range obtained is not too varied. Mean for age is 10.93 years with a minimum value of 8 years and a maximum value of 13 years. In this study it was considered that the age was in the same year range so it was not taken into account in the model made.

For the variable smoking habits found that 100% of children, was non-active smokers but 79% of children are passive smokers, namely children who have active smokers in their families. This is reasonable because the research subjects are in elementary school age. Even if it is seen that in the rocket there is thiocyanate which can affect the performance of the Thyroid gland. But because the study aims for elementary school age children, the smoking habit (active smoker) is not included as a factor that has a relationship with the occurrence of Goiter.

Results obtained using Chi Square in each Variable can be seen in table 2. Nutritional status (BMI> 18.5 and BMI <= 25) was not significant to goiter occurrence, but we dicide to include this variable to predicting goiter occurance. While children playing in the agricultural area, and children having the habit of washing hands after playing in the area agriculture has no significant relationship with the occurrence of goiter. The variable is not continued to be included in the model. From this method we obtain a reduced variable(9 variable), namely

1. Parents as Farmers / Laborers of Farmers: If the wrong father / mother has a job as a farmer or farmer worker
2. Children were involved in agricultural activities: for example helping parents to produce onions, also spraying pesticides
3. Children usually visit agricultural drug stores: children are often told by their parents to buy medicine at a farm shop
4. Habit of storing crops in the house: onions that have been harvested are stored in the house, near the kitchen or place to eat
5. Habit of spraying pesticides on crops: The crops inside the house will be sprayed with pesticides to make them more durable
6. Nutritional Status (BMI <18.5 / thin): children classified as thin with a ratio of BB / TB <18.5
7. Passive smoking: There are family members who are active smokers
8. Use mosquito repellent / spray: routinely use mosquito coils / spray
9. Plastics or newspapers to wrap food: children often wrap food with plastic bags or used newspapersGoiter.

Table 1 . Data Characteristics

Variable	Frequency (n = 101)	%
Parents as Farmers / Farm Labor	31	30.7
Children play in farming area	17	16.8
Children are involved in agricultural activities	15	14.9
Children usually visit agricultural drug stores	11	10.9
Children have the habit of eating vegetables without washing	10	9.9
Children habbit not washing hands after following agricultural activities	7	6.9
The habit of storing crops in the house	23	22.8
The habit of spraying pesticide on harvest	14	13.9
Nutritional Status (BMI <18.5 / Thin)	92	91.1
Female Gender	58	57.4
Passive smokers	79	78.2
Use mosquito repellent / spray	62	61.4
Plastics or newspapers to wrap food	95	94.1

Table 2 . Variable Selection Using Chi Square

Variable		Positif (n=53)	Negatif (n=48)	p-value	OR (95%CI)
Parents as Farmers / Farm Labor	Yes No	23 30	8 40	0.009	3.7 (1.5 – 9.5)
Children are involved in agricultural activities	Yes No	14 39	1 47	0.001	17.3 (2.2 – 137.7)
Children usually visit agricultural drug stores	Yes No	11 42	0 48	0.002	Not Availlable[a]
The habit of storing crops in the house	Yes No	21 32	2 46	0.001	15.1 (3.3 – 68.9)
The habit of spraying pesticide on harvest	Yes No	14 39	0 48	0.001	Not Availlable[b]
Nutritional Status (BMI <18.5 / Thin)	Yes No	48 5	44 4	1.000*	1.1 (0.3 – 4.6)
Passive smokers	Yes No	47 6	32 16	0.015	4.0 (1.4 – 11.1)
Use mosquito repellent / spray	Yes No	42 11	20 28	0.001	5.4 (2.2 – 12.9)
Plastics or newspapers to wrap food	Yes No	53 0	42 6	0.010*	Not Availlable[c]
Children have the habit of eating vegetables without washing	Yes No	8 45	2 46	0.096*	4.1 (0.8 – 20.3)
Children habbit not washing hands after following agricultural activities	Yes No	6 47	1 47	0.115*	6.0 (0.7 – 51.8)
Female Gender	Yes No	32 21	26 22	0.668	1.3 (0.6 – 2.8)
Children play in farming area	Yes No	12 41	5 43	0.170	2.5 (0.8 – 7.8)

* Fisher Exact

[a,b,c] Risk were not available because the data came from Case-Control Study and have no children within it's variable(zero value).

From our reduced variable(9 variable) we processed on WEKA, it's Se, Sp, PPV, NPV sequentially 0.774, 0.708, 0.745 and 0.739(tabel 3). This means that with Logistic Regression, it can only correctly predict goiter children with existing variable around 77%. PPV values describe the proportion of the children that realy have a disease with all the positif result by the model. So if you get a high Se and PPV number, the prediction for Disease Positif results will be good.

To complement this study we experimenting using full variable that exist on our data(13 Variable). With WEKA the result for Se, Sp, PPV, NPV was 0.698, 0.688, 0.712, and 0,673. Although that our reduced variable not significantly higher than full variable but we can see that reduced variable have more promising result on predicting goiter.

Table 3 . Result with Logistic Regression on WEKA

	Se	Sp	PPV	NPV
Full Variable(13 Variable)	0.698	0.688	0.712	0.673
Reduced Variable(9 Variable)	0.774	0.708	0.745	0.739

Discussion

Many research for prediction of a diseases using Machine Learning has been done. Machine Learning is a mechanism for pattern recognition and building intelligence into a machine(computer) that has the ability to learn, this means that a Machine Learning will be able to do something better in the future based on data or experience (training)[8]. The main purpose of the implementation of Machine Learning for disease prediction means that developing an algorithm that has the purpose of predicting the disease so that the results are as accurate as possible.

A research that using machine learning technique to predict liver cancer form Diabetic Mellitus Type2 Patient[9], and their result was very promising. They are using multiple Machine Learning, which is Logistic Regression, Decission Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN) Back propagation. The highest results was obtained are using ANN model, with sensitivity, specificity, Negative (NPV) and positive Predictive Value (PPV) respectively 75.7%, 75.5%, 79%, and 73%.

Another research was conducted for detecting Ischemic Stroke based in EEG[10], they are using more sophisticated method on Machine Learning called 1D Convolutional Neural Network and resulting good performance. It's sensitivity average was 86.1%. They are also using Logistic Regression on their studi but it's result was not as good as the first method.

In another study on predictions of hypothyroidism[11] a more complex (Artificial Neural Network) algorithm was used to resolve the case. A chance to increase accuracy is an aim for this area of research. The use of other algorithms to classify this disease can be used as material for further research.

With our reduced variable method, our experiment shows that with Chi Square method we can slightly increase it's accuracy. This reduced Variable method was also used in other disease prediction study[12]. Although there are many method to reduce variable for machine learning technique, our chi square method was suite for our research because of our categorical data.
With variable that have been investigated by Rasipin, in this study yielded a fairly high Se and Sp value which is above 70%, then this variable can be used as a verbal screening which is quite accurate. Data that only can be obtained from Laboratory such as Iodine Urine Excretion (EIU)[13] as in the previous study are indeed very important to do and can still be used as gold standard, but this will require time and money.

Like in some district in India[14], Brebes still using Universal Salt Iodization to prevent Goiter occurrence. But research conducted by Rasipin in Bulakamba Sub-district shows that almost certainly Goiter not because lack of iodine that on a child body. The major risk can came from the pesticides that majorly used in Brebes. We can use this to take precautionary measures on children. On recent study conducted in bulakamba[15], it's resulting that pesticide is still affecting the farmer on that sub district. Although that research were not focusing on children but we can see that pesticide is an issue on that sub district.

Technology that used to predict a disease can be used as a screening tool. Screening was not a diagnostic procedure. But more intends to dispose of a large portion of the intended population of interventions by minimizing the false positive results of a population[16]. If a screening tool can identify someone who has the potential for an illness, this patient can be immediately followed up with a specific diagnosis of an illness according to the screening tool used[17].With this verbal screening model that constructed, we can predict the diseases that can affect some individual, based on that we can constructed some focused preliminary interventions to that individual. Because the purpose of screening is to discover asymptomatic, affected individuals so that they can receive appropriate treatment[18]. Other research a prediction model was implemented to a computer program[19]. This study we do not implement this model on a computer program, but for further research, this model can be implement with computer program and combined with variable suggestions that can ultimately be more focused on changing the behavior of the individual.

IV. CONCLUSION

Results of this study with reduced Variable are sequentially Se, Sp, PPV, NPV at 0.774, 0.708, 0.745 and 0.739. Based on evaluation

parameters it can be seen that this study can be used for Verbal Screening. This means that this method can identify the patient and the Positive and Negative above 70%.

ACKNOWLEDGMENT

The data used in this research was acquired from Dr. dr. Rasipin. M.Kes.

REFERENCES

[1] Kocak M, Erem C, Deger O, Topbas M, Ersoz HO, Can E. "Current prevalence of goiter determined by ultrasonography and associated risk factors in a formerly iodine-deficient area of Turkey". Endocrine. 47(1):290–8.

[2] Pandian BG, Sireesha P, Ping NY, Parashuram N. "Monitoring the prevalence of metabolic syndrome among hypothyroid patients and assessing the effect of anti-hypothyroid treatment on it among the south indian population". J Young Pharm [Internet]. 8(2):104–7.

[3] Budiyono, Suhartono, Kartini A, Hadisaputro S, P TGD, Soetadji A, et al. "Pesticide Metabolites, Anti-Thyroid Peroxidase and Thyroid Stimulating Hormone Status in School Children: A Preliminary Study in Agriculture Areas in Indonesia". Int J Sci Basic Appl Res. 22(1):1–12.

[4] Who Regional Regional Office for South-East Asia. "WHO Regional Committee for South-East Asia-Report of the Seventieth Session". Who Publ. (September):6–10.

[5] Rasipin. "Faktor-Faktor Yang Berhubungan Dengan Kejadian Goiter Pada Siswa-Siswa SD di Wilayah Pertanian (Penelitian di Kecamatan Bulakamba Kab.Brebes)" [Internet].

[6] Revathi, K.K. dan Kavita KK. "Comparison of classification techniques on heart disease data set". Int J Adv Res Comput Sci. 8(0976):276–81.

[7] Kumar N, Khatri S. "Implementing WEKA for medical data classification and early disease prediction". 2017 3rd Int Conf Comput Intell Commun Technol [Internet]. :1–6.

[8] Gollapudi S. "Practical Machine Learning" [Internet]. 1st ed. Agrawal R, Jain R, Kamoshida R, Kankanala RT, Yi DJ, editors. Birmingham B3 2PB, UK: Packt Publishing Ltd.; 468 p.

[9] Rau HH, Hsu CY, Lin YA, Atique S, Fuad A, Wei LM, et al. "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network". Comput Methods Programs Biomed [Internet]. 125:58–65.

[10] Giri EP, Fanany MI, Arymurthy AM, Wijaya SK. "Ischemic stroke identification based on EEG and EOG using ID convolutional neural network and batch normalization". 2016 Int Conf Adv Comput Sci Inf Syst ICACSIS 2016. (November 2017):484–91.

[11] Khiew K, Wang T, Lin MYS, Jiang Y. "Prediction of Hypothyroidism Disease by Data Mining Technique". 14:97–116.

[12] Liu T, Xu J, Yang C. "Machine Learning Techniques for Thyroid Cancer Diagnosis". :2012.

[13] Xiu L, Zhong G, Ma X. "Urinary iodine concentration (UIC) could be a promising biomarker for predicting goiter among school-age children: A systematic review and meta-analysis". PLoS One [Internet]. 12(3):1–11.

[14] Manjunath B, Suman G, Hemanth T, Shivaraj NS, Murthy NS. "Prevalence and Factors Associated with Goitre among 6-12-year-old Children in a Rural Area of Karnataka in South India". Biol Trace Elem Res. 169(1):22–6.

[15] Susilowati DA, Suhartono S, Widjanarko B. "Hubungan Antara Faktor Pengetahuan, Sikap Dan Perilaku Penggunaan Pestisida Dengan Kadar Serum Che Pada Petani Penyemprot: Studi di Desa Dukuhlo Kecamatan Bulakamba Kabupaten Brebes" [Internet].

[16] Segundo U, Aldámiz-Echevarría L, López-Cuadrado J, Buenestado D, Andrade F, Pérez TA, et al. "Improvement of newborn screening using a fuzzy inference system". Expert Syst Appl [Internet]. 78:301–18.

[17] Kanne SM, Carpenter LA, Warren Z. "Screening in toddlers and preschoolers at risk for autism spectrum disorder: Evaluating a novel mobile-health screening tool". Autism Res. 11(7):1038–49.

[18] Eastman CJ. "Screening for thyroid disease and iodine deficiency". Pathology. 44(2):153–9.

[19] Adyatmaka I. "Model Simulator Risiko Karies Gigi" [Internet]. Universitas Indonesia Depok Jawa Barat Indonesia;

AUTHORS

First Author – Maula Ismail Mohammad, Medical Record and Health Information Study Program, Tasikmalaya Health Polytechnic Ministry of Health , Jl. Pemuda No 38 Cirebon, Indonesia.
Second Author – Besral, Biostatistic and Population Studies Department Faculty of Public Health Universitas Indonesia, A Building 2nd Floor Kampus Baru UI Depok 16424, Indonesia.

Third Author – Milla Herdayati, Biostatistic and Population Studies Department Faculty of Public Health Universitas Indonesia, A Building 2nd Floor Kampus Baru UI Depok 16424, Indonesia.

Correspondence Author – Maula Ismail Mohammad, maula.ismail.m@gmail.com, +62 853 2811 1435.