EMITTER International Journal of Engineering Technology
Vol. 4, No. 1, June 2016
ISSN: 2443-1168

Comparison of The Data-Mining Methods in Predicting The Risk Level of Diabetes

Andri Permana Wicaksono, Tessy Badriyah, Achmad Basuki

Electronic Engineering Polytechnic Institute of Surabaya
Jl. Raya ITS Politeknik Elektronika, Kampus ITS Sukolilo, Surabaya
(031) 5947280
E-mail:andri.perman4@gmail.com, tessy.badriyah@pens.ac.id, basuki@pens.ac.id

Abstract

Mellitus Diabetes is an illness that happened in consequence of the too high glucose level in blood because the body could not release or use insulin normally. The purpose of this research is to compare the two methods in The data-mining, those are a Regression Logistic method and a Bayesian method, to predict the risk level of diabetes by web-based application and nine attributes of patients data. The data which is used in this research are 1450 patients that are taken from RSD BALUNG JEMBER, by collecting data from 26 September 2014 until 30 April 2015. This research uses performance measuring from two methods by using discrimination score with ROC curve (Receiver Operating Characteristic). On the experiment result, it showed that two methods, Regression Logistic method and Bayesian method, have different performance excess score and are good at both. From the highest accuracy measurement and ROC using the same dataset, where the excess of Bayesian has the highest accuracy with 0,91 in the score while Regression Logistic method has the highest ROC score with 0.988, meanwhile on Bayesian, the ROC is 0.964. In this research, the plus of using Bayesian is not only can use categorical but also numerical.

Keywords: diabetes, logistic regression, Bayesian.

1. INTRODUCTION

Mellitus Diabetes (MD) is an illness that happened in consequence of a high glucose level in blood because the body could not release or use insulin normally. The glucose level in human body per day is various, it is increasing after having a meal and become normal again in two hours. In order to control a diet and therapy for this illness, there is a must thing to do to identify the early signs of diabetes on the individual. Early identification in diagnosing this illness could be done in a hospital that has the complete facility and an available internist doctor. According to *American Diabetes Association* (ADA, 2005) (1), Mellitus diabetes is an illness that causes insulin secretion disorder, insulin performance disorder, or both that are also a metabolic illness with hyperglycemic characteristic. Diabetes is an illness that causes death and rank high in the world, including Indonesia. The Chief of the center Indonesian Diabetes Union (Persadia), Professor Sidartawan Soegondo said that diabetes in Indonesia tends to increase from years to

years. Most of the diabetes patients are caused by the unhealthy lifestyle and genetical factor. Mellitus diabetes is also a perennial illness that becomes the people's health problems in Indonesia. According to Perkeni census on 2011, it is suggested that on 2020 there will be 21,3 millions of Mellitus Diabetes patients that must be managed well in order to prevent a burdening complication problem, and for every second there is one patient died because of it. (The result on Perkeni census 0n 2011) (2). So it is very important to build prediction model using risk factor from the data-mining technique for intervention that is related to the development of diabetes.

Health data could be used to make a decision supporting system and diagnose illness. One of them is to use the data-mining method that is aimed for extracting and finding a pattern from a set of valuable information. In the data-mining method, there are various learning methods that could be used to compare two methods from the data of diabetes patient, so that in health sector it could be used for predicting diabetes in Jember Municipality, every learning method has also different characteristic model.

On Health Department of Jember Municipality, the scope and handling of diabetes was only done by giving diabetes medical treatment without looking at the genetical trait and the diet of the patient. The treatment implementation was done by the Health Department of Jember Municipality when the patient is doing checked up in the public region hospital. Therefore, there was still less action of early prevention to overcome the illness by the Health Department of Jember Municipality, especially in a remote region. So the diabetes cases are continuously increases from years to years. It could be seen and inspected on a data of diabetes cases in all of Jember Municipality area in 2011 until 2013.

2. RELATED WORKS

Badriyah, T, et al [3], did the research by using Decision Tree method to solve a certain problem that is commonly used by Regression Logistic as a solution. The model result of Hematology and Biochemical (BHOM) dataset that is taken from Portsmouth Hospitals NHS from 1 January to 31 December 2001 is divided into four parts of compilation. One part of the training data is used to create a model, and then the taken model was applied on the three dataset tests. The performance of every model from two methods then is compared by using calibration (test $\chi 2$ or chi-test) and discrimination (area under ROC curve). The experiment showed that two methods had the proper result in c-index case. However, in some cases of calibration score ($\chi 2$) it got a high enough result. After doing an experiment and observing the advantage and disadvantage of every method, we can conclude that Decision Tree method could be seen as an alternative method for Regression Logistic in the data-mining sector.

Kumari, M,et al [4] did the research by applying the data-mining method in predicting diabetes by using *bayesian network* method. The knowledge finding of the medical dataset was important for an effective

medical diagnose. The aim of the data-mining was to extract knowledge from information that is saved in the dataset and creating a clear and easy description for understanding the system. Mellitus Diabetes is a chronical illness and the society's main health problem, challenging the whole world. Using a the data-mining method to help people to predict diabetes has got big popularity. In this journal, *Bayesian Network* was suggested to predict whether a person got diabetes or not. The dataset that is in used were collected from the hospital which is collected from both people with and without diabetes. We use Weka tool for the experiment and analysis. Classification Algorithm was applied on stored data from the hospital.

Xue-Hui, Meng, et al (5) did a research to compare regression Logistic Process, artificial neural networks (ANNs), and decision tree models to predict diabetes or prediabetes using general risk factor. The participants are from two communities in Guang Zhou, China; 735 patients were confirmed having diabetes or prediabetes and 752 of normal control were recruited. The standard questioner was given to get an information about demographic characteristic, family diabetes history, anthropometry measure, and lifestyle risk factor. Then we developed three model predictions by using 12 inputs variable and one output variable from questioner information; we evaluated three models for their accuracy, sensitivity and specificity. Regression Logistic Model reached 76,13% of accuracy classification with 79,59% in sensitivity and 72,74% in specificity. ANN model reached 73,23 % of accuracy classification with 82,18 % in sensitivity and 64,49 % in specificity; and decision tree (C5.0) reached 77,87 % of accuracy classification with sensitivity from 80,68 % and 75,13 % in specificity. Decision tree models (C5.0) has the best accuracy classification, followed by regression logistic model, and ANN gives the lowest accuracy.

Tapak, L, et al [6] this research was to compare two traditional methods (logistic regression and Fisher linear discriminant analysis) and the four classifications of learning-machine (neural networks, support vector machines, fuzzy c-mean, and random forests), in order to classify whether the patient got diabetes or not. This research showed that on the sensitivity, the specificity, and the whole accuracy classification, supporter of vector machine model on the first level of all classification are tested in a diabetes prediction. Therefore, this approach was a promising classifier to predict diabetes and it needs an advanced research for predicting the other illnesses.

Vijiyarani, S, et al [7] did the data-mining research for sorting a very big number of data for useful information. Some important and popular the data-mining methods are rules of the association, classification, group, prediction and sequential system. The the data-mining method is used for various applications. In health industry treatment, the data-mining played important roles to predict illness. To detect numbers of illness, a test should be done by the patient. But using mining method, the test numbers should be decreased. Decreasing the test played important roles for timing and performance. This method has a surplus and the lack. The report of this research analyzed how

the data-mining method is used for predicting various illnesses. This paper mainly concentrated on the discussion of predicting heart, diabetes, and breast cancer.

Maroco J, et al [8] did research to modernize the hypothesis in new statistic classification method from the data-mining and *machine* learning methods like *Neural Networks, Support Vector Machines* dan *Random Forests* that could increase accuracy, sensitivity, and prediction specificity which is obtained from the neuropsychological test. The seven non-parametic classification was derived from the data-mining method (*Multilayer Perceptrons Neural Networks*, *Radial Basis Function Neural Networks, Support Vector Machines, CHART, CHAID* dan *QUEST Classification Trees* dan *Random Forests*) that was compared to three traditional classification (Analysis of *Linear Discriminant*, Analysis of Quadratic Discriminant and Regression Logistic) in aspect of accuracy to the overall classification, specificity, sensitivity that is located under ROC curve and Press'Q. Model predictor on the 10 tests of neuropsychology today is used in dementia diagnosis. Classification parameter statistic distribution was taken from a 5 times *cross*-validation and then was compared by using non-parametrical by Friedman.

3. ORIGINALITY

From some relatively small scale researches, it has considered that using the data-mining method to build the appropriate prediction model for diabetes and still use the available tools. From some the used models, in regression logistic method, there still is an accuracy value and ROC value under 90%. And the Bayesian method still doesn't fully use a numerical data with the same attribute yet. In this research, we proposed comprehensive comparison by using Logistic Regression method and Bayesian method with a web-based application and by using some attributes and data that have not been researched before that is taken from RSD Balung JEMBER. Determining factors of the diabetes patients are from the unhealthy lifestyle and the genetical factor. According to the medic, the most influence factor on diabetes is from a personal data and a blood check-up. Some parameters from the personal data and the blood check-up are a sexuality, an age, a *Hemoglo*in data (g/dl), a *white cell count data* (10^3/ul), a glucose data in time(mg/dl), a *creatin serum* data(mg/dl), an urea(mg/dl), a *Cholesterol total data*(gr/dl), and a *Triglyceride* data (gr/dl). Therefore, the attributes that is used in this research are attributes from internal factors of the diabetes patients. After knowing the prediction and the risk level of diabetes, the next step is analyzing some of the best methods from the compared method.

This system is built to be able to be used on the data-mining technique and collecting valuable information from some collected information on diabetes that is taken from RSD Balung Jember, and to predict how high the risk level of diabetes patients so that it is useful for the nurse or doctor in RSD Balung Jember and also for science.

4. SYSTEM DESIGN

In this session, we will explain the whole phases of this research by predicting diabetes using Logistic Regression and Bayesian method. Our research data was taken from RSD BALUNG JEMBER by collecting data starting from 26 September 2014 until 30 April 2015, then we analyzed the data for a data processing in the web-based application that we made. Those data would be calculated by using Logistic Regression and Bayesian method. The computation result of Logistic Regression and Bayesian method then is used for analysis to predict diabetes.

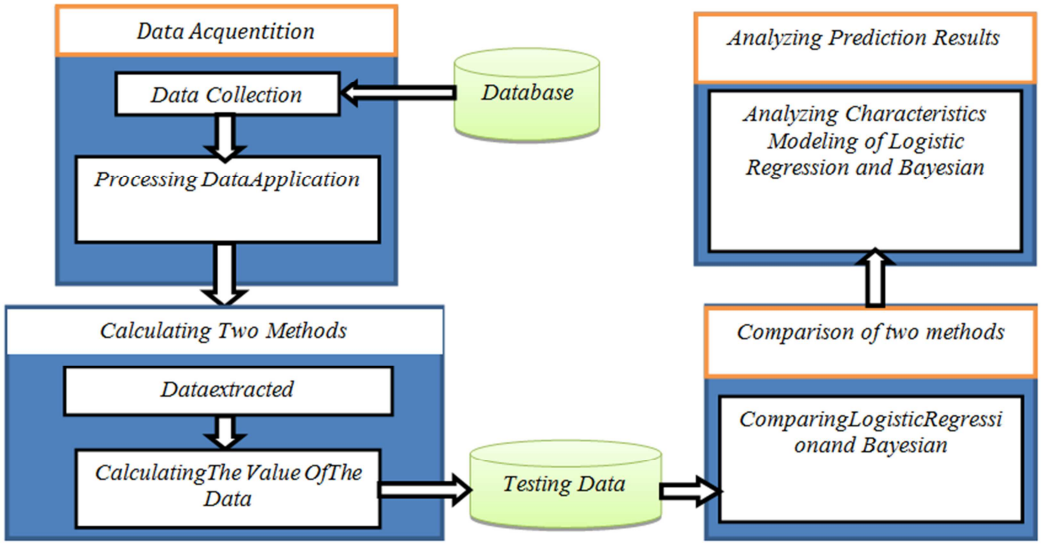

Figure 1. System Design

4.1 Original Parameter Data

This research uses patients data which is obtained from RSD Balung Jember then from those data would be displayed the historical data of patients who are diabetes affected or unaffected. The below table contains data that is used in this experiment, including: Gender, Age, *Hemoglobin* data (g/dl), *White Cell Count* data (10^3/ul), When blood sugar (mg/dl), *Creatin serum* data (mg/dl), Urea(mg/dl), *Cholesterol total* data (gr/dl), *Triglyiceride* data (gr/dl), and for diabetes attribute used *input variable*.

Table 1. Original Sample of Data Patients

Gender	Age	Hemoglobin	White Cell Count	When sugar blood	Creatin Serum	Urea	Cholesterol Total	Triglyceride	Diabetes
1	62	7.9	12.6	311	1.2	63	252	148	1
0	68	12.4	11.8	154	0.2	25	171	118	0
1	87	9.9	1.1	185	0.2	25	197	114	0
0	46	13.5	17.1	290	1.2	56	112	163	1
0	66	10.2	7.9	355	0.2	48	262	109	1

4.2 *Logistic Regression*.

Logistic Regression is a common method for creating probability prediction model of an incident such as a linear regression. The logistic Regression is used only if the output variable of a used model is defined as a binary category [8]. The difference of this logistic regression method is the prediction of a dependent variable in dichotomy scales. It means that dichotomy scale is a nominal scale that has two categories, such as Yes or No, or High and Low. In equation formula, $\mathbf{Pb_j}$ is the predicted probability by encoding it as 1, and $\mathbf{(1- Pb_j)}$ is predicted probability by another decision and is encoded as 0.

$$log\left(\frac{Pb_j}{1-\ Pb_j}\right) = \ \alpha + \ \beta_1.X_{1j} + \beta_2.X_{2j} + \cdots + \beta_n.X_{nj}$$

Notation in Formula Logistics, wherein:

- α isthe Intercept,
- X_{1j}…X_{nj} are indepedent attributes in the record –j,
- β_1…β_n are slopes for independent attributes,
- $n$is the number of independent attributes,
- $j$is the number of records in the dataset.

On the characteristic of ROC curve is to measure the modeling that is designed for determining the probability. On ROC curve there are two axises, the axis Y is called as True Positive and axis X is called as False Positive. For calculating the probability, we use AUROC (Area Under the ROC Curve) by calculating the predicted data and the result of the data that is in prediction. AUROC has scored between 0.0 until 1,0 because AUROC score would be more powerful for classification.

4.3 Bayesian

Naive Bayes is a classification by using probability method and statistic that is suggested by a scientist from England, Thomas Bayes, according to his experience in the previous period, he could predict an opportunity in the future so that finally he was known for the Bayes theory [4]. *Bayesian* theory is a probability condition theory that calculates the probability where the incident (hypothesis) depended on another incident (hypothesis).

The Statement of *Bayes* theory:

The following data would explain Bayes method in using numerical data type.

Table 2. Adult dataset that contains the numeric data type

The Numeric Adult data With Summary statistic										
	BI-RADS		Age		Shape		Margin		Density	
	malignant	benign	malignant	benign	malignant	benign	malignant	benign	malignant	benign
	5	5	34	66	4	4	5	4	2	4
	4	4	42	36	1	3	1	1	1	2
	6	6	40	41	3	2	3	1	4	3
	2	2	59	55	1	1	1	3	3	1
	0	0	69	58	4	4	5	4	3	3
	3	3	42	51	2	1	1	1	3	4
Mean	3.33	3.33	47.67	51.17	2.50	2.50	2.67	2.33	2.67	2.83
Stdev	2.16	2.16	13.37	11.09	1.38	1.38	1.97	1.51	1.03	1.17

A New Data				
BI-RADS	Age	Shape	Margin	Density
1	53	2	1	4

Assumed here, the numerical data type was distributed normally. For example X_1, X_2, ..., X_{11} are numerical data type on dataset, so the formula used are the following :

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$f(w) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(w-\mu)^2}{\sigma^2}}$$

If the new data that would be predicted for the score from the target attribute were (*Bi-Rads=1, Age=53, Shape=2, Margin=1, Density=4*)

$$f(BI-RADS = 1 \mid malignant) = \frac{1}{\sqrt{2\pi(2.16)}} e^{-\frac{(1-3.33)^2}{2.16^2}} = 0{,}103$$

So the attribute probability (c*lass=malignant*):
0,103 x 0,028 x 0,271 x 0,142 x 0,168 = 0,0000183
and attribute probability (*class=benign*):
0,103 x 0,035 x 0,271 x 0,179 x 0,207 = 0,0000367
Because the probability score (*class=benign*) > probability (c*lass=malignant*), so the prediction score of *Class* attribute for the new data (*Bi-Rads=1, Age=53, Shape=2, Margin=1, Density=4*) is '*benign* '.

5. EXPERIMENT AND ANALYSIS

On the experiment in this research, there are 1450 patients data which is affected and unaffected by diabetes, those data is taken from RSUD BALUNG JEMBER, by collecting data started from 26 September 2014 until 30 April 2015. This research used some attributes as following: Gender, Age, *Hemoglobin* data (g/dl), *White Cell Count* data (10^3/ul), When blood sugar (mg/dl), *Creatin serum* data (mg/dl), Urea(mg/dl), *Cholesterol total* data (gr/dl), *Triglyceride* data (gr/dl), and for diabetes attribute used *input variable*.

To compare the difference betwen the two models by using the curve of the receiver operating characteristic (ROC) that resulted from the real data result and prediction result. For the sensitivity, specificity, and accuracy is calculated according to the following formula:

- Sensitivity = TP / (TP+FN)
- Spesificity = TN / (TN+FP)
- Accuracy = (TP+TN) / (TP+FP+TN+FN)

information:

TP = True Positive
FP = False Postive
TN = True Negative
FN = False Negative

5.1 Graph of Patient Attribute Data

On this below graphic, it shows how much the influence data among the attribute tables data of a patient that is used to predict diabetes.

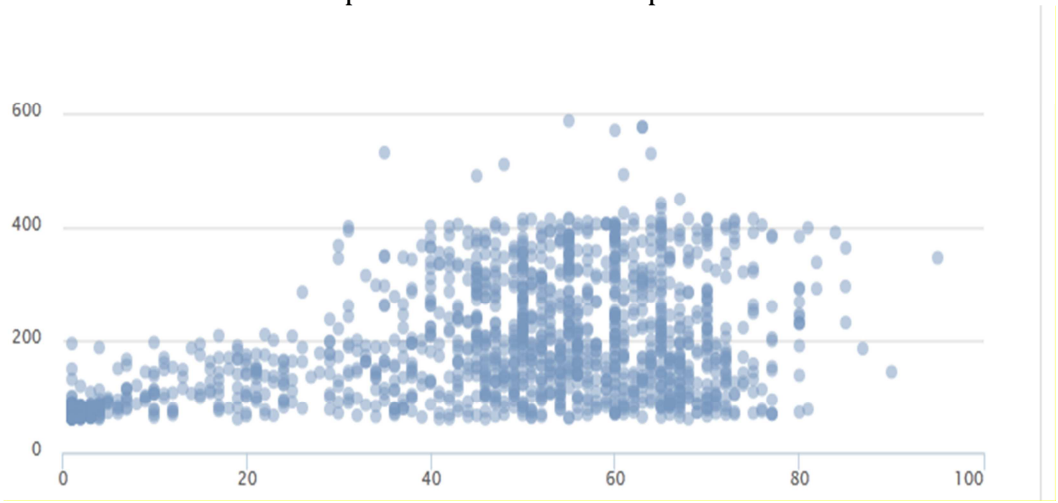

Figure 2. Patient Data Graph

On the experiment on figure 2 above, I choose the age for x-axis and the glucose for y-axis by using patient data on diabetes, so that the above pattern figures shows that the most dots was found on 40 years old until 80 years old.

From the analyzed result above, it showed that the patients that are prone to diabetes infection are average among 40 years old above.

5.2 Logistic Regression Method

Logistik RegressionMethod is one method to analyze multivariate that is useful for predicting base on the *independent variable*.

a. The Concept of Logistic Regression

From table 3 and table 4, it is seen that Risk attribute shows the level risks of diabetes that are obtained from the calculation of *logistic regression* method with the calculation of the predicted attribute is equal to 0,0, so that the risk level is classified as healthy, if equal to 0,5 then the risk level is classified as low, if equal than 0,7 so the level risk is on average, and if equal to 1 then the level risk is high.

Table 3. Predicting and Determining The Risk Level by using The Logistic Regression Method.

Gender	*Age*	*Hemoglobin*	*White Cell Count*	*When blood sugar*	*Creatin Serum*	*Urea*	*Cholesterol Total*	*Triglyceride*	*Predicted*	*Observed*	*Risk*
Female	*50*	*8.6*	*15.2*	*268*	*1.2*	*61*	*155*	*80*	*1.0*	*1*	*high*
Female	*51*	*15.4*	*17.3*	*300*	*1.2*	*32*	*285*	*197*	*1.0*	*1*	*high*
Female	*49*	*14.3*	*9.8*	*303*	*1.2*	*29*	*208*	*180*	*1.0*	*0*	*high*
Female	*47*	*14.6*	*3.5*	*315*	*1.2*	*42*	*215*	*122*	*1.0*	*0*	*high*
Male	*49*	*12.5*	*7.1*	*226*	*1.2*	*48*	*148*	*82*	*0.9*	*1*	*moderate*
Male	*68*	*13.4*	*10.6*	*213*	*0.2*	*42*	*275*	*221*	*0.8*	*1*	*moderate*

Table 4. Conversion of Computing from prediction to the risk level by using the Logistic Regression Method

Modelling
1. Computing prediction $log\left(\frac{Pb_j}{1-Pb_j}\right) = \alpha + \beta_1.X_{1j} + \beta_2.X_{2j} + \cdots + \beta_n.X_{nj}$ 2. Determining the level risk from computing the level risk If the score computingof the level risk. >=0,0 so that healthy, >=0,5 so that low, >=0,7 so that on average, and >= 1 so that high.

To describe how the calculation of *ROC* by using *logistic regression method*, this experiment has 1450 data, where the score on the class attribute is predicted attribute and the result are monitored attribute. On the discredit score with a number prediction from 0,0 until 1,0, a binary result with 0 and 1 as the scores.

Table 5. Sample of datasets to show the area under ROC curve

Record	Predicted	Observed
21	0.9	1
22	1.0	1
23	1.0	1
24	0.8	1
25	0.9	1
26	1.0	1

From table 5, for every result, we can choose score N where prediction score is "1". If the score is more than or equal to N (N is a score from various attribute score, then it would result in a new table that is base on table 5 to calculate: *False Positive, False Negative, True Positive, True Negative, sensitivity* and *(1-spesificity)* like is shown in table 6.

Table 6. Dot set in sensitivity and 1-spesificity to create ROC curve

	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity	1-Specificity		ROC
0.0	545	905	0	0	1	0	1		0.861
0.1	537	120	785	8	0.985	0.867	0.133		0.073
0.2	534	53	852	11	0.98	0.941	0.059		0.023
0.3	531	32	873	14	0.974	0.965	0.035		0.012
0.4	529	21	884	16	0.971	0.977	0.023		0.008
0.5	528	14	891	17	0.969	0.985	0.015		0.002
0.6	525	12	893	20	0.963	0.987	0.013		0.003
0.7	522	9	896	23	0.958	0.99	0.01		0.004
0.8	510	5	900	35	0.936	0.994	0.006		0.002
0.9	491	4	901	54	0.901	0.996	0.004		0.002
1.0	452	2	903	93	0.829	0.998	0.002		0.988

Table 6 above, it is seen that ROC score that has been highlighted by a yellow shows that the data number of ROC is 0,988%, this data is so close to the perfect accuracy level, that is 100%.

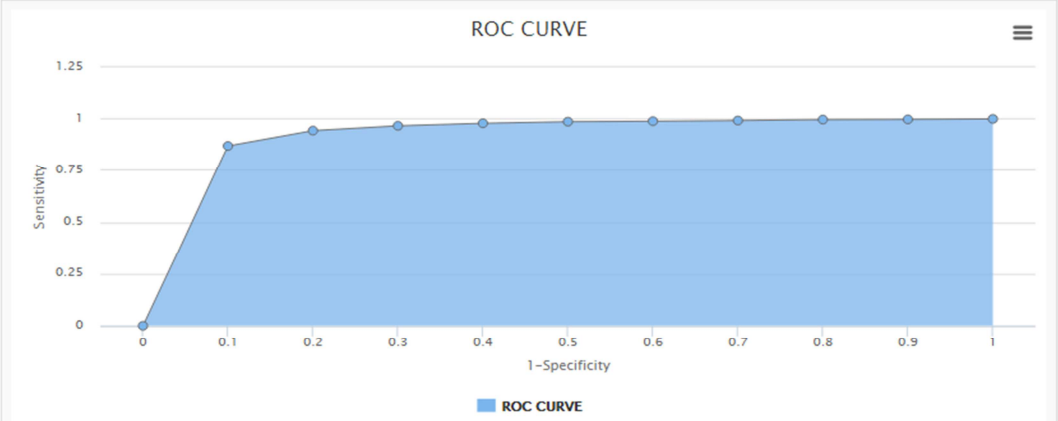

Figure 3. ROC Curve of the Logistic Regression method

On figure 3 above, it shows that ROC curve which in the every sensitivity is showing that the curve is almost reaching number 1, that is 0,988%. This curve shows a very good accuracy score.

5.3 Bayesian Method

The experiment of Bayesian method would display the ROC calculation by using Bayesian calculation method, this experiment also has same noted data as data which is used on Logistic Regression method, where the score on the class attribute is the predicted attribute and the result is the monitored attribute. This result is used to predict and determine the risk level of the diabetes illness.

Table 7. Bayesian computation

Keterangan	*Gender*	*Age*	*Hemoglobin*	*White Cell Count*	*When blood sugar*	*Creatin Serum*	*Urea*	*Cholesterol Total*	*Triglyceride*
YES	*0.4257*	*0.001*	*0.062*	*0.071*	*0.003*	*0.4183*	*0.031*	*0.0057*	*0.0069*
NO	*0.4773*	*0.002*	*0.007*	*0.008*	*0*	*0.8963*	*0.009*	*0.0073*	*0.0093*

Keterangan	*Probability / Predicted*
YES	*0.00000000000000007207320720*
NO	*0.00000000000000000000000*

Table. 7 above shows a Bayesian computation diagram where the computation predicts diabetes base on the Bayesian method, so that from the chosen attribute with a result of gender male, age 85, hemoglobin 7,3, White Cell Count 15,4, When Sugar blood 363, Creatin Serum 0,2, Urea 43, Cholesterol Total 161, Triglyceride 165 then the computation of Bayesian method of the probability result or a prediction on the Yes column of probability is printed in yellow color and No is not printed, so that the prediction result is yes, therefore the patient is classified as affected by diabetes.

Table 8. Predicting and Determining the level risk of Bayesian method

Gender	*Age*	*Hemoglobin*	*White Cell Count*	*When blood sugar*	*Creatin Serum*	*Urea*	*Cholesterol Total*	*Triglyceride*	*Diabetes*	*Predicted*
Female	*48*	*10.4*	*12.3*	*155*	*0.2*	*24*	*187*	*180*	*0*	*0*
Female	*64*	*13.9*	*5.1*	*313*	*1.2*	*43*	*128*	*58*	*1*	*1*
Female	*87*	*9.9*	*1.1*	*185*	*0.2*	*25*	*197*	*114*	*0*	*0*
Female	*62*	*7.9*	*12.6*	*311*	*1.2*	*63*	*252*	*148*	*1*	*1*
Male	*46*	*7.4*	*12.3*	*277*	*1.2*	*51*	*210*	*75*	*1*	*1*
Male	*22*	*6.1*	*17.2*	*174*	*0.2*	*41*	*116*	*79*	*1*	*1*

Yes	*No*	*Risk Computation*	*Risk*
0.0000000000000010867	*0.000000000000409704954390*	*0.02583876250*	*healthy*
0.0000000000000010113	*0.000000000000000000645242*	*0.99999361967*	*moderate*
0.000000000000000077	*0.000000000000004362467518*	*0.01724426224*	*healthy*

0.0000000000000060162	*0.000000000000000000000004*	*0.99999999993*	*moderate*
0.0000000000000075208	*0.000000000000000000014069*	*0.99999981292*	*moderate*
0.0000000000000000110	*0.000000000000000005212798*	*0.95483870106*	*moderate*

Table 9. Conversion Computation from prediction to level risk with Bayesian method

Modeling	Computing level risk	Risk Category Conversion
1. Computing prediction $f(w)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(w-\mu)^2}{\sigma^2}}$ 2. Find Prediction score from the result of Computing .Prediction If the prediction score : Yes >= No = 1. and Yes <= No = 0.	3. Find Computation level risk from prediction score. p = a / (a+b) Info. If prediction score: P =risk level computation a = Yes dan b = No.	4. Determining level risk from computing level risk F the computation score of risk level. >=0,0 so that healthy, >=0,5 so that low, >=0,7 so that average, and >= 1 so that high.

Table 8 and table 9 above show the prediction of patient data where the computation predicts diabetes base on Bayesian method. On the diabetes, attribute is the original data and on the predicted data is the data from computation result that using Bayesian method, which would get yes and no score. If "the yes" result is more than "the no", then the score is 1. And if the computation result of "the no" is more than "the yes", then the score is 0. Meanwhile, on the risk attribute computation (risk level) there is p=a/(a+b) which is "a" is for "yes" and "b" is for no. And on the risk attribute it shows the risk level of diabetes, if risk computation (risk level) equal to 0,0, then the risk level is healthy. If it is equal to 0,5, then the risk level is low. If it is equal to 0,7, then the risk level is average. And if it is equal to 1, then the risk level is high.

The next computation stage is to find the ROC score that is obtained from the Bayesian Computation method, showed on table 8.

Table 10. Dot set in *sensitivity and 1-specificity* to create ROC curve

	True Positive	False Positive	True Negative	False Negative	Sensitivity	Specificity	1-Specificity		ROC
0.0	545	905	0	0	1	0	1		0.838
0.1	533	121	784	12	0.978	0.866	0.134		0.067
0.2	533	59	846	12	0.978	0.935	0.065		0.021
0.3	532	39	866	13	0.976	0.957	0.043		0.01
0.4	528	30	875	17	0.969	0.967	0.033		0.005
0.5	525	25	880	20	0.963	0.972	0.028		0.006
0.6	525	20	885	20	0.963	0.978	0.022		0.008
0.7	524	13	892	21	0.961	0.986	0.014		0.004
0.8	521	9	896	24	0.956	0.99	0.01		0.004
0.9	515	5	900	30	0.945	0.994	0.006		0.002
1.0	505	4	901	40	0.927	0.996	0.004		0.964

On table 10 above, it is seen that ROC score is blocked in yellow which shows that the data numbers of ROC are 0,964. This data show the almost perfect accuracy.

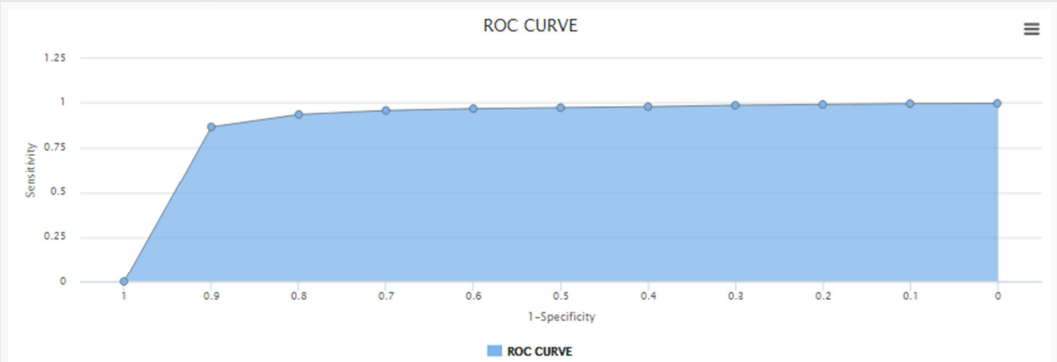

Figure 4. ROC Curve Bayesian Method

5.4 The Comparison Between The Logistic Regression and the *Bayesian Methods*

The results from the two computation result (logistic regression and Bayesian methods) will be compared for predicting the risk level of diabetes and for classification which is evaluated in an aspect of discriminative accuracy by the sensitivity (the proportion patients contagious diabetes), the specificity (The proportion patients uncontagious diabetes), the accuracy and the ROC.

Table 11. Comparison of Logistic Regression and Bayesian methods for predicting the risk level of diabetes

Gender	Age	Diabetes	Logistik Regression Method		Bayesian Method	
			Predicted	Risk	Predicted	Risk
Male	47	0	0.3	healthy	0	Healthy
Male	43	1	0.6	low	0	Healthy
Male	44	1	0.7	moderate	1	moderate
Male	60	1	1	high	1	moderate
Female	40	0	0	healthy	0	healthy
Female	64	1	0.5	low	1	moderate
Female	49	1	0.9	moderate	1	moderate
Female	55	1	1	high	1	moderate

Table 11 above shows the result of several same attribute contents with a different modeling, it seems that the risk level of diabetes is varying from logistic regression and Bayesian methods. So, it could be known where the logistic regression method that almost precise with diabetes infected patient or the uninfected patients.

Table 12. Comparison of logistic regression and Bayesian methods for sensitivity, specificity, accuracy and ROC

	Logistik Regression Methods	Bayesian Methods
Sensitivity	0.951	0.965
Specificity	0.882	0.876
Accuracy	0.908	0.910
ROC	0.988	0.964

On table 12 above, it seems that almost all of the algorithm give a result with more sensitivity (more than 90%). However, the specificity (LR, 0,882; Bayesian 0,876) is the lowest. From the experiment that we got, the highest sensitivity of Bayesian method is (0,965). The whole discriminative ability on table 12 is represented by ROC scores that on logistic regression method (0,988)

From all of the computation process, it has more than 90% in accuracy. The highest accuracy is reached by Bayesian method. Therefore, the Bayesian approach appears to be better than the logistic regression method.

6 CONCLUSION

This research discussed the comparison of the data-mining technique by using two different methods, those are Logistic Regression Method and Bayesian Method which is using a discrimination score with the sensitivity, specificity, accuracy, and ROC (Receiver operating characteristic) to measure the performance. The each two methods have it own plus. From the accuracy measurement result and the highest ROC with the same dataset, the plus of the Bayesian method is it has the highest accuracy with 0,91 in the score. Meanwhile, the plus of Logistic Regression is it has the highest ROC score with 0,988%, the Bayesian's ROC score come as the second with the score of 0,964%. In research, the Bayesian method also has a plus because it is not only using categorical but also numerical.

Therefore, it can be concluded that the two methods have good discrimination score to predict diabetes and each of them have it own plus. The prominent advantage of the two methods is the near perfect 100% ROC score (the Logistic Regression method) and the highest accuracy score for the Bayesian method that also could use numerical data.

REFERENCES

[1] Andriani Anik, **Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree,** Jurnal Bianglala Informatika Vol. 1 No 1 September 2013.

[2] Wuryandari Aryati, Trisnawati Depi, **Aplikasi Sistem Pakar untuk Diagnosa Penyakit Diabetes Mellitus menggunakan Metode Dhemster Shafer**. Magistra No. 85 Th. XXV September 2013 ISSN 0215-9511.

[3] Badriyah, T., Briggs, Jim S., and Prytherch, Dave R., **Decision Trees to predict death risk using collected routine data**, World Academy of Science, Engineering and Technology Vol:6 2012-02-24.

[4] Kumari, M, Vohra, R, Arora, A,**Diabetes Prediction using Bayesian Network**, International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178

[5] Xue-Hui Meng, Yi_Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu, **Comparison if three the data-mining models for predicting diabetes or prediabetes by risk factors,** Kaohsiung Journal of Medical Sciences (2013) 29, 93-99

[6] Tapak Lily, Mahjub Hossein, Hamidi Omid, Poorolajal Jalal, **Real-Data Comparison of The data-mining Methods in Prediction of Diabetes in Iran.** The Korean Society of Medical Informatics 2013.

[7] Triajianto, J, Purwananto, Y, Soelaiman, R, **The Implementation System of Fuzzy Classification Based on ant optimation colony to Diagnose Diabetes,** Jurnal Teknik POMITS vol. 2, no. 1, 2013, 2337-3539

[8] Vijiyarani, S, Sudha, S,**Prediction of Internal Ilness by Using The data-mining Method - A Survey**, International Journal of Computer Science and Information Technologies, Vol. II, Issue I, January 2013,

[9] Trisnawati, S., Widarsa, T., Suastika, K. **The Risk Factors od Mellitus Diabetes Type 2 for Patient Undertreatment in Local Government Clinic Subdistrict South Denpasar,**Public Health and Preventive Medicine Archive, Volume 1, Nomor 1, Juli 2013.

[10] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. **The data-mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificityof linear discriminant analysis, logistic regression, neural networks, support vector machines, classificationtrees and random forests**. BMC Res Notes 2011; 4:299.

[11] Kim S, Kim W, Park RW. A **comparison of intensive care unit mortality prediction models through the use of the data-mining techniques.**Healthc InformRes2011;17(4):232-43.

[12] Han, J. dan Kamber, M., ***The data-mining: Concepts and Techniques (2)****, Elsevier Inc*.2006.