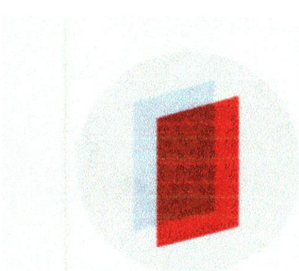
Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)

by

Journal of Physics: Conference Series

Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)



The 3rd International Conference On Science IOP Publishing
Journal of Physics: Conference Series **1341** (2019) 092009 doi:10.1088/1742-6596/1341/9/092009

Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)

D R Ente[1], S Arifin[1], Andreza[2] and S A Thamrin[1]

[1]Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, Indonesia, 90245
[2]Medical Education, Faculty of Medicine, Hasanuddin University, Makassar, Indonesia, 90245
E-mail: Dewii.ente@gmail.com

Abstract. Diabetes mellitus (DM) is one of the chronic and deadly diseases that are widely observed in various countries today. This disease continues and is increasing to a very alarming stage. Indonesia ranks fourth in the world with the highest DM after the United States, India and China. The method used in this study is data collection, variable selection, classification methods, validation and evaluation and decision making. The algorithm used in this study is C4.5 Algorithm and Naive Bayesian Method using a dataset obtained from the results of Hasanuddin University hospital medical records. The results of calculations that have been done obtained accuracy on the C4.5 algorithm of 100% and on the Bayesian naive method obtained at 90%. From these results it can be concluded that to diagnose DM disease it is recommended to use the C4.5 Algorithm.

1. Introduction

Diabetes mellitus (DM) is a chronic disease characterized by hyperglycemia and glucose intolerance that occurs because the pancreas gland cannot produce insulin adequately or because the body cannot use insulin produced effectively [1].

WHO predicts an increase in the number of people with diabetes in Indonesia from 8.4 million in 2000 to around 21.3 million in 2030. International Diabetes Federation (IDF) in 2009 also predicted an increase in the number of DM patients from 7.0 million to 12.0 million in 2030. Although there are differences in prevalence rates, the second report shows an increase in the number of people with diabetes as much as 2-3 times in 2030. This makes Indonesia ranked 4th in the world after the United States, India and China [2]. Based on the results of the 2018 Basic Health Research (Riskesdas) through examination of blood sugar, DM prevalence in Indonesia rose from 6.9% in 2013 to 8.5% in 2018. This is a large amount to be handled by Diabetes experts [3]. The high statistical figures, of course, should be anticipated by health service providers such as hospitals to prevent the explosion of diabetes patients [4].

In the field of medicine, there are many records of disease sufferers, one of which is DM. Very much data cannot be used if there is no information or conclusion from the data. Even a lot of data can actually become garbage and useless. Therefore it is necessary to do an extraction process to find information in data that has not been previously known. One method that can be used for this extraction process is machine learning.

Based on previous research, the machine learning approach can increase the risk of prediction on health output rather than the conventional approach undertaken by Selya and Anshutz [5]. Meanwhile Yusa and Sindu [6] used the C4.5 Decision Tree algorithm model for the classification of obesity. DeGrogory [7] has proven that machine learning algorithms provide a unique overview of the stages of data analysis applications in obesity. The research conducted by Farid Nurhidayat, determining the accuracy of the classification method using C4.5 based on particle swarm optimization algorithm on predictions of diabetes mellitus with the aim of getting the rule in predicting diabetes mellitus. The results of this study can be concluded that the C4.5 algorithm based on particle swarm optimization has accuracy and the AUCR value is higher than the C4.5 algorithm with the difference in accuracy value of 3.28% and the AUC value of 0.12%. Therefore, researchers feel also need to do this research by implementing data mining using the C4.5 and naïve Bayesian algorithm to diagnose DM disease

2. Material and Methods

2.1 Data Source

The data of this study used the results of medical records of patients with diabetes mellitus in Hasanuddin University Hospital Makassar City, with 127 patients. The variables of this study were gender, age, weight, height, fasting blood glucose, HDL cholesterol, LDL cholesterol, total cholesterol, triglycerides and DM status. The age interval for people with DM is around the age of 26-82 years.

2.2 Data Mining

Data mining is defined as the process of finding patterns in data. This process must be automatic or usually semi-automatic. The resulting pattern must mean that the pattern provides several advantages. The pattern is identified, validated, and used to make a prediction [8].

2.3 Classification

Classification is the process of finding a model (function) that describes and distinguishes a data class or concept that aims to be used to predict the class of objects whose label class is unknown [9]. Data classification consists of 2 steps process. The first is learning (training phase), the classification algorithm is made to analyze training data and then represented in the form of classification rules. The second process is the classification of testing data used to estimate the accuracy of the classification rules. The classification process is based on four components [10]. First, the class is a dependent variable in the form of categorical which represents the "label" contained in the object. Second, predictors are independent variables represented by data characteristics (attributes). Third, the training dataset is a data set that contains the values of the two components above which are used to determine the suitable class based on predictors. Fourth, testing the dataset contains new data that will be classified by the model that has been made and the classification accuracy evaluated. Classification is the process of finding a set of models (functions) that describe and distinguish concepts or classes of data, with the aim that the model can be used to predict the class of an object or data whose label class is unknown.

2.4 Decision Tree

Decision trees are prediction models using tree structures or hierarchical structures. Apart from being relatively fast development, the results of the models built are also easy to understand, so this Decision Tree is the most popular classification method used. Decision Tree is a flow chart like a tree structure, where each internal node shows a test on an attribute, each branch shows the results of a test and leaf node showing classes or class distribution.

2.5 C4.5 Algorithm

The basic concept of C4.5 Algorithm is to convert data into decision trees and rules. The C4.5 algorithm maps attributes to classes that can be applied to new classifications [9]. The advantages of C4.5 Algorithm are easy to understand, flexible and interesting because they can be visualized in images.In general, the C 4.5 algorithm for building decision trees is as follows [11]:

1. Select the attribute as root based on the highest gain value on each attribute. The formula calculates gain :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy(Si) \quad (1)$$

Information :

S : Set of cases
A : Attribute
n : Number of partition attributes A
$|Si|$
Number of cases on the partition to -i
$|S|$: Number of cases in S
Meanwhile to calculate the entropy value can be used with the formula:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (2)$$

Information :

S : Set of cases
A : Features
n : Number of partitions S
Pi : Proportion of S_i to S

2. Make a branch on each value
3. The cases in the branch.
4. Repeat the process for each branch until all cases in the branch have the same class.

The advantages of C4.5 Algorithm are easy to understand, flexible and interesting because they can be visualized in images.

2.6 Naive Bayesian Method

Bayes is a simple probabilistic based prediction technique based on the application of the Bayes theorem. The use of Bayes theorem on the Naive Bayes method is by combining prior probability and conditional probability in a formula that can be used to calculate the probability of each possible classification. This independence model produces the best solution. The equation of the Bayes theorem is:

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)} \quad (3)$$

Information :

X : Data with unknown classes
H : The data hypothesis is a specific class
$P(H|X)$: Probability of hypothesis H based on condition hypothesis X
$P(H)$: Probability of hypothesis H
$P(X|H)$: Probability of hypothesis X based on condition hypothesis H
$P(X)$: Probability of hypothesis X

Naïve Bayesian is a classifier with a statistical approach which can predict the probability of each class. The advantage of this Bayes grouping is that there is a high level of accuracy and speed in large data usage. Grouping Naive Bayesian assumes that the attribute values on the class label are independent of other attribute values which can facilitate the calculation [12].

3. Result

The amount of data used in this study is 127 data.

Table 1. DM Sample Tables

JK	AGE	BB	TB	GDP	HDL	LDL	KOL TOTAL	TG	RESULTS
1	35	40.5	145	84	43	93	165.6	148	Normal
0	41	38.25	145	95	44	93	163.4	132	Normal
0	52	40	148	230	41	146	231	151	DM
0	41	41.65	149	86	42	97	163.4	122	Normal
1	53	44.1	149	90	42	94	161.8	129	Normal
0	64	52	150	414	35	191	281	165	DM
0	73	50	150	54	67	1254	1077	374	DM
0	71	45	150	212	48	136	258	182	DM
0	68	65	150	229	38	162	207	118	DM
0	47	60	150	212	22	138	167	106	DM
0	53	60	150	352	32	170	205	100	DM
0	67	50	150	137	45	135	201	88	DM
0	52	55	150	150	79.8	112	261	391	DM
0	62	50	150	132	32	99	169	197	DM
0	50	43	150	420	64	169	266	166	DM
0	59	70	150	135	21	147	199	128	DM
0	38	45	150	368	56	118	203	143	DM
0	53	60	150	154	28	111	149	113	DM
0	54	79	150	225	29	174	209	247	DM
0	76	60	150	356	26	102	175	172	DM
0	68	48	150	132	34	154	221	207	DM
1	28	45	150	80	40	96	164	140	Normal
0	54	42.5	150	103	44	96	167.4	137	Normal
1	54	45.9	151	74	41	98	164.2	126	Normal
1	69	61	152	184	52	103	175	81	DM

From Table 1 the DM disease data obtained is not accompanied by a description that specifically explains the intent of each attribute obtained. This is a reference for researchers to be used as an initial step, which is analyzing the purpose of data with information retrieval. The information obtained is listed in Table 2.

Table 2. Variable datasets and descriptions

No	Variable	Scale	Information
1	Gender	Nominal	0 : women

2	Age	Ordinal	Men 1: 26 – 32 Years 2: 33 – 39 Years 3: 40 – 46 Years 4: 47 – 53 Years 5: 54 – 60 Years 6: 61 – 67 Years 7: 68 – 74 Years 8: 75 – 81 Years
3	Weight	Ratio	-
4	Height	Ratio	-
5	GDP	Ordinal	1: Low, if the level is < 70 mg/dL 2: Normal, if the level is 70 – 100 mg/dL 3: Pre DM, if the level is 101-126 mg/dL 4: Height, if the level is > 126 mg/dL
6	HDL	Ordinal	1: Normal, if the level is > 65 mg/dL 2: Low, if the level is < 65 mg/dL
7	LDL	Ordinal	1: Normal, if the level is < 110 mg/dL 2: Height, if the level is > 110mg/dL
8	Kolestrol Total	Ordinal	1: Normal, if the level is < 200 mg/dL 2: Height, if the level is > 200 mg/dL
9	Trigliserida	Ordinal	1: Normal, if the level is < 150 mg/dL 2: Height, if the level is > 150 mg/dL

3.1 Variable Selection

The number of variables in the data used in this study can cause data dimensions, which result in overfitting and underfitti[18]. To overcome this problem, variable selection is carried out. To select the variables used by testing the Chi-Square Test (χ^2). The Chi-Square test is applied to each variable, and is measured by the p-value. The most informative variables will be identified by sorting each variable based on the p-value. Determination of variable selection is by comparing the p-value with a significant level of 5%.

The hypothesis used is:

H_0 : There is no influence between Variables - n with DM disease

H_1 : There is an influence between Variables - n with DM disease

Table 3. Value of the p-value of each attribute

Variabel	P-Value	Hasil
Gender	0.7014	No effect
Age	1.01E-06	Take effect
Weight	6.52E-03	Take effect
Height	0.004644	Take effect
GDP	2.2E-16	Take effect
HDL	0.08883	No effect
LDL	6.76E-13	Take effect

Total cholesterol	8.49E-06	Take effect
Triglycerides	8.49E-06	Take effect

The results obtained from table 3 are 7 variables that influence the age, weight, height, GDP, LDL, total cholesterol and Trigleserida and 2 variables that are not influential, namely Gender and HDL.

3.2 Process Method for Classification, Validation and Model Evaluation

From the results of calculations and trials using the R-Studio version 3.5.1 application with the C4.5 Algorithm and the Naive Bayesian Method produces the accuracy with the comparison found in Table 4.

Table 4. Comparison of accuracy of C4.5 and Naive Bayesian Algorithms

Criteria	Algorithm C4.5	Naive Bayesian
Accuracy	100%	92%
Sensitivity	100%	85%
Specificity	100%	94%
AUC	100%	90%

The AUC value has a range between 50% and 100%. Interpretation of the AUC value can be classified into five different parts, namely 50% - 60% (incorrect accuracy), 61% - 70% (weak accuracy), 71% - 80% (moderate accuracy), 81% - 90% (high accuracy), and 91% - 100% (very high level of accuracy) [13].

For C4.5 Algorithm the performance of the model classification obtained criteria for Accuracy, Sensitivity, Specificity and AUC which is 100%, meaning that the model obtained is very good with a very high degree of accuracy while for the Naive Bayesian method the AUC value obtained is 90% which means the model got good and high accuracy.

4. Conclusion

From the results of this study it can be concluded that the C4.5 Algorithm has the best level of accuracy compared to using the Naive Bayes method with a difference of 10% accuracy.

Acknowledgement

Thank you for saying to the Ministry of Research, Technology and Higher Education that has funded our research

References

[1] Evi Kurniawaty Bella Y 2016. *Faktor-Faktor Yang Berhubungan Dengan Kejadian Diabetes Melitus Tipe 2*. Universitas Lampung. Jurnal Majority volume 5 nomor 2.

[2] PERKENI, 2011. *Konsensus Pengelolahan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia*. Jakarta. PB Perkeni.

[3] Riset Kesehatan Dasar (Riskesdas). 2018. Diakses pada tanggal 20 Juni 2019 melalui website http://www.depkes.go.id/article/view/18110200003/potret-sehat-indonesia-dari-riskesdas-2018.html

[4] Rodiyatul, F. S, Tama, B. A dan Mulya, M. 2010. *Pengembangan Perangkat Lunak Diagnosa Penyakit Diabetes Mellitus Tipe II Berbasis Teknik Klasifikasi Data*. Prosiding Seminar Nasional. 13-14 Desember 2010.

[5] Selya, A.S and Anshutz, D. (2018). Machine Learning for the Classification of Obesity from Dietary and Physical Activity Patterns in P. J. Giabbanelli et al. (eds.) Advanced Data Analytics in Health, Smart Innovation, Systems and Technologies 93. https://doi.org/10.1007/978-3-319-77911-9_5

[6] Yusa, M dan Sindu, W. 2015. Evaluasi Model Decision Tree C4.5 Guna Prediksi Posibilitas Resiko Obesitas. Seminar Nasional Informatika, 1(1): 147-152
[7] DeGregory, K. W., Kuiper, T. DeSilvio, J. D. Pleuss, R. Miller, J. W. Roginski, C. B. Fisher, D. Harness, S. Viswanath, S. B. Heymsfield, I. Dungan and D. M. Thomas. 2018. *A review of machine learning in obesity*. Obesity reviews, 19(5): 668-685
[8] Witten I. H., Frank E.,Hall M. A.2011. Data mining practical Machine Learning Tools and Techniques (3 rd ed). USA: Elsevier
[9] Han, Jiawei dan Kamber, Micheline. (2006). Data Mining : Concept and Techniques Second Edition, Morgan Kaufmann Publishers.
[10] Gorunescu, F. (2011). Data Mining Concepts, Model and Techniques (Vol. 12). Springer: Berlin.
[11] Adhika N., Isni O.2017. *Penerapana Algoritma klasifikasi data mining C4.5 pada dataset cuaca wilayah bekasi*. Universitas Gunadarma. Jurnal Format volume 6 Nomor 2. ISSN 2089-5615
[12] Han J., Kamber M., Pei J. 2012. Data Mining: Concepts and Techniques. Ed ke-3. Massachussets (US): Morgan Kaufmann Publishers.
[13] Hikma Widayu, Surya DN, Natalia S, Mesran.2017. *Data mining untuk memprediksi jenis transaksi nasabah pada koperasi simpan pinjam dengan algoritma C4.5*. Medan: Media Informatika Budidarma, Vol 1 No 2. ISSN: 2548-8368 hal 32-37

Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)

ORIGINALITY REPORT

%20	%11	%15	%13
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Achmad Solichin. "Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation", 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2019 Publication	%2
2	Amri Danades, Devie Pratama, Dian Anggraini, Diny Anggriani. "Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status", 2016 6th International Conference on System Engineering and Technology (ICSET), 2016 Publication	%2
3	Submitted to Universitas Atma Jaya Yogyakarta Student Paper	%1
4	Sony Puji Triasmoro, Vita Ratnasari, Agnes Tuti Rumiati. "Comparison performance between	%1

rare event weighted logistic regression and truncated regularized prior correction on modelling imbalanced welfare classification in Bali", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018
Publication

5	repositori.usu.ac.id Internet Source	%1
6	Des Suryani, Ause Labellapansa, Eka Marsela. "Chapter 95 Accuracy of Algorithm C4.5 to Study Data Mining Against Selection of Contraception", Springer Science and Business Media LLC, 2018 Publication	%1
7	e-jurnal.pelitanusantara.ac.id Internet Source	%1
8	link.springer.com Internet Source	%1
9	Submitted to Program Pascasarjana Universitas Negeri Yogyakarta Student Paper	%1
10	prism.ucalgary.ca Internet Source	%1
11	Mujiono Sadikin, Fahri Alfiandi. "Comparative Study of Classification Method on Customer	%1

	Candidate Data to Predict its Potential Risk", International Journal of Electrical and Computer Engineering (IJECE), 2018 Publication	
12	Submitted to University of Melbourne Student Paper	%1
13	ejournal3.undip.ac.id Internet Source	%1
14	Submitted to Mesa Community College Student Paper	<%1
15	Submitted to Sogang University Student Paper	<%1
16	journal.unhas.ac.id Internet Source	<%1
17	www.autospost.com Internet Source	<%1
18	slideplayer.com Internet Source	<%1
19	publikasiilmiah.ums.ac.id Internet Source	<%1
20	Reza Firsandaya Malik, Eko Pratama, Huda Ubaya, Rido Zulfahmi, Deris Stiawan, Kemahyanto Exaudi. "Object Position Estimation Using Naive Bayes Classifier Algorithm", 2018 International Conference on	<%1

	Electrical Engineering and Computer Science (ICECOS), 2018 Publication	
21	www.ncbi.nlm.nih.gov Internet Source	<%1
22	Submitted to Pasundan University Student Paper	<%1
23	Submitted to Oxford Brookes University Student Paper	<%1
24	www.prisonlegalnews.org Internet Source	<%1
25	ijcrar.com Internet Source	<%1
26	Submitted to Universitas Prima Indonesia Student Paper	<%1
27	Radius Tanone, Hendra Budi Prasetya. "Designing and Implementing an Organoleptic Test Application for Food Products Using Android Based Decision Tree Algorithm", International Journal of Interactive Mobile Technologies (iJIM), 2019 Publication	<%1
28	educationdocbox.com Internet Source	<%1
29	Submitted to Monash University Sunway	

	Campus Malaysia Sdn Bhd Student Paper	<%1
30	Tities Anggraeni Indra, Aida Lydia, Dyah Purnamasari, Siti Setiati. "Asosiasi antara Status Vitamin D 25(Oh)D dengan Albuminuria pada Pasien Diabetes Melitus Tipe 2", Jurnal Penyakit Dalam Indonesia, 2017 Publication	<%1
31	Submitted to Universitas Nasional Student Paper	<%1
32	Submitted to University of Central Lancashire Student Paper	<%1

EXCLUDE QUOTES ON
EXCLUDE BIBLIOGRAPHY ON
EXCLUDE MATCHES < 5 WORDS