ISSN: 2078-0273 Baku, Azerbaijan

Journal of HEALTH, SPORT AND TOURISM

Sağlamlıq, idman və turizm jurnalı

Vol. 5. No. 1 / 2014

JHST

"SAĞLAMLIQ, İDMAN VƏ TURİZM" jurnalı
ISSN: 2078-0273

JOURNAL OF HEALTH, SPORT AND TOURISM

Vol. 5. No. 1, May, 2014

Member of crossref

DOI: 10.7813/jhst.2014

Indexed by:

"UlrichsWeb", "Academic Search Complete-EBSCO", "Genamics"
"IndexCopernicus International", "Norwegian Social Science Data Services",

"PROGRESS" IPS LLC
Baku-2014

JOURNAL OF HEALTH, SPORT AND TOURISM

JHST, Vol. 5. No. 1, May, 2014

DOI for issue: 10.7813/jhst.2014/5-1

ISSN: 2078-0273
National reg. No. **3818**

"Sağlamlıq, İdman və Turizm" jurnalı Azərbaycan Respublikası Ədliyyə Nazirliyi tərəfindən rəsmi Dövlət Qeydiyyatına alınıb (**№ 3818**). "Sağlamlıq, İdman və Turizm" jurnalı müstəqil şəkildə Ümumdünya ISSN Mərkəzində (Paris, Fransa) qeydiyyatdan keçərək mətbu orqan kimi **ISSN: 2078-0273** nömrəsi ilə beynəlxalq nəşr statusu qazanıb.

"Sağlamlıq, İdman və Turizm" jurnalı "Norwegian Social Science Data Services", "UlrichsWeb", "Academic Search Complete", "Education Research Complete" (EBSCO), "IndexCopernicus International", Genamics və s. elmi indeksləşdirmə bazalarına daxil edilib.

"Sağlamlıq, İdman və Turizm" jurnalı ildə 2 dəfə-May və Dekabr aylarında dərc olunur. Bütün hüquqları qorunur. Redaksiyanın yazılı icazəsi olmadan materialların təkrar nəşri, tərcümə edilərək yayılması qadağandır. Məqalələr bir qayda olaraq Beynəlxalq Redaksiya Heyətinin yekun qərarı ilə dərc olunur. Məqalələr unikal DOI ilə nömrələnir.

About the journal

The Journal of Health, Sport and Tourism is published two times per year in May and December. Journal publishes original papers in health, sport and tourism in general, but giving a preference to those in the areas of health, sport and tourism represented by the editorial board. All submitted papers are considered subject to the undersanding that they have not been published and are not being considered for publication elsewhere. To be publishable, papers must treat new research, be well written, and be of interest to a significant segment of the science community. Papers may be written in English, French, Spanish or Russian. All articles are subjected to a double blind peer-review process. Manuscripts (and minireviews) are invited from academicians, researchers and practitioners.

Ünvan: Bakı şəh. İ. Qutqaşınlı küç., 97/2
Veb: www.ijar.lit.az/medicine.php
E-mail: edit@ijar.lit.az, subijar@gmail.com
Tel.: +994 12 5100280

Növbəti buraxılış tarixi: 30.12.2014
Materiallar toplanıb: 15.12.2013-20.05.2014. Çapa imzalanıb: 10.06.2014.
Format: 62x84 $^1/_8$. Şrift: Arial. Səhifələrin sayı: 54. f.ç.v. 7. Tiraj: 100

TABLE OF CONTENTS

R. Hiola, Bambang W. Otok. Ensemble method of classification tree on diabetes mellitus patient Dr. M.M. Dunda Hospital of Gorontalo. Journal of Health Sport and Tourism 2014; 5(1), 43-51. ***DOI:*** *10.7813/jhst.2014/5-1/7*

ENSEMBLE METHOD OF CLASSIFICATION TREE ON DIABETES MELLITUS PATIENT DR. M.M. DUNDA HOSPITAL OF GORONTALO

Rama Hiola[1], Bambang Widjanarko Otok[2]

[1]Faculty Science Health and Sportmanship, University Country of Gorontalo, Gorontalo,
[2]Laboratory of Environmental and Health Statistic, Sepuluh Nopember Institute of Technology (ITS), Surabaya **(INDONESIA)**
E-mails: hiola.rama@gmail.com, bambang_wo@statistika.its.ac.id

DOI: 10.7813/jhst.2014/5-1/7

Received: 15 Jan, 2014
Accepted: 16 May, 2014

ABSTRACT

Health is one of factor that contribute an important role in realizing qualify human resources. Diabetes Mellitus (DM) is one of factor of health problem in which affected on productivity is not just influence individually, but health system of country. The risk factors influence on of occurrence DM are heredity factor, diet, obesity, lack of sport activity, age and gender. The objective of this research is to determine accuracy of grouping on patient of occurrence of diabetes mellitus based on risk factors. The result of this research showed that classification trees obtained accuracy of grouping of occurrence of diabetes mellitus based on risk factors is 94,8 percent. Heredity factor, diet, lack of sport activity and obesity is most influence factors in the determination of occurrence of diabetes mellitus. The development of classification tree through bootstrap ensemble was able to increase accuracy of classification tree up to 2,96 percent to be 97,80 percent compare with classification tree method. Flexibility of the bootstrap ensemble of classification tree approach can increase accuracy of classification up to 3,3 percent and 6,4 percent compare with logistic regression and discriminant analysis method.

Key words: Accuray of classification, classification tree, *diabetes Melitus*, ensemble

1. INTRODUCTION

Health problems is one factor that contribute an important role in realizing qualify human resources. By establishment in health, is expected to further increase the level of public health and health services can be felt by all levels of society adequately. But on the other hand much nutritional excessive problem with all the consequences begin to show the trend of increased prevalence of degenerative diseases such as diabetes mellitus, hypertension, coronary heart disease, stroke, peripheral vascular disease and other heart diseases (Depkes RI, 2010). According to the International Federation of Diabetes Mellitus (FDI) (2010), showed that 285 million people who suffer diabetes afflict more on young people, and about 50 percent of them aged between 20-60 years. The data above indicates that 10 (ten) largest state of diabetes mellitus are India, China, United States, and Indonesia. Ministry of Health of the Republic of Indonesia (2010) in the results of basic health research (RISKESDAS) showed that the prevalence of non-communicable disease incidence of 5.7 percent suffer from diabetes mellitus. While the region has suffered a high prevalence of diabetes mellitus among NAD (8.5%), Riau (10.4%), the Pacific Islands (8.6%), Central Java (7.8%), West Kalimantan (11.1%), North Sulawesi (8.1%), Gorontalo (7.7 %) and North Maluku (11.1%).

Diabetes Mellitus (DM) is one of factor of health problem in which affected on productivity is not just influence individually, but health system of country. Some of the risk factors that affect the incidence of diabetes is a gene factor, diet, obesity, lack of sport activity, age, gender, smoking, lack of sleep, frequent stress, mostly snacking, a history of delivering a baby > 4 kg, history of hypertension, history of cardiovascular disease, TGT history, history of cholesterol, an avid soda, using the contraceptive pill, lifestyle, due to drugs / chemicals and less education counseling.

The data of occurrence of diabetes mellitus showed that cases of diabetes mellitus over the years there has been a considerable increase in the number of patients significantly to the treatment of cases of diabetes mellitus in the General Dr. M.M Dunda Hospital Limboto. In year 2008, 146 people, or 13.75 % of 1062 patient are persons with diabetes mellitus, in year 2009 there are 113 people, or 10.12% of 1116 patient are persons with diabetes mellitus and in year 2010 amounted to 148 people, or 12.89 % of the 1148 patient are persons with diabetes mellitus. To deal this very complex diabetes mellitus desease, the hospitals, especially in the area of Gorontalo, are found not to have a clear picture to make a conclusion or decision making strategic moves in an effort to control the occurrence of acute complications for people with diabetes is increasingly overflowing well in private and government hospitals. This issue would require scientific study to determine the incidence of diabetes mellitus classification accuracy over the ensemble of classification tree method, which in turn is expected to be a material consideration in determining policies to reduce the incidence of diabetes mellitus.

2. METODOLOGY

Data is obtained from Dr. M. M. Dunda General Hospitals Limboto. Sampling is done to 560 patient on diabetes mellitus cases (DM type I and Type II). Then the data is divided into 2 (two) part, namely learning data for model verivication (get the best model) and testing data for validation of model. Division of data is unlimited so was done a simulation with subdivision ([95%, 5%], [90%, 10%], [85%, 15%], [80%, 20%], [75%, 25%]). On data learning was done modeling cart with a response variable is diabetes mellitus (0 = DM type I, 1 = DM type II), with predictor variable is (X1 = A factor of the acts of heredity, X2 = diet, X3 = obesity, X4 = lacking activity sports, X5 = age, X6 = gender), because response variable is categoryc hence used classification tree model.

Conceptual of research are served as follows:

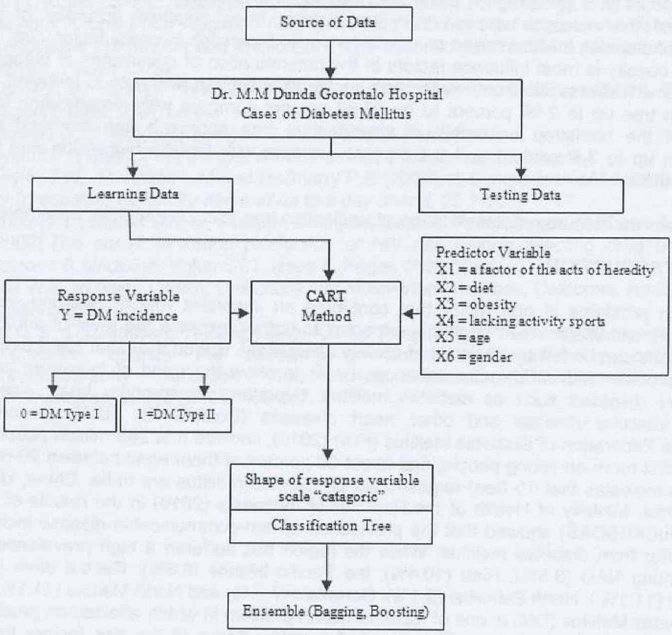

***Fig. 1.** Research Framework*

CART is one of the methods of Machine Learning, developed by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone in 1984. This is a method of decision tree classification algorithm with overall insulation technique recursive binary sorting is done on a group of data collected in a space called the node into two nodes. Each child node can then be divided again into two child nodes again, and so on until it meets certain criteria. All the nodes are determined based on the criteria of the most variable capable of sorting out the se-so the child node to be more homogeneous than the initial node.

The best sorting criteria measured from the Gini index which is formulated as follows:

$$i(t)=\sum_{i,j=-1} p(j\mid t)p(i\mid t) \quad , i\neq j$$

where p (j | t) is the proportion of classes j in node t and p (i | t) is the proportion of class i at the knot t. evaluation of sorting used goodness of split $\phi(s,t)$ of the parser node s on t-define as a decrease in heterogeneity:

$$\phi(s,t)=i(t)-p_L i(t_L)-p_R i(t_R)$$

where
i(t) = t node heterogeneity on the function
p_L = the proportion of observations left node
p_R = the proportion of observations to the right knot
$i(t_L)$ = function node left child on the heterogeneity i (t_R) = function of heterogeneity on the right child node

The parser generates $\phi(s,t)$ is higher is the best because the parser reduces the heterogeneity was higher.

Ensemble Methods

Ensemble methods developed with the hope of improving the accuracy of classification of a single classifier. The basic idea is to use much the same method of classifier and later to combination through the process of voting to gain a final classification conjecture (Wezel and Potharst, 2007).

Bagging

Bagging is an acronym of Bootstrap Aggregatting introduced by Breiman (1994) with the aim of reducing the variance of the Predictor. The basic idea of this method is to use a resampling ensemble of random data set returns early so that a new set of data obtained for classification trees evoke with many versions which are then combined to obtain the final prediction. The combination of multiple versions of the classification tree is expected to improve the accuracy of classification classification single trees.

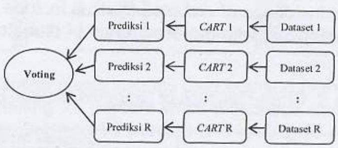

Fig. 2. Bagging CART

Boosting

Boosting is a family ensemble which includes a lot of AdaBoost algorithm, which is one of the most popular. This method was first developed by Freund and Schapire in 1995. Boosting in general focus to make a series of classification trees. The Dataset that is used on every tree classification relies on the previous classification trees and focus on the wrong data predicted. The wrong Data is predicted to be repaired by continuous tree-tree classification classification next. Cao Xu, Liang, Zhang, and Li (2010) States that the Boosting method is one of the best ensemble. The following is the algorithm AdaBoost. M1 for binary classification:

Input: Dataset $D = \{(x_1, y_1), (x_2, y_2),..., (x_m, y_m)\}$
Algoritma Base Learning L;
Count of Replication R;

Process:
(i) the sample weighting data Initialization $D_1(i) = 1/m$
(ii) For $r = 1,..., R$:

a. Do predictions h_r for a dataset D has been weighted D_r : $h_r = L(D, D_r)$
b. Calculate the weights D_r from data wrong classified ε_r
c. Calculate the weights voting

$$\alpha_r = \frac{1}{2} \ln\left(\frac{1-\varepsilon_r}{\varepsilon_r}\right)$$

d. Updated weights the sample data $$D_{r+1}(i) = \begin{cases} \frac{D_r(i)}{Z_r} \times \exp(-\alpha_r) & , h_r(x_i) = y_i \\ \frac{D_r(i)}{Z_r} \times \exp(\alpha_r) & , h_r(x_i) \neq y_i \end{cases}$$

where Z_r is a normalization factor to $\sum_{i=1}^{m} D_{r+1}(i) = 1$

(iii)Finish

Output: Voting classification prediction:

$$H(x) = sign\left(\sum_{r=1}^{R} \alpha_r h_r(x)\right)$$

Illustration AdaBoost.M1 Indicated in the figure 3 as follows:

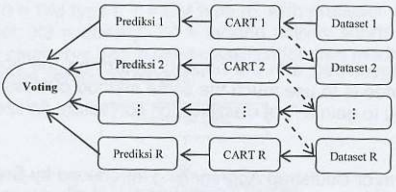

Fig. 3. Boosting CART

3. RESULTS AND DISCUSSION

Relationship of Predictor Variable with Diabetes Melitus Incidences

The characteristics of the incidence of diabetes mellitus include A factor of the acts of heredity, obesity, gender and age its relationship with the occurrence of diabetes mellitus in detail is presented in the table 1 below.

Table 1. Characteristics of Diabetes Melitus Patient

<table>
<tr><td rowspan="2" colspan="2">Predictor Variable</td><td colspan="4">Diabetes Mellitus Incidences</td></tr>
<tr><td colspan="2">DM Type I</td><td colspan="2">DM Type II</td></tr>
<tr><td rowspan="2">A factor of the acts of heredity</td><td>Doesn't Exist</td><td colspan="2">44 (7.9)</td><td colspan="2">48 (8.6)</td></tr>
<tr><td>Exist</td><td colspan="2">0 (0)</td><td colspan="2">468 (83.6)</td></tr>
<tr><td colspan="6">Pearson Chi-Square = 242.912 with Asymp. Sig. = 0.000</td></tr>
<tr><td rowspan="2">Obesity</td><td>No</td><td colspan="2">27 (4.8)</td><td colspan="2">46 (8.2)</td></tr>
<tr><td>Yes</td><td colspan="2">17 (3.0)</td><td colspan="2">470 (83.9)</td></tr>
<tr><td colspan="6">Pearson Chi-Square = 98.381 with Asymp. Sig. = 0.000</td></tr>
<tr><td rowspan="2">Gender</td><td>Male</td><td colspan="2">27 (4.8)</td><td colspan="2">230 (41.1)</td></tr>
<tr><td>Female</td><td colspan="2">17 (3.0)</td><td colspan="2">286 (51.1)</td></tr>
<tr><td colspan="6">Pearson Chi-Square = 4.603 with Asymp. Sig. = 0.032</td></tr>
<tr><td rowspan="3" colspan="2">Age</td><td colspan="2">DM Type I</td><td colspan="2">DM Type II</td></tr>
<tr><td>Mean</td><td>Stdev</td><td>Mean</td><td>Stdev</td></tr>
<tr><td>53.91</td><td>10.124</td><td>53.91</td><td>10.124</td></tr>
</table>

The percentage of patients with diabetes mellitus occurrence Hospital Dr. M.M. Dunda Gorontalo achieved 92,1 percent. Meanwhile, the percentage of patients there is no incidence of diabetes mellitus found amounted to 7,9 per cent. The percentage of patients who did not have descendants of diabetes mellitus and type I DM amounted to 7,9 per cent, amounting to DM occurs next 8.6 percent. The percentage who are descendants of diabetes mellitus and DM type I is 0 percent, then the DM type II amounted to 83,6 percent. This shows that the majority of patients who have diabetes mellitus descent tend to be exposed to the DM. this trend is also confirmed by the test results of Pearson Chi-Square of 242,912 or Sig. = 0.0000 are smaller than α = 5 per cent, which means that the offspring (gene) there is a connection with the Genesis DM.

The percentage of patients who are not obese and DM type I amounted to 4,8 percent, then occurred case of DM of 8,2 percent. The percentage of obese patients with DM type I is 3,0 percent, then DM type II amounted to 83,9 percent. This shows that the majority of obese patients are likely to be exposed to the DM. This trend is also confirmed by the test results of Pearson Chi-Square of 98,381 or Sig. = 0,0000 are smaller than α = 5 percent, which means that obesity is having relationship with the Genesis DM.

Based on the table 1 it appears that the average of age of the patients who had DM type I amounting to 54 years with standard deviation of 10.124, while the age of the patients who had DM have average of age of 54 with standard deviation of 10.968. It showed that the average age of the patients who happen DM or not is same. The percentage of a patient with the male sex and DM type I reached 4,8 %, next occurring DM of 41.1 percent. Meanwhile the percentage of a patient with the female sex and DM type I of 3,0 percent next DM type II of 51,1 percent. It goes to show that between the patient of the male and female are having same exposed to DM, but viewed from the trends of patients, patient women more than patient man. This trend is also strengthened by the outcome of a test of the person chi-square of 4,603 or Sig. = 0,032 smaller than α = 5 percent, that can be meant that gender there is a relationship with the DM.

Determination of Learning and Testing Data

The application of a method of trees classifications in the classification of patients DM type I and DM type II in the province of Gorontalo done by first divide into two groups, sample data namely data learning and data testing. Breiman (1993) declaring that no particular provision pertaining to learning and testing, the proportion of data it ' s just data learning must be greater than data testing. Hence to research this time is tried combination of the proportion of learning and testing data among others (95 percent: 5 percent), (90 percent: 10 percent), (85 percent: 15 percent), (80 percent: 20 percent), (75 percent: 25 percent) and (70 percent: 30 percent). Exactness classifications data used as the basis for testing could describe kindness tree classifications formed to classify in new data

Table 2. Comparison of Accuracy of Classification in Each Combination of Data

Combination of Data (%)		*Accuracy of Classification (%)*		*Number of Terminal Node*
Learning	*Testing*	*Learning*	*Testing*	
95	5	91,5	89,3	11
90	10	91,3	92,9	9
85	15	91,8	89,3	9
80	20	91,3	92,0	9
75	25	91,9	90.0	8
70	30	91,3	91,7	8

Table 2 implies that exactness classifications data learning highest given by the combination of data learning 75 percent and testing 25 percent while exactness classifications data testing highest given by the combination of data learning 90 percent and testing 10 percent. Because it was chosen combination of data trees yielding classifications more that have value exactness classifications testing is high. A combination of data selected is 90 percent for data learning and 10 percent for data testing. This combination trees yielding 9 terminal, with a knot but having the value of exactness classifications data testing high namely 92,9 percent. A combination of this is to be used for analysis next.

Classification of Diabetes Mellitus Incidences Using Classification Tree

Any knot terminal point is the end of an sorting presently trees, this knot can ' t be sorted back into a knot another or in other words a knot terminal is a knot containing the observations that homogeny and will eventually be included as a class of certain. A picture in detail of a noose of was as follows.

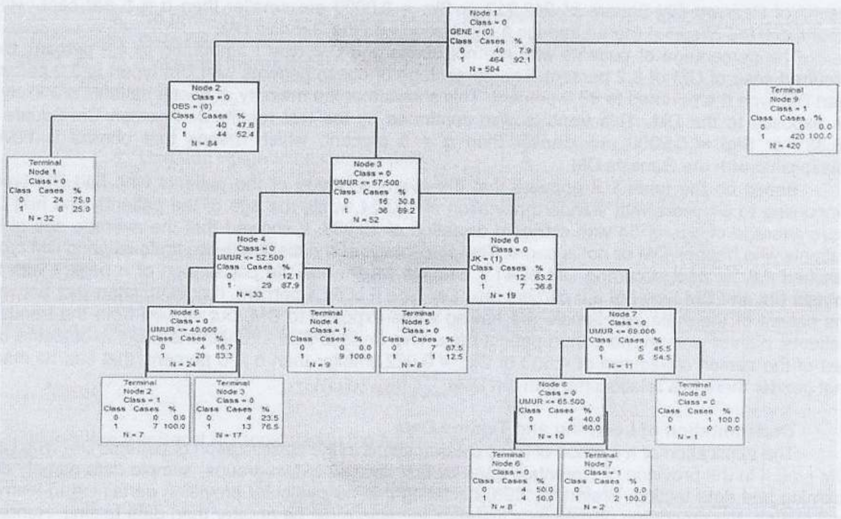

***Fig. 4.** Optimum Classification Tree of Diabetes Mellitus Patient*

Figure 4 can be interpreted in each carpel the knots of trees classifications scene dm as follows: a knot terminal 1: consisting of 32 observation that it is predicted as a group of patients dm type i as many as 24 and patients dm type ii as much as 8. Characteristic of the patient from this knot is as follows.

Obesity (X3): categories (0: not obesity).

A factor of the acts of heredity (X1): categories (0: no factor the acts of descent).

A knot terminal 2 consisting of 7 observation (patients) that it is predicted as a group of patients dm type II. Characteristic of the patient from this knot is as follows.

Age (X5): X5 <=40 years

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 3 comprises 17 observation (patients) that it is predicted as a group of patients dm type I as much as 4 and dm type ii as much as 13. Characteristic of the patient from this knot is as follows.

Age (X5): 40 < X5 <= 52.5 years

Obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 4 consisting of 9 observation (patients) that it is predicted as a group of patients dm type II. Characteristic of the patient from this knot is as follows.

Age (X5): 52.5 < X5 <=57.5 years

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of the descent)

A knot terminal 5 consisting of 9 observation (patients) that it is predicted as a group of patients dm type I there were 7 and dm type ii as much as 1. Characteristic of the patient from this knot is as follows.

Gender (X6): categories (1: male)

age (X5): X5 >57.5 years

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 6 consists of eight observation (patients) that it is predicted as a group of patients dm type I as much as 4 and dm type ii as much as 4. Characteristic of the patient from this knot is as follows.

Age (X5): 57.5 < x5 & < = 65.5 years and gender (X6): categories (0: female)

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 7 consisting of 2 observation (patients) that it is predicted as a group of patients dm type II. Characteristic of the patient from this knot is as follows.

Age (X5): 65.5 < X5 < = 69 years and gender (X6): categories (0: female)

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 8 consisting of 1 observation (patients) that it is predicted as a group of patients DM type I. Characteristic of the patient from this knot is as follows.

Age (X5): X5 >= 69 years and gender (X6): categories (0: female)

obesity (X3): categories (1: obesity)

A factor of the acts of heredity (X1): categories (0: no of the acts of descent)

A knot terminal 9 consisting of 420 observation (patients) that it is predicted as a group of patients DM type II. Characteristic of the patient from this knot caused by the acts of heredity (X1): categories (1: there is the history of the descendants)

Based on a tracing of maximum tree above seen that patients existing terminal 2, in a knot 4, of 7 and 9 is in a patient that is indicated as a group of patients with the dm. While the patient who belongs to a knot terminal 8 identified as a group of patients with no scene dm, and patients who belongs to a terminal 1, a knot 3, 5 and 6 were identified as a group of patients with no scene dm and there was an incident DM.

Table 3. Accuracy of Classification on *Learning* Data on Maximum Tree

Class Actual	Class Prediction		Accuracy (%)
	DM Type I	DM Type II	
DM Type I	40	0	100,00
DM Type II	26	438	94,37
Accuracy Total (%)			94,84

Table 3 indicate the great exactness classifications on data learning is 94,84 percent it means tree classifications formed capable of classify observations with right of 94,84 percent. Error classifications occurs in all classes observation, good group patient DM type I and patient DM type II. Patient DM type I wrong classified into a group of patients DM type II of 0. While 26 patient DM type II wrong classified into a group of patients DM type II.

Maximum model obtained next been validated to know feasible or not model formed in classify in new data. Test validation done with insert data testing as many as 56 into model tree classifications maximum has formed formerly of data learning.

Table 4. Accuracy of Classification on *Testing* Data on Maximum Tree

Class Actual	Class Prediction		Accuracy (%)
	DM Type I	DM Type II	
DM Type I	4	0	100,00
DM Type II	4	48	92,31
Accuracy Total (%)			92,86

Total accuracy of classifications on data testing is 92,86 percent means classifications tree formed able to classify observations with proper 59,22 into 92,86 percent. Patient DM type I any classified into groups patients DM type II amounted to 0. While 4 patients DM type II any classified into groups patients DM type I.

Classification of Diabetes Melitus Incicdences Using Ensemble Method

The application of bagging technique on classification tree used to increase accuracy of classifications tree produced by the method classifications tree as usual. Sampling of Bootstrap process was done to produce a learning data set loans that would be used to construct classification tree. Many bootstrap replication samples made is 50, 100, 150, 200 and 250 times. Table 5.19 presenting results testing exactness classifications resulting from replication samples bootstrap.

Table 5. Comparison of Accuracy of Classification of Ensemble Method Application

Number of Replication	Accuracy (%)		
	Case by Case Sample	Ensemble Bagging	Ensemble Arcing
50	97,8	97,8	91,1
100	95,6	97,8	96,3
150	97,8	97,8	96,3
200	91,1	97,8	97,8
250	97,8	97,8	97,8

Table 5 indicating exactness classifications tending to rise as you get the number replication until obtained exactness classifications optimum on replication 50 times as that of 97,8 percent. Exactness classifications unchanged although the replication be increased to up 200 times.

Many trees classifications formed by algorithms bagging tree of classifying according to many replication samples bootstrap made namely 50 fruit trees classification. Testing exactness tree bagging classifications tree classifications in classifies cases done by means of data running testing through 50th classifications the tree and then compute bagging misclassification rate.

4. CONCLUSION

Results showed patients with diabetes mellitus influenced by the A factor of the acts of heredity, diet, obesity, lack of exercise activity, age and gender, with a classification accuracy of 94,84 percent. Grouping or classification of patients with diabetes mellitus type I and type II are grouped into 9 groups involving interaction predictor variables. History descent is the dominant factor in classifying diabetes mellitus, diet and obesity later and less active sports. Development of a classification tree by bootstrap bagging ensemble can improve classification accuracy up to 2,96 percent, which became 97,80 per cent compared with the classification tree method. Flexibility bootstrap approach to ensemble classification tree also supported by other classification methods, such as logistic regression and discriminant analysis. The results of empirical studies show that the bootstrap aggregating and combine adaptive bootstrapping can work well in determining the selection of the best model is applied to the classification tree models. The empirical results related to the comparison between the predictive accuracy of classification tree models and other methods of classification that indicates there is a tendency that the classification tree models (new hybrid) gives better results, i.e. instead of the logistic regression and discriminant analysis, classification tree models bootstrap ensemble improves the classification accuracy of up to 3,3 percent and 6,4 percent. Moreover, empirical studies of the effect of the initial processing of the data concerning the determination of the data in-sample and out-sample, many observations and the determination of the minimum in the index group and ensemble methods which, if chosen correctly can improve the accuracy of classification. It can be seen from both the classification accuracy of the data in-sample and out-sample the data on the grouping of patients with diabetes mellitus hospital Dr. M.M. Dunda Gorontalo.

REFERENCES

1. Banerjee A.K., Neelima A., U.S.N Murty. (2008). Classification and Regression Tree (CART) Analysis for Deriving Variable Importance of Parameters Influencing Average Flexibility of CaMK Kinase Family. *Electronic Journal of Biology, Vol 4(1) 27-23.*
2. Bauer E. and Ron K. (1998). An Empirical Comparasion of Voting Classification Algorithm: Bagging, Boosting, and Variants. *Machine Learning*, vv, 1 – 138. Boston Manufactured in The Netherlands.
3. Boinee P., Alessandro De A., and Gian L.F. (2005). Ensembling Classifiers – An Application to Image Data Classification from Cherenkov Telescope Experiment. World Academy of Science, Enggieering and Techonolgy, 12.
4. Breiman L., Friedman J., Olshen R. and Stone C. (1984). *Classification and regression trees.* Chapman Hall, New York – London.
5. Bühlman P. dan Yu B.(2002), Analyzing Bagging, *The Annals of Statistics*, Vol. 30 no. 4, hal 927-961.
6. Clarke R.T. dan Bintercourt H.R. (2003), *Use of Classification And Regression Trees (CART) to Classify Remotely_Sensed Digital Images,* Research Report, Centro stadual de

Pesquisas em Sensoriamento Remoto Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, Brazil.

7. Cao D.S, Xu Q.S., Liang Y.Z, Zhang L.X., dan Li H.D., (2010). The Boosting: A New Idea of Building Models. *Chemometrics and Intelligent Laboratory Systems*, Vol. 100, hal. 1-11
8. Departemen Kesehatan RI. (2010). *Peran Diet Dalam Penanggulangan Diabetes*. Jakarta
9. Dillon W.R. (1978). On the Performance of Some Multinomial Classification Rules, *Journal of American Statistical Association*, 73, pp.305-313
10. Efron B. dan Tibshirani R.J. (1993). *An Introduction to the Bootstrap.* Chapman Hall, New York.
11. Freund Y. (1995), Boosting a weak learning algorithm by majority. *Inform. Comput*, Vol.121, hal 256-285.
12. Freund Y., and Schapire R., (1996). Experiments with a New boosting algorithm, *The Annals of statistics*, Vol 26, No.3, hal 801 - 849.
13. Hastie T., Tibshirani R., Friedman J., (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York
14. Jonathan S. et. al. (1999). *The Epidemiology of Diabetes: a World-Wide Problem*. University of Sydney.
15. Kavitha K., Sarojamma R.M. (2012). *Monitoring of Diabetes with Data Mining via CART Method.* International Journal of Emerging Technology and Advanced Engineering, Vol.2, Issue 11.
16. Khardori R. (2011). Type 2 Diabetes Melitus. *http://emedicine.medscape.com /article/117853-overview#a0104.* Diakses pada tanggal 25 September 2012 pukul 20.00.
17. Lewis and Roger J. (2000), An Introduction to Classification And Regression Trees (CART) Analysis. *Departement of Emergency Medicine Harbor-UCLA Medical Center*, Torrance, California.
18. Michie D., Spiegelhalter D.J. and Taylor C.C. (1994). *Machine Learning, Neural and Statistical Classification.*
19. Okun O., Valentini G., Re M. (2011). *Ensembles in Machine Learning Applications*. New York:Springer Science+Bussines Media, LLC
20. Otok B.W. (2008), *"Pendekatan Bootstrap pada Model Multivariate Adaptive Regression Spline"*, Desertasi, Universitas Gadjah Mada, Yogyakarta.
21. Otok B.W. dan Sumarmi (2009). Bagging CART Pada Klasifikasi Anak Putus Sekolah. *Seminar Nasional Statistika IX.* Institut Teknologi Sepuluh Nopember. Surabaya.
22. Prasad A.M., Louis R.I., and Andy L. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystem*, 9: 181-199.
23. Seng C.Y. (2006). Classification and Regression Trees (CART) for SPAM Prediction. *Journal of Computational Intelligent: Method and Applications.* Assignment. 2
24. Schapire R. (1990). The Strength of weak learnability. *Machine Learning.* 5, hal 197 - 227.
25. Schwarzer G., Nagata T., Mattern D., Schmelzeisen, R., and Schumacher. (2003). Comparison of Fuzzy Inference, Logistic Regression, and Classification Trees (CART). *Methods Inf Med*; 42:572-7.
26. Raina R., Yirong S., Andrew Y.Ng, and Andrew McCallum. (2004). Classification with Hybrid Generative/Discriminative Models. Department of Computer Science. Stanford University.
27. Seewald A.K. Johann P. and Gerhard W. (2001). Hybrid Decision Tree Learner with Alternative Leaf Classifiers: An Empirical Study. *Proceedings. American Association for Artificial Intelligence.*
28. Sutton Clifton D. (2005). Classification and regression trees, Bagging, and Boosting. *Handbook of statistics*, Vol. 24.
29. Steinberg, D., dan Colla, P., (1995). *CART: Tree-structured nonparametric data analysis.* San Diego, Calif., U.S.A.: Salford Systems.
30. Steinberg D. dan Phillip, C., (2005). *CART – Classification and Regression Trees.* CA: Salford System, San Diego
31. Suyono S. (2007). *Patofisiology Diabetes Melitus,* Balai Penerbit FKUI, Jakarta.
32. Tandra Hans. (2007). *Diabetes Panduan Lengkap Mengenal dan Mengatasi Diabetes dengan Cepat dan Mudah.* Jakarta
33. Waspadji S. (2007). *Pengobatan dengan Insulin pada NIDDM.*, Jakarta
34. Wezel M.V. dan Potharst R., (2007). Improved Customer Choice Predictions using Ensemble Methods. *European Journal of Operational Research*, Vol. 181, hal.436-452.