ISSN 2579-7069

Comparison of Cart and Naive Bayesian Algorithm Performance to Diagnose Diabetes Mellitus

Irfan Santiko [a,1,*], Pungkas Subarkah [b,2]
[a] Informatic Departement STMIK AMIKOM Purwokerto, Let.Jend. Pol Sumarto, Purwokerto, 53113, Indonesia
[b] Informatic Departement STMIK AMIKOM Purwokerto, Let.Jend. Pol Sumarto, Purwokerto, 53113, Indonesia
[1] irfan.santiko@amikompurwokerto.ac.id *; [2] subarkah18.pungkas@gmail.com
* corresponding author

Abstract

Based on Indonesia's health profile in 2008, Diabetes Mellitus is the cause of the ranking of six for all ages in Indonesia with the proportion of deaths of 5.7% under stroke, TB, hypertension, injury and perinatal. This is reinforced by WHO (2003), Diabetes Mellitus disease reached 194 million people or 5.1 percent of the world's adult population and in 2025 is expected to increase to 333 million inhabitants. In particular, in Indonesia, people with Diabetes Mellitus are increasing. In 2000, Diabetes Mellitus sufferers have reached 8.4 million people and it is estimated that the prevalence of Diabetes Mellitus in 2030 in Indonesia reaches 21.3 million people. This allows researchers and practitioners to focus their attention on detecting/diagnosing diabetes mellitus and to prevent it because the disease can cause complications. The method used in this research was problem identification, data collection, pre-processing stage, classification method, validation and evaluation and conclusion. The algorithm used in this research was CART and Naïve Bayes using dataset taken from UCI Indian Pima database repository consisting of clinical data of patients who detected positive and negative diabetes mellitus. Validation and evaluation method used was 10-cross validation and confusion Matrix for the assessment of precision, recall and F-Measure. The result of calculation has been done, got the accuracy result on CART algorithm equaled to 76.9337% with precision 0.764%, recall 0.769%, and F-Measure 0.765%. While the diabetes dataset was tested with the Naïve Bayes algorithm, got an accuracy of 73.7569% with precision 0.732%, recall 0.738%, and F-Measure 0.734%. From these results it can be concluded that to diagnose diabetes mellitus disease it is suggested to use CART algorithm.

Keywords: Performance; Diagnosis; Algorithm;Diabetes Mellitus

1. Introduction

Diabetes Mellitus disease is one of the serious health threats and can cause death both in Indonesia and in the world. According to a survey conducted by the World Health Organization (WHO) in 2005, Indonesia as a lower-middle income country ranks fourth with the largest number of Diabetes Mellitus sufferers in the world after India, China and the United States (MOH RI, 2009). Based on Indonesia's health profile in 2008, Diabetes Mellitus is the cause of the ranking of six for all ages in Indonesia with the proportion of deaths of 5.7% under stroke, TB, hypertension, injury and perinatal. This is reinforced by WHO (2003), Diabetes Mellitus disease reached 194 million people or 5.1 percent of the world's adult population and in 2025 is expected to increase to 333 million inhabitants. In particular, in Indonesia, people with Diabetes Mellitus are increasing. In 2000, Diabetes Mellitus sufferers have reached 8.4 million people and it is estimated that the prevalence of Diabetes Mellitus in 2030 in Indonesia reaches 21.3 million people [1].

The Entitled Implementation of Fuzzy Classification System Based on Optimization of Ants Colony to Diagnose Diabetes Disease. The result was obtained by the algorithm accuracy of the ant colony optimization algorithm of 78.55%. Accuracy of Naïve bayes algorithm of 74.32% and C.4.5 of 85.13%. The same thing was done using J48 algorithm and got the accuracy equal to 74,72% [7][13].

From reviewing the research that has been done, by the authors would like to conducted similar research using the CART and Naïve Bayes algorithms. The CART algorithm is an algorithm whose high accuracy is seen in related research and the Naïve Bayes algorithm has a Bayesian classification algorithm having similar classification capabilities to the decision tree, neural network [6].

2. The Proposed Method/Algorithm

2.1. Literature Review

Entitled Performance Comparison of Decision Tree J48 and ID3 In Classification Diagnosis of Diabetes Mellitus Disease. The aim of the study was to extract information from the Pima Indian diabetes dataset used as a decision support system to predict the diagnosis of diabetes mellitus. The result showed that the accuracy of J48 algorithm was 74.72% higher than ID3 algorithm which was only 72.64% [8][14].
The title Determination of Large Accuracy Classification Method Using Algorithm C4.5 Particle-Based Swarm Optimazation On Predicted Diabetes Disease. The purpose of this study was to get the rule in predicting diabetes mellitus and to provide an accurate value. The result obtained an accuracy value and AUC value was higher than C4.5 algorithm with accurary difference of 3.28% and AUC 0.012 value [9].
The title Implementation of Fuzzy Classification System Based on Optimization of Ants Colony for Diagnosis of Diabetes Disease. The purpose of the study was to predict diabetes using expert systems. The result of this research was the implementation into expert system software built using Microsoft Visual Studio 2012 with the result of accuracy performance of 78.55% [12].
The title Comparison of C4.5 and CART Algorithm Performance in Classification data Student Value Prodi Computer Engineering Polytechnic Padang. The purpose of this study was to predict the most crucial lecture knowledge in determining students' graduation in the first semester. The result showed that C4.5 algorithm was 85.61% higher than CART algorithm 84.95% [11].

2.2. Diabetes Mellitus

Diabetes Mellitus is a chronic metabolic condition because the pancreas does not produce enough insulin or the body can not use insulin produced effectively. Insulin is a hormone that measures the balance of blood sugar levels, consequently if insulin is not balanced then there is an increase in glucose concentration in the blood or hyperglycemia [10][15].

2.3. Data of Indian Pima Diabetes

The dataset in this study was taken by UCI Indian Pima database repository (http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). The Pima Indian dataset consists of 768 clinical data, all of which are of female gender with age of at least 21 years [9].

2.4. Data Mining

Data mining is a mean that aims to find patterns automatically or semi-automatically from existing data in a database or other data source used to solve a problem through various process rules [12].

2.5. Cart Algorithm

CART (classification and regression trees) algorithm is one method or method of algorithm from one of data exploration technique that is decision tree technique. This algorithm recursively divides records into exercise data into subsets that have the same target (class) attribute value. This method was developed by Leo Breiman, Jerome H. Friedman, Richard A. Olshen and Charles J. Stone around 1980. Here are some of the advantages of CART, among others [11]:

1. CART is a method that is nonparametric / suitable for data type of numeric.
2. CART does not require a variable to be selected first.
3. CART generates an invariant for independent variable transformations.

The CART algorithm develops a decision tree by selecting the most optimal branch for each node. The classification of CART (Classification and Regression Tress) algorithms, a record will be classified into one of the classifications available in the destination variable based on the values of the predictor variables. Characteristic of CART algorithm is node of decision which is always bifurcated or branched binary [11].

The steps in the CART algorithm are as follows:

1. Arrange candidate branch (candidate split). This arrangement is done on all predicted or predicted variables (exhaustive). The list of candidate branches is called the latest branch candidate.
2. Give an overall assessment of the latest branch candidates by calculating the value of the magnitude $\emptyset$ (s|t).
3. Determine which candidates will actually be made a branch by choosing a candidate for a branch of the biggest good value $\emptyset$ (s|t). After that, draw branches. If no more decision nodes, the implementation of the CART

algorithm will be terminated. Let Ø (s | t) be the "good" value of candidate branch s on the decision node t, then the value Ø (s | t), can be identified by the following equations [7]:

$$\emptyset(s|t) = 2P_LP_RQ(s|t) \quad (1)$$

$$Q(s|t) = 2\, P_LP_R \sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)| \quad (2)$$

Where the Equation is:

t_L = The left branch candidate of the decision node t

t_R = The right branch candidate of the decision node t

$$P_L = \frac{\text{Number of records on the left branch candidate t\_L}}{\text{The total number of records in the exercise data}}$$

$$P_R = \frac{\text{Number of records on the right branch candidate } t_R}{\text{The total number of records in the exercise data}}$$

$$P(j|t_L) = \frac{\text{The number of records j on the left branch candidate } t_L}{\text{The number of records on the decision node t}}$$

$$P(j|t_R) = \frac{\text{The number of records j on the right branch candidate } t_R}{\text{The number of records on the decision node t}}$$

2.6. Naive Bayesian Algorithm

The naïve bayes algorithm is a simple probability-based prediction technique based on the application of Bayes rules with the assumption of strong independence. In addition, naïve bayes can also analyze the variables that most influence it in the form of opportunities [2].

Naïve Bayes is the most effective and efficient algorithm or method for learning machine design and data mining. Here are some advantages of Naïve Bayes Algorithm among others [3]:

1. Naïve bayes algorithm is easy to use for machine learning data.
2. Naïve bayes algorithm requires only one scan of training data.
3. It is used for handling missing attribute values and continuous data.

The Naes Bayes classification is a classification of statistics that can be used to predict the probability of membership of a class. Bayesian classification is based on Bayes's theorem, derived from the name of a mathematician who is also British Prebysterian minister [4]. The Bayesian classification has similar classification capabilities to the decision tree and neural network [6].

The general form of the bayes theorem is as follows [3]:

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

Where :

X : Data with unknown class

H : The hypothesis of X data is a specific class

P(H|H) : The probability of hypothesis H is based on condition X (*posteriori probability)*

P(H) : Hypothesis Probability H (*prior probability*)

P(X|H) : The probability of X is based on the condition in hypothesis H

P(X) : The probability of X.

2.7. Weka

Weka (waikato environment for knowledge analysis) is a Java-based data mining platform application. This application was first developed by Waikato University of New Zealand before becoming part of Pentaho. Weka consists of a collection of machine learning algorithms that can be used to generalize or formulate a collection of sampling data [7].

2.8. Accuracy

Accuracy is the value of the degree of proximity of the quantity measurement to the true value (rule). Accuracy values is obtained from the results of the rule generated from the calculation of CART and it is tried on the data testing and generates degrees of accuracy of the rule after the test on the data testing. Here is the formula of accuracy value, accuracy formula [4]:

$$\text{Accuracy} = \frac{\text{number of TP+number of TN}}{\text{number of TP+FP+FN+TN}}$$

Where :

TP : *True Positive*

TN : *True Negative*

FP : *False Positive*

FN : *False Negative*

2.9. Confusion Matrix

Evaluation by means of confusion matrix produces precision value, recall, F-Measure and accuracy. Here is a confusion matrix in Table 1.

Table 1 Confusion Matrix

Correct Classification	Classified as	
	+	-
+	True positives	False negatives
-	False positives	True negatives

(Source : Han & Kamber, 2006)

The formula for calculating *precision, recall, F-Measure* dan sebuah *accuracy* (Han & Kamber, 2006) is as follows :

$$Precision = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

$$Recall = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

$$F\text{-}Measure = \frac{\boldsymbol{2 \times precision \times recall}}{\boldsymbol{precision + recall}}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

3. Method

3.1. Method of Collecting Data

Data collection method used in this research was literature study by taking secondary data.

1. Literature Studies
   Literature study is a method of collecting data in order to collect information in the form of written sources (books, scientific magazines, archives, official documents and scientific papers), drawings and electronic documents supporting the writing process.
2. Secondary Data
   Secondary data is the source of research data obtained by researchers indirectly through intermediate media (obtained, recorded, or have been investigated by others). Secondary data is generally in the form of evidence, records or reports and historically compiled in published and unpublished (documentary) files. This data is used because the data source is taken from the UCI Indian Pima Diabetes repository.

3.2. Research Concept

The research flow is as follows:

1. Identification of Problems
   The process of identifying the problem was done as an effort to find out the problems and methods appropriate for this research.
2. Data Collection
   In this study, the secondary data used were taken from the UCI Indian Pima Diabetes database repository consisting of 768 clinical data.
3. Pre-processing stage.
   This stage was done to get the data clean and ready to use. The pre-processing data stage included identification and attribute identification and selection, handling missing values, and value discretization process.
4. Use of Classification Methods
   The steps performed in the CART algorithm were:
   a. The Pima Indian dataset was classified using the CART (SimpleCart) algorithm in the Weka app.
   b. Training and testing process using the 10-cross validation method were performed.
   c. Classifier results were obtained and confusion matrix was generated.
   d. The result of decision tree decision from output result of CART algorithm can be seen
   e. From the confusion matrix results, the precision value, recall, F-Measure can be calculated by describing confusion matrix to be of confusion.

   While the steps performed in the Naïve Bayes algorithm were:
   a. The Indian Pima dataset was classified using the Naïve Bayes (NaiveBayes) algorithm in the Weka app.
   b. The training and testing process using the 10-cross validation method were performed.
   c. Classifier results were obtained and confusion matrix was generate.

d. From the confusion matrix results can be calculated precision value, recall, F-Measure by describing confusion matrix to be of confusion.

5. Validation and Evaluation
   In this stage validation and measurement of the accuracy of the results were achieved by the model using the techniques contained in the application weka namely confusion matrix and cross-validation.
6. Withdrawal Conclusion
   The next step is to conclude the results obtained from the research. The CART or Naïve Bayes algorithm provided the best accuracy results for diagnosing diabetes mellitus based on precision, recall, F-Measure values of each algorithm.

4. Results and Discussion

4.1. Identification of Problems

In this study the authors conducted a study of literature related to the research and got the appropriate algorithms for this research they are CART and Naïve Bayes.

4.2. Data Collection

In this study the data used was taking from the Pima Diabetes Indian repository consisting of 768 clinical data all of which were from female genitalia with a minimum age of 21 years. Below is table 2 Diabetes Mellitus Dataset Indian Pima:

Table 2 Diabetes Mellitus Indian Pima Dataset

No	*Pregnant*	*Glucose*	DBP	TSFT	INS	BMI	DPF	*Age*	*Class*
1	6	148	72	35	0	33.6	0.627	50	Positive
2	1	85	66	29	0	26.6	0.351	31	Negative
…	…	….	….	….	….	….	….	….	….
768	1	93	70	31	0	30.4	0.315	23	Negative

4.3. Pre-processing stage

After analyzing the Pima Indian dataset, it was known that not all attributes had a complete value, where the attribute attribute value greatly affected the classification. Attributes that had an incomplete amount of data were pregnant as many as 111, 5 glucose attributes, 35 DBP attributes, 2227 TSFT attributes, 374 attributes of INS and 11 attributes of BMI 11. While attribute age and class had a complete value. To handle missing value:

1. The zero value in the pregnant attribute can be assumed that the value indicated the patient had never given birth, so this was possible in accordance with the actual conditions.
2. Data with zero values on the attributes of glucose, DBP and BMI can be removed because the amount was not too much so it did not really affect the classification results.
3. Because the TSFT and INS attributes had a non-existent number of values, these two attributes can not be removed and can not be used in the classification. Therefore, in this study the attributes of TSFT and INS were not used.

After the process of handling the incomplete value (missing value) was done with the above rules, we got 724 data (249 positive and 475 class negative) from 768 original data and ready to be processed further with attributes of pregnant, glucose, DBP, BMI, DPF, Age and class.

However, first discrete attribute process was done. The goal was to facilitate the grouping of values based on predetermined criteria. It also aimed to simplify the value of problems and to improve accuracy in the learning process (Lesmana, 2012). The glucose attribute was divided into three, namely low, medium and high. DBP attributes were divided into three, namely normal, normal-to-high and high (Patil, et al, 2010). While the attributes of BMI were grouped into four, namely low, normal, obese, and severely-obese (Patil, et al, 2010). DPF attributes were divided into two groups: low and high. Class attributes were divided into two groups, namely positive diabetes and negative diabetes.

4.4. Process of Classification Method

From the results of calculations and trials using weka applications with CART algorithm, it was produced an accuracy of 76.9337%. The accuracy value was obtained from the calculation of precision, recall, and F-measure. The result of calculating accuracy value of confusion matrix is presented in Table 3 as follows:

Table 3 value of accuracy based on confusion matrix

Class	*Precision*	*Recall*	*F-Measure*
Tested_negative	0.804	0.857	0.83
Tested_positive	0.688	0.602	0.642
Weighted Avg	0.764	0.769	0.765

Whereas if the classification using Naïve Bayes algorithm using weka application gave an accuracy of 73.7569%. Accuracy value was obtained from the calculation of precision, recall and F-Measure. The result of calculating the accuracy value of confusion matrix is presented in Table 4 as follows:

Table 4 value of accuracy based on confusion matrix

Class	*Precision*	*Recall*	*F-Measure*
Tested_negative	0.783	0.829	0.806
Tested_positive	0.633	0.562	0.596
Weighted Avg	0.732	0.738	0.734

Here is the difference of accuracy results obtained using CART and Naïve Bayes algorithm in table 5.

Table 5 Comparison of Accuracy Results of CART and Naïve Bayes

Algorithm	Accuracy Results	*Precision*	*Recall*	*F-Measure*	Time
CART	76.9337 %	0.764	0.769	0.765	0.16 *second*
Naïve Bayes	73.7569 %	0.732	0.738	0.734	0.04 *second*

The difference of accuracy obtained by using CART and Naïve Bayes algorithm was 3.1768%. The time spent on running the dataset on Weka applications also differed between the CART and Naïve Bayes algorithms of 0.12 second difference. In the CART algorithm there was a difference that recursively divided records into exercise data into subsets that have the same target attribute value (class), which caused long time for compiling system.

4.5. Validation and Evaluation

Tables 6 and 7 are tables of confusion matrix results from dataset testing using CART and Naïve Bayes algorithms with 10-fold cross validation.

Table 6 Confirm CART matrix

	Positive Diabetes	Negative Diabetes
Positive Diabetes	407	68
Negative Diabetes	99	150
724	506	218

Tabel 7 *Confusion matrix Naïve Bayes*

	Positive Diabetes	Negative Diabetes
Positive Diabetes	394	81
Negative Diabetes	109	140
724	503	221

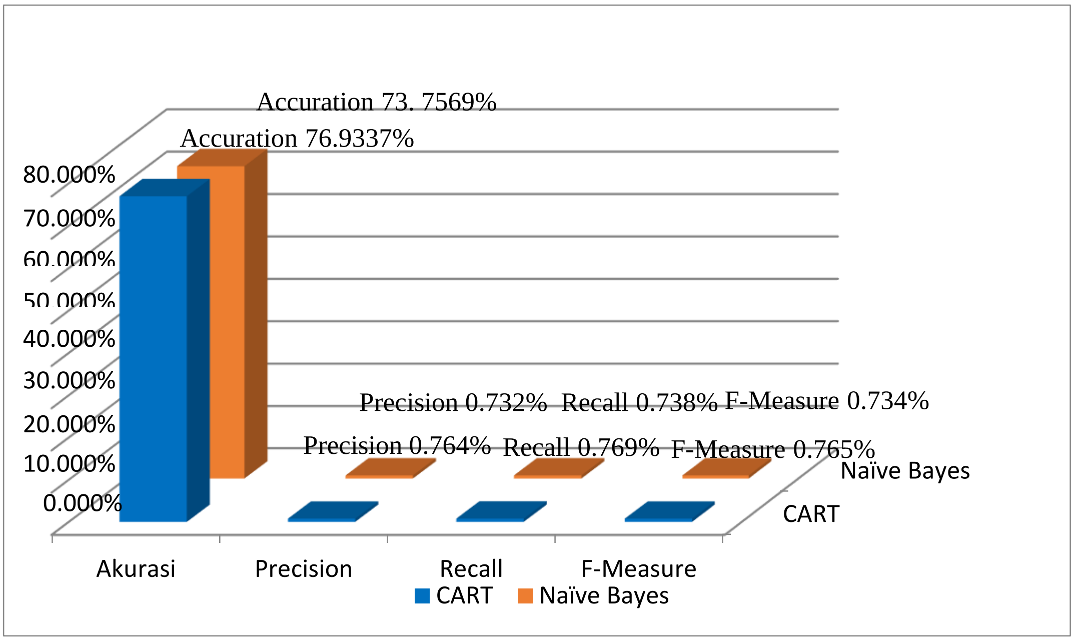
From table 4.5 it can be seen that the amount of data formation of the rule affected by diabetes mellitus was similar to the data testing that also affected by diabetes as much as 407. Then, the amount of data formation of the rule that was not affected by diabetes mellitus with data testing that affected by diabetes was as much as 68. Next , The amount of data formation of the rule affected by diabetes and data testing that was not exposed to diabetes was as much as 99. Finally, the amount of data formation of the rule that was not affected by diabetes was the same with the data testing that was also not affected by diabetes as much as 150.

From table 4.6 it is seen that the amount of data formation of the rule affected by diabetes mellitus was the same with the data testing which was also affected by diabetes as much as 394. Then, the amount of data formation of the rule that was not affected by diabetes mellitus with data testing affected by diabetes as much as 81. Furthermore, the amount of data formation of rule affected by diabetes and data testing that was not exposed to diabetes was as much as 109. Finally, the amount of data formation of rule that was not affected by diabetes was the same as the data testing that was also not affected by diabetes as much as 140.

4.6. Final Result

From the calculation that has been done on both algorithm, we got the accuracy result from each algorithm that was 76.9337% with precision value 0.764%, Recall 0.769% and F-Measure 0.765% on CART algorithm and 73.7569% on Naïve Bayes algorithm with precision value 0.732% , Recall 0.738% and F-Measure 0.734%. Here is a graph of accuracy results from calculations on the CART and Naïve Bayes algorithms in Figure 4.1

Figure 4.1 Comparison of Accuracy Results of CART and Naïve Bayes

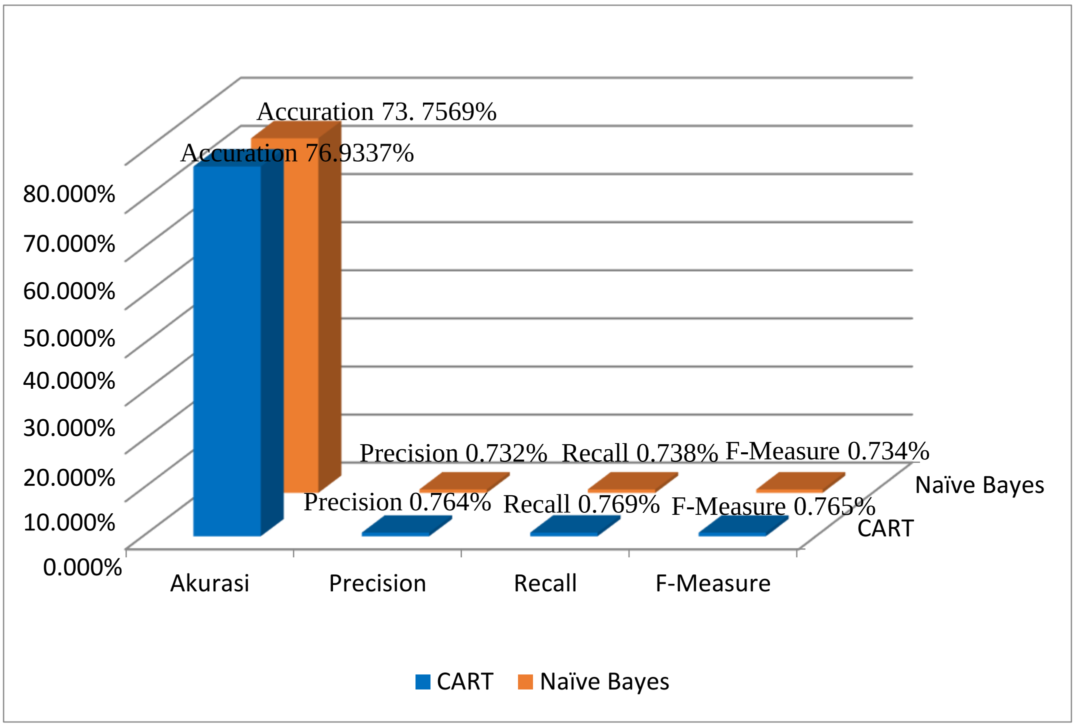


This after seeing the above calculation results, to determine the diagnosis of mellitus disease, it is better to use CART algorithm because the accuracy is higher than naive bayes algorithm..

5. Conclusion

After doing the pre-processing step and calculating by using two algorithms namely CART and Naïve Bayes, and the evaluation stage with confusion matrix, we obtained accuracy as much as 76.9337% with 0.764% precision value, 0.769% recall value and 0.765% F-Measure value on the algorithm CART and while the accuracy of Naïve Bayes was as much as 73.7569% with 0.732% precision value, 0.738% recall value and 0.734% F-Measure. Furthermore, confusion matrix resulted from two algorithms namely CART and Naïve Bayes algorithm can be seen with graph of accuracy result from calculation of two algortima in picture 5.1

Figure 5.1 Comparison of CART and Naïve Bayes algorithm accuracy results



From these results it can be concluded that to diagnose diabetes mellitus disease it was suggested using CART algorithm.

Acknowledgment

From the analysis that has been done by the authors to diagnose diabetes mellitus disease it was obtained the accuracy of two algorithms they were on CART and naïve bayes algorithm. The author hopes for further research, the results of this analysis can be implemented into a system. Higher accuracy results were CART algorithm. Furthermore, to get the results of better accuracy, the authors provide the following suggestions:

1. handling of missing values on each attribute.
2. Use another algorithm between ID3, C4.5 and KNN by looking at a higher level of accuracy.

References

[1] Diabetes Care. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030.

[2] Gorunescu,F.2011. *Data mining Concepts, Models and Techniques*. Verlan Berlin Heidelberg: Spinger

[3] Han,J., & Kamber,M.2006. *Data Mining Concepts And Techniques*. Verlag Berlin Heidelberg : Spinger

[4] Jayalskshmi, T., Santhakumaran, A., " *Impact of Prepocessing for diagnosis of diabetes mellitus using artificial neural network*," Machine Learning and Computing (ICMLC),2010 Second International Conference on, vol., no., pp.109-112,9-11 Feb.2010.

[5] Kemenkes RI.2014.Situasi dan Analisis Diabetes. Jakarta : Kemenkes RI

[6] Kusrini, & Lutfhi,E. T. 2009.*Algoritma Data Mining*.Yogyakarta:Andi Offset.

[7] Larose, D. T., 2005. *Discovering Knowledge In Data : An Introduction To Data Mining*. New Jersey : Wiley-nterscience.

[8] Patil, B.M., Joshi,R.C., Toshniwal,D.2010. Assosiation rule for classification of type 2 diabetic patients. *Machine Learning And Computing (ICMLC),pp.330-334*

[9] Pima Indians Diabetes Dataset, UCI Machine Learning Repository , diambil dari http://archive.ics.edu/ml/datasets/Pima+Indians+Diabetes. Diakses 29 Agustus 2016

[10] RISKESDAS, Indonesian Ministry of Health's Health Research and Development Agency, 2013.

[11] Timofeev, Roman.2004.*Classification and Regression Trees (CART) Theory and Aplications*.Humboldt University :,Berlin

[12] WEKA, Machine Learning Group at University of Waikato, from http://www.cs.waikato.ac.nz/ml/weka/ . Access at 29 Agustus 2016.

[13] International Diabetes Federation. Retrieve 3 July 2015, from http://www.idf.org/diabetesatlas/update-2014.

[14] World Health Organization. Retrieve 18 June 2015, from http://www.who.int/diabetes/en/.

[15] Han J, Kanber M. Pei J. Data Mining: Concepts and Techniques, 3rd ed. USA: Morgan Kaufman; 2012.