International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 363-369, 2018, ISSN:-2319–8656

Analysis of Comparison of Fuzzy Knn, C4.5 Algorithm, and Naïve Bayes Classification Method for Diabetes Mellitus Diagnosis

Putri Elfa Mas`udia
Departement of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Ridwan Rismanto
Departement of Technology Information
State Polytechnic of Malang
Malang, Indonesia

Abdullah Mas`ud
Departement of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Abstract: Early detection of diabetes mellitus (DM) can prevent or inhibit complication. There are several laboratory test that must be done to detect DM. The result of this laboratory test then converted into data training. Data training used in this study generated from UCI Pima Database with 6 attributes that were used to classify positive or negative diabetes. There are various classification methods that are commonly used, and in this study three of them were compared, which were fuzzy KNN, C4.5 algorithm and Naïve Bayes Classifier (NBC) with one identical case. The objective of this study was to create software to classify DM using tested methods and compared the three methods based on accuracy, precision, and recall. The results showed that the best method was Fuzzy KNN with average and maximum accuracy reached 96% and 98%, respectively. In second place, NBC method had respective average and maximum accuracy of 87.5% and 90%. Lastly, C4.5 algorithm had average and maximum accuracy of 79.5% and 86%, respectively.

Keywords : Fuzzy KNN, C4.5 Algorithm, Naïve Bayes Classifier, Diabetes Mellitus

1. INTRODUCTION

Diabetes mellitus (DM) is a disease marked by high level of blood sugar caused by impaired insulin secretion, insulin disruption, or both.DM is a heterogeneous group marked by increase on glucose level in the blood or hyperglycemia [1].

There are various classification methods, such as K-nearest neighbor (KNN), fuzzy KNN (F-KNN), decision treemethod using C4.5 algorithm, Naïve Bayes classifier (NBC) method, and many other methods. In previous studies, one of this methods was used to classify a problem without analyzing which classification method produce the best result. Yanita Selly conducted a study in 2013 to compare KNN and F-KNN methods. The result showed that F-KNN method is better than KNN method, as accuracy of F-KNN reached 98% while KNN only had 96% accuracy [12].

The result was then further analyzed in this study, where F-KNN, decision tree method using C4.5 algorithm, and Naïve Bayes classifier (NBC) method were compared. The results of these three methods were analyzed to obtain the best classification method.

A. *Research Objective*

1. To apply fuzzy KNN method, decision tree method using C4.5 algorithm, and Naïve Bayes classifier (NBC) methodin diagnosing DM.
2. To create a software to compare the three methods based on accuracy, time, precision, and recall.

B. *Related Results from Previous Studies*

Yanita Selly dkk compared DM classification using K-nearest neighbor (KNN) and fuzzyKNN methods. KNN is a classification method that perform strict prediction on tested data based on k nearest neighbor. Meanwhile, F-KNN predicts tested data based on membership value of tested data in each class, and then class data with highest membership was selected as resulting predicted class. The study results showed that F-KNN method is better than KNN method, as accuracy of F-KNN reached 98% while KNN only had 96% accuracy [12].

Other study conducted by Parida Purnana regarding detection of Type II DM using Naïve Bayesbased on particle swarm optimization. In the study, particle swarm optimizationwas used to improve accuracy in detecting DM. The study result showed that this method had 98.16% accuracy and 0.99 AUC, thus it can be classified as ‘excellent classification’ [9].

Larissa dkk conducted study regarding classification of client using C4.5 algorithm as creditingbasis. This study classify clients of a bank, so that when a problem occurs, the bank could easily obtain rules from the resulting decision tree. With decision tree method using C4.5 algorithm, process of gathering information was faster and more optimal with larger number of data, therefore the error in decision making could be minimized [4].

2. SYSTEM PLANNING

The steps of this research were:

1. Studying literatures regarding fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier methods.
2. Studying dataset from Indian Pima Diabetes that were used as trainingdata
3. Designing software to perform classification in accordance with tested methods.
4. Applying fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier methodsto diagnose DM.

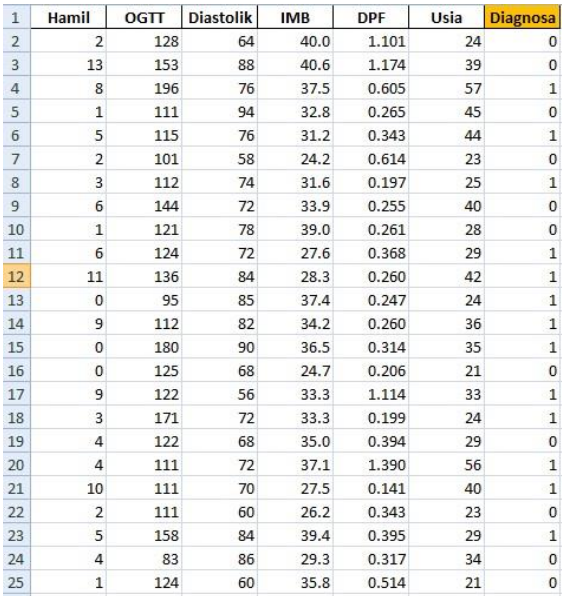
5. Testing and analyzing the results of each method and calculating the accuracy.

2.1 Data Preprocessing Method

This study used dataset that were then classified as training data and testingdata. These data were obtained from UCI machine learning repository database: Indian Pima Diabetes in http://archieve.ics.uci.edu. There are 768 clinical data in Indian Pima database but not all attributes are completely available.

768 clinical data obtained from Indian Pima Database were preprocessed, which means that insignificant data was deleted to maximize classification result. Missing valueor data with incomplete attributes was treated using rules from [LES-12], since the classification result is highly influential to training data.

From 8 parameter of data, parameters of TSFT and INS were deleted since the missing value was very large. The preprocessed data are displayed in Figure 1.



1	Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
2	2	128	64	40.0	1.101	24	0
3	13	153	88	40.6	1.174	39	0
4	8	196	76	37.5	0.605	57	1
5	1	111	94	32.8	0.265	45	0
6	5	115	76	31.2	0.343	44	1
7	2	101	58	24.2	0.614	23	0
8	3	112	74	31.6	0.197	25	1
9	6	144	72	33.9	0.255	40	0
10	1	121	78	39.0	0.261	28	0
11	6	124	72	27.6	0.368	29	1
12	11	136	84	28.3	0.260	42	1
13	0	95	85	37.4	0.247	24	1
14	9	112	82	34.2	0.260	36	1
15	0	180	90	36.5	0.314	35	1
16	0	125	68	24.7	0.206	21	0
17	9	122	56	33.3	1.114	33	1
18	3	171	72	33.3	0.199	24	1
19	4	122	68	35.0	0.394	29	0
20	4	111	72	37.1	1.390	56	1
21	10	111	70	27.5	0.141	40	1
22	2	111	60	26.2	0.343	23	0
23	5	158	84	39.4	0.395	29	1
24	4	83	86	29.3	0.317	34	0
25	1	124	60	35.8	0.514	21	0

Figure 1 Preprocessed Data

Used training data had six parameters, which were hamil, ogtt, diastolik, IMB, DPF and Usia. These parameters were used in classification process. The value of each parameter was used to determine diabetes diagnosis, where value of '1' means positive diabetes and '0' means negative diabetes

2.2 System Description

This study compared three classification methods, which were fuzzy KNN, C4.5 algorithm, and Naïve Bayes classifier. Classified data were generated from Indian Pimadatabase. This study designed a software for classification process and the results were used to determine the best method. Process of the study is displayed in Figure 2.

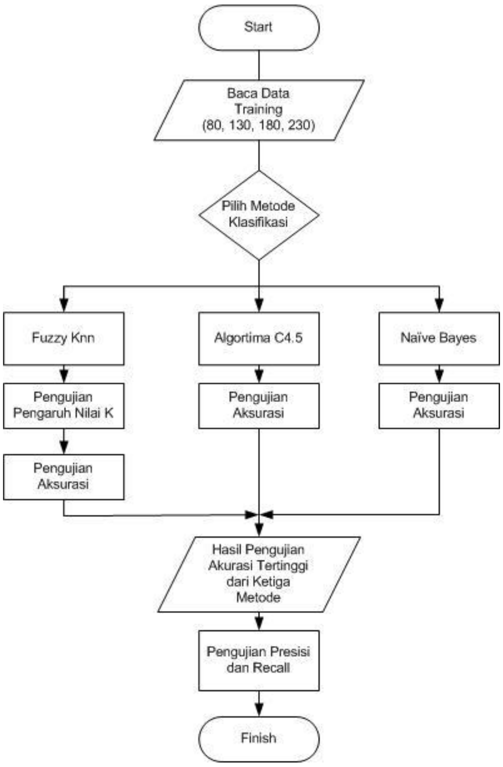

Start
Baca Data Training (80, 130, 180, 230)
Pilih Metode Klasifikasi
Fuzzy Knn
Algortima C4.5
Naïve Bayes
Pengujian Pengaruh Nilai K
Pengujian Aksurasi
Pengujian Aksurasi
Pengujian Aksurasi
Hasil Pengujian Akurasi Tertinggi dari Ketiga Metode
Pengujian Presisi dan Recall
Finish

Figure 2. Flowchart of the Study

2.3 Classification using Fuzzy KNN Method

In general, classification using fuzzy KNN method was conducted following these steps:

1. Input normalized training data.
2. Determine k value as initial parameter.
3. Determine weight exponent (m), this study used m = 2.
4. Calculate distance between new record data and each record training data using Euclidian distance.
5. Calculate membership value of each class, class with the highest membership value then used to determine new target.
6. Output was the result of the class with the highest membership value.

2.4 Classification using Fuzzy KNN Method

In Anyanwu journal, Podgorelec explains that C4.5 algorithm is a development of ID3 algorithm that is used in generating decision tree. C4.5 algorithmis not limited to binary number and is able to generate decision tree with multiple variables. Attributes in C4.5 algorithm generate one branch for every attribute branch in default [7]. Steps of classification process using decision tree C4.5 in general can be expressed as flowchart that is displayed in Figure 3.

Flowchart of decision tree C4.5 algorithm can be explained as:

1. Training data was required for classification process
2. Since data hamil, Ogtt, Diastolik, IMB, DPF and Usia were numerical data, early classification was done to minimize branch for further selection
3. Training data was required for classification process

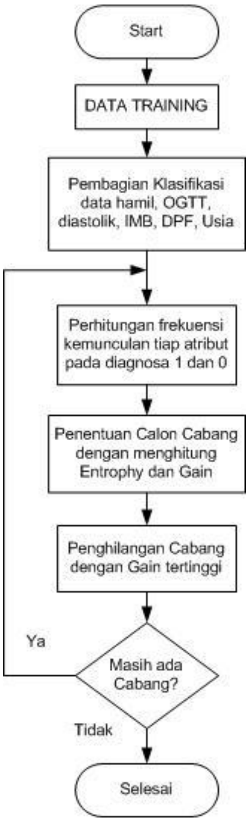

Start
DATA TRAINING
Pembagian Klasifikasi data hamil, OGTT, diastolik, IMB, DPF, Usia
Perhitungan frekuensi kemunculan tiap atribut pada diagnosa 1 dan 0
Penentuan Calon Cabang dengan menghitung Entrophy dan Gain
Penghilangan Cabang dengan Gain tertinggi
Ya
Masih ada Cabang?
Tidak
Selesai

Figure 3. Flowchart of C4.5 Algorithm

4. Frequency of occurrence of each data in positive and negative diabetes diagnosis was calculated.
5. Branch was determined by calculating entropy and gain according to aforementioned formula.
6. If initial branch/root had been determined, then the second branch was determined by removing parameter of the obtained branch. This process was repeated until there was no branch candidate.
7. If there was not any branch candidate, then the process was finished and decision tree had been generated.

2.4 Classification Using Naïve Bayes

Han (2006) explains that NBC uses Bayesian algorithm to calculate total probability. In NBC, probability of one word will be classified as one category(posterior probability),and it is based on the highest previous probability (prior probability).Naïve Bayes works by calculating the number of occurrence of specific attribute in particular category.

In Naïve Bayes with non-numerical data, probability of occurrence of specific category can be directly calculated, then it is multiplied with every attribute. However, with numerical data, this cannot be done as the data is continuous. For numerical data, the probability is calculated using Gaussian equation. Flowchart of Naïve Bayes is shown in Figure 4.

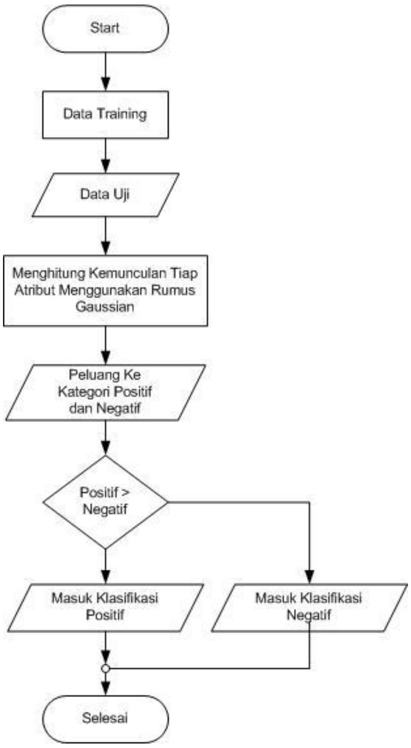
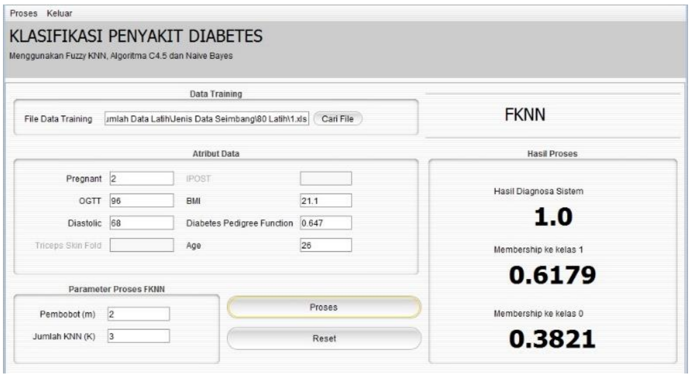

Start
Data Training
Data Uji
Menghitung Kemunculan Tiap Atribut Menggunakan Rumus Gaussian
Peluang Ke Kategori Positif dan Negatif
Positif > Negatif
Masuk Klasifikasi Positif
Masuk Klasifikasi Negatif
Selesai

Figure 4. Flowchart of Naïve Bayes Classifier

3. RESULT AND DISCUSSION

3.1 Classification Data Using Fuzzy Knn

When fuzzy KNN method was selected as feature process, then the diagnosis could be conducted in individual or collected data. Figure 5 shows classification using individual testing data.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Training

File Data Training: ımlah Data Latih\Jenis Data Seimbang\80 Latih\1.xls

FKNN

Atribut Data

Pregnant: 2
IPOST:
OGTT: 96
BMI: 21.1
Diastolic: 68
Diabetes Pedigree Function: 0.647
Triceps Skin Fold:
Age: 26

Parameter Proses FKNN

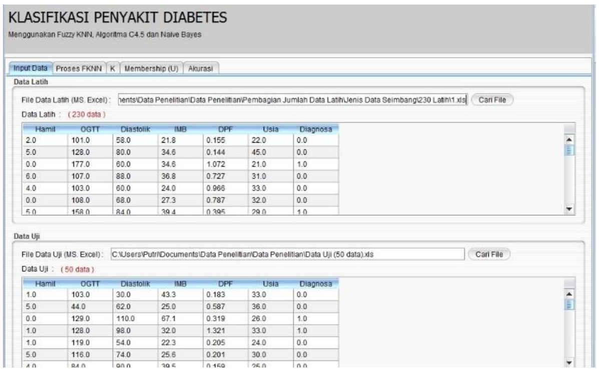
Pembobot (m): 2
Jumlah KNN (K): 3

Hasil Proses

Hasil Diagnosa Sistem
1.0
Membership ke kelas 1
0.6179
Membership ke kelas 0
0.3821

Figure 5.Display of Fuzzy KNN Classification using Individual Testing Data

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process FKNN, which was for entering the k and m parameter values and for clarification. Display of the clarification of the testing data collection is shown in Figure 6.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Latih

File Data Latih (MS. Excel) : nents\Data Penelitian\Data Penelitian\Pembagian Jumlah Data Latih\Jenis Data Seimbang\230 Latih\1.xls

Data Latih : (230 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
2.0	101.0	58.0	21.8	0.155	22.0	0.0
5.0	128.0	80.0	34.6	0.144	45.0	0.0
0.0	177.0	60.0	34.6	1.072	21.0	1.0
6.0	107.0	88.0	36.8	0.727	31.0	0.0
4.0	103.0	60.0	24.0	0.966	33.0	0.0
0.0	108.0	68.0	27.3	0.787	32.0	0.0
5.0	158.0	84.0	39.4	0.395	29.0	1.0

Data Uji

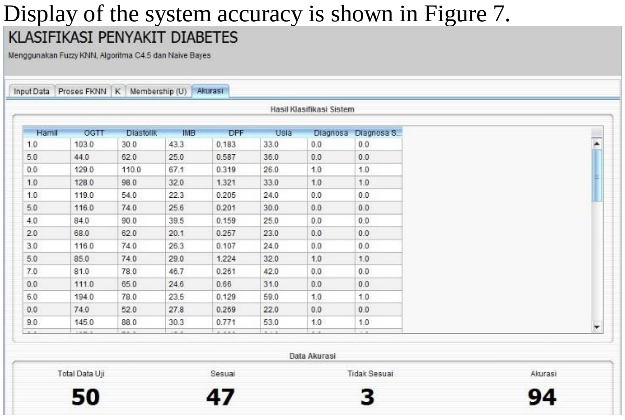
File Data Uji (MS. Excel) : C:\Users\Putri\Documents\Data Penelitian\Data Penelitian\Data Uji (50 data).xls

Data Uji : (50 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
1.0	103.0	30.0	43.3	0.183	33.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0

Figure 6 Display of the clarification of the testing data collection Fuzzy KNN

Feature 'k' was used to observed results in accordance with the number of k and feature 'membership' was used to observe the membership value, while feature 'accuracy' was used to calculate the system accuracy whether the results fit the previous theories. Display of the system accuracy is shown in Figure 7.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	0.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	1.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	0.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

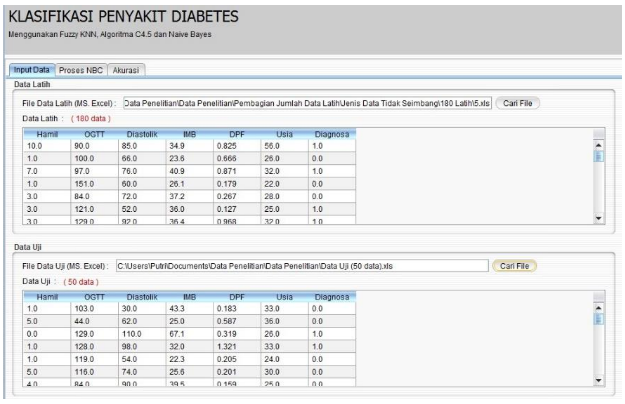
Data Akurasi

Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
50	47	3	94

Figure 7 Display of Feature 'System Accuracy FKNN'

3.2 Classification Data Using Naïve Bayes

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process NBC, which was for clarification process. Display of the clarification of the testing data collection is shown in Figure 8.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Data Latih

File Data Latih (MS. Excel) : Data Penelitian\Data Penelitian\Pembagian Jumlah Data Latih\Jenis Data Tidak Seimbang\180 Latih\5.xls

Data Latih : (180 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
10.0	90.0	85.0	34.9	0.825	56.0	1.0
1.0	100.0	66.0	23.6	0.666	26.0	0.0
7.0	97.0	76.0	40.9	0.871	32.0	1.0
1.0	151.0	60.0	26.1	0.179	22.0	0.0
3.0	84.0	72.0	37.2	0.267	28.0	0.0
3.0	121.0	52.0	36.0	0.127	25.0	1.0
3.0	129.0	92.0	36.4	0.968	32.0	1.0

Data Uji

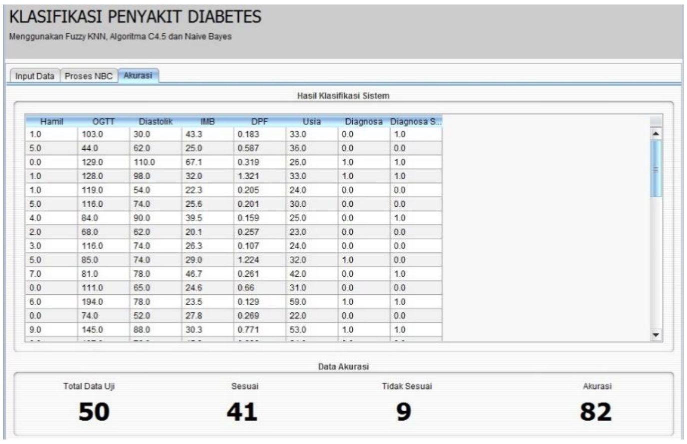
File Data Uji (MS. Excel) : C:\Users\Putri\Documents\Data Penelitian\Data Penelitian\Data Uji (50 data).xls

Data Uji : (50 data)

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa
1.0	103.0	30.0	43.3	0.183	33.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0

Figure 9 Input Data in Naïve Bayes

Feature 'Proses NBC' was used for clarification, while the feature 'accuracy' was used to calculate the level of system accuracy. The display of the system accuracy is shown in Figure 10.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	1.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	1.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	0.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	1.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

Data Akurasi

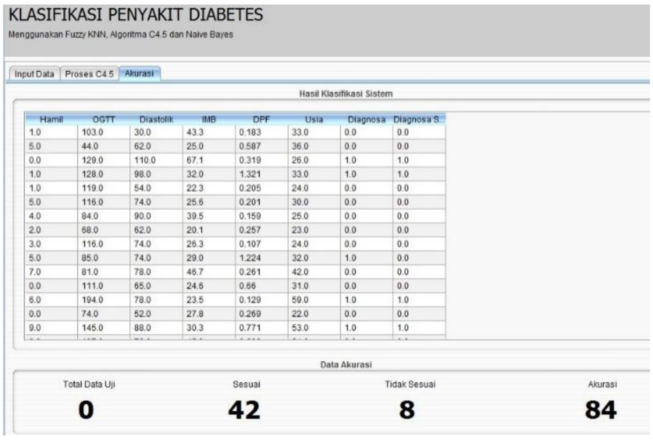
Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
50	41	9	82

Figure 10 Display of the System Accuracy of Naïve Bayes

3.3 Classification Data Using Decision Tree C4.5

There were several features in the classification using testing data collection: 1) input data, which was for entering the training data and testing data collection; 2) process C4.5, which was for clarification process. Display of the clarification of the testing data collection is shown in Figure 10.

Feature 'accuracy' was used to observe the level of accuracy of the algorithm C4.5 classification by the system and the results were compared with the previous theories. The display of the results of the accuracy of the algorithm C4.5 is shown in Figure 11.



KLASIFIKASI PENYAKIT DIABETES

Menggunakan Fuzzy KNN, Algoritma C4.5 dan Naive Bayes

Hasil Klasifikasi Sistem

Hamil	OGTT	Diastolik	IMB	DPF	Usia	Diagnosa	Diagnosa S...
1.0	103.0	30.0	43.3	0.183	33.0	0.0	0.0
5.0	44.0	62.0	25.0	0.587	36.0	0.0	0.0
0.0	129.0	110.0	67.1	0.319	26.0	1.0	1.0
1.0	128.0	98.0	32.0	1.321	33.0	1.0	1.0
1.0	119.0	54.0	22.3	0.205	24.0	0.0	0.0
5.0	116.0	74.0	25.6	0.201	30.0	0.0	0.0
4.0	84.0	90.0	39.5	0.159	25.0	0.0	0.0
2.0	68.0	62.0	20.1	0.257	23.0	0.0	0.0
3.0	116.0	74.0	26.3	0.107	24.0	0.0	0.0
5.0	85.0	74.0	29.0	1.224	32.0	1.0	0.0
7.0	81.0	78.0	46.7	0.261	42.0	0.0	0.0
0.0	111.0	65.0	24.6	0.66	31.0	0.0	0.0
6.0	194.0	78.0	23.5	0.129	59.0	1.0	1.0
0.0	74.0	52.0	27.8	0.269	22.0	0.0	0.0
9.0	145.0	88.0	30.3	0.771	53.0	1.0	1.0

Data Akurasi

Total Data Uji	Sesuai	Tidak Sesuai	Akurasi
0	42	8	84

Figure 11 Results of the Accuracy of the Algorithm C4.5

3.4 System Testing Method

The methods for system testing were:

1. Training Data Testing, the test was done with equal amount of the testing data, which was 50, but with various training data: 80, 120, 160, 200 data training.
2. Results of the weight exponent (m) in Fuzzy KNN. This was because m determined how much the distance weight between each neighbor to the membership value.

3. Duration test for the clarification process between fuzzy KNN, C4.5 algorithm and Naïve Bayes classifier
4. Accuracy test between fuzzy KNN, C4.5 algorithmandNaïve Bayes classifier. This test used accuracy formula.
5. Precision test amongfuzzy KNN, C4.5 algorithm and Naïve Bayes classifier
6. Recall test fuzzy KNN, C4.5 algorithm and Naïve Bayes classifier

3.5 Testing and Analysis Results in FKNN

The results of the test using fuzzy KNN method on the balanced and unbalanced training data are displayed in Table 1 and Table 2.

Table 1. Result of Fuzzy KNN Test On Balance Training Data

K	System Accuracy (%)				
	80 training data	130 training data	180 training data	230 training data	Average
2	70	84	86	88	82
4	80	86	92	94	88
6	92	92	92	94	92.5
8	90	90	94	94	92
10	90	96	96	96	94.5
12	92	96	98	98	96

Table 2. Result of Fuzzy KNN Test On Unbalance Training Data

K	System Accuracy (%)				
	80 training data	130 training data	180 training data	230 training data	Average
2	76	82	82	86	81.5
4	84	86	86	94	87.5
6	84	86	86	94	87.5
8	84	94	96	96	92.5
10	88	94	94	96	93
12	92	96	98	98	96

The results shows that the more training data, the higher system accuracy. This means that as the number of training data increases, the number of record with distance near the predicted data class also increases, which in turn improves the accuracy.

The test results of balanced training data show that accuracy tended to increase, except for k=8 where it slightly decreased. Meanwhile, for 180 and 230 training data, all system accuracy increased from k=2 to k=12.

Test results of unbalanced training data show that for 80, 130, and 230 training data, all system accuracy increased from k=2 to k=12. Meanwhile for 180 training data, system accuracy slightly decreased on k=10.

Test results of both balanced and unbalanced training data show that the number of training data is directly proportional to system accuracy. The slight decrease in several tests was insignificant and system accuracy was tended to be stable.

3.6 Testing and Analysis Results in C4.5

This test was aimed to observe which type of training data generated the best results. Each training data was tested five times to obtain the best results.

The test results of balanced training data shows that the best data was obtained from the second experiment with respective average and maximum accuracy of 78% and 84% for 130 training data. Meanwhile, the test results of unbalanced training data shows that the best data was obtained from the third experiment with respective average and maximum accuracy of 79.5% and 86% for 80 training data. All test results are displayed in Table 3

Table 3a. Test Result of C4.5 Algorithm Test on Balance Training Data

Training Data	System Accuracy (%)					
	Balanced Training Data					
	1	2	3	4	5	Avg
80	70	80	80	88	80	79.6
130	72	84	78	78	80	78.4
180	74	74	74	68	62	70.4
230	74	74	74	74	74	74
Avg	72.5	78	76.5	77	74	**75.6**

Table 3b. Test Result of C4.5 Algorithm Test on Unbalance Training Data

Training Data	System Accuracy (%)					
	Unbalanced Training Data					
	1	2	3	4	5	Avg
80	80	86	86	80	82	82.8
130	82	84	84	80	80	82
180	62	62	74	74	74	69.2
230	74	74	74	74	74	74
Avg	74.5	76.5	79.5	77	77.5	**77**

Each training data was tested five times to observe the results change, then average of system accuracy was taken. The results showed that the number of training data is inversely proportional to system accuracy. This probably caused by the increasing number of training data makes it more difficult to generate decision tree.

The test of this study was done on two type training data. Balanced training data means that the diagnosis was evenly distributed on positive and negative results. Meanwhile, in unbalanced training data, the diagnosis was random, which means that there was no record of the number of positive and negative results. In this test, average of system accuracy from two types of training data was calculated. The average results are shown in Table 4.

Table 4. Average Accuracy of Training Data Type

Number of Training Data	Balanced Training Data (%)	Unbalanced TrainingData (%)
80	79.6	82.8
130	78.4	82
180	70.4	69.2
230	74	74
Average	**75.6**	**77**

The results show that unbalanced training data had better accuracy with 77% compared with balanced training data (75.6%).

3. 7 Testing and Analysis Results in NBC

This test was aimed to observe which type of training data generated the best results. Each training data was tested five times to obtain the best results. The results can be seen at Table 5.

Table 5a. Test Result of Classification Using Naïve Bayes

Training Data	**System Accuracy (%)**					
	Balanced Training Data					
	1	**2**	**3**	**4**	**5**	**Average**
80	90	80	86	84	80	84
130	86	88	88	84	82	85.6
180	86	86	90	82	86	86
230	86	86	86	86	86	86
Average	87	85	87.5	84	83.5	**85.4**

Table 5b. Test Result of Classification Using Naïve Bayes

Training Data	**System Accuracy (%)**					
	Unbalanced Training Data					
	1	**2**	**3**	**4**	**5**	**Average**
80	84	86	80	86	86	84.4
130	84	84	84	88	86	85.7
180	86	86	86	90	84	86.4
230	86	86	86	86	86	86
Average	85	85.5	84	87.5	85.5	**85.5**

The test results of balanced training data shows that the best data was obtained from the third experiment with respective average and maximum accuracy of 87.5% and 90% for 180 training data. Meanwhile, the test results of unbalanced training data shows that the best data was obtained from the fourth experiment with respective average and maximum accuracy of 87.5% and 90% for 180 training data. The results show that the number of training data is directly proportional to system accuracy.

The test of this study was done on two type training data. Balanced training data means that the diagnosis was evenly distributed on positive and negative results. Meanwhile, in unbalanced training data, the diagnosis was random, which means that there was no record of the number of positive and negative results. In this test, average of system accuracy from two types of training data was calculated. The average results are shown in Table 6.

Table 6. Average Accuracy of Training Data Type

Number of Training Data	**Balanced Training Data (%)**	**Unbalanced Training Data (%)**
80	84	84.4
130	85.6	85.7
180	86	86.4
230	86	86
Average	**85.4**	**85.5**

As can be seen from Table 6, there was no significant difference between both data types with the accuracy difference only 0.01%, with unbalanced training data had slightly higher accuracy than balanced training data.

3. 8 Testing and Analysis Results in Naïve Bayes

In the previous tests, accuracy of fuzzy KNN, C4.5 algorithm, and Naïve Bayes had been tested in detail. From the test result, the best accuracy of each method was compared with other methods without considering training data type. The comparison result of accuracy of all methods is displayed in Table 7.

Table 7. Comparison of Accuracy of The Three Methods

Training Data	**Accuracy (%)**		
	Fuzzy KNN	**C4.5 Algorithm**	**Naïve Bayes**
80	92	86	86
130	96	84	88
180	98	74	90
230	98	74	86
Average	**96**	**79.5**	**87.5**

The result shows that fuzzy KNN method had the highest accuracy with 96%.

4. CONCLUSSION

From the results, it can be concluded that:

1. The system was able to classify DM diagnosis using fuzzy KNN, C4.5 algorithm, or Naïve Bayes with average accuracy of all methods was 88.5%.
2. In classification using fuzzy KNN method, the highest accuracy was obtained in 180 training data and k=12, with accuracy of 98% and average accuracy of all training data of 96%. The results show that the more training data, the higher system accuracy. This means that as the number of training data increases, the number of record with distance near the predicted data class also increases, which in turn improves the accuracy.
3. In classification using C4.5 algorithm, the highest accuracy was obtained in 80 training data, with accuracy of 86% and average accuracy of all training data of 79.5%.
4. In classification using Naïve Bayes method, the highest accuracy was obtained in 180 training data, with accuracy of 90% and average accuracy of all training data of 87.5%. The results show that the more training data, the higher system accuracy.
5. Based on the results of accuracy test, the best classification method was fuzzy KNN with average accuracy of 96%, followed by Naïve Bayes method with 87.5%, and lastly C4.5 algorithm with average accuracy of 79.5%.
6. Based on the results of precision and recall test, the best classification method was fuzzy KNNwith precision and recall of 0.94 and 1, respectively. This result shows that the accuracy is directly proportional to precision and recall.

5. REFERENCES

[1] Brunner and Suddarth. 2002. *Buku Ajar Keperawatan Medikal Bedah*, edisi 8 volume 2. Jakarta : EGC.

[2] Keller, James. 1985. *A Fuzzy K-Nearest Neighbor*. IEEE vol. SMC-15, No. 4

[3] Kusrini, 2007. *Design and implementation of building decision tree using C4.5 algorithm.* Proceedings of SEAMS-GMU Conference 2007.

[4] Larrisa Navia Rani, 2015. *Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit*. Jurnal Kom Tek Info Fakultas Ilmu Komputer Volume 2 No. 2. ISSN : 2356-0010

[5] Li D, Deogun JS, Wang K (2007) Gene Function Classification Using Fuzzy K-Nearest Neighbor Approach.

[6] Manning, D. Cristopher, Prabakhar Raghavan dan Hinrich Schutze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

[7] Maimon, O. dan Last, M. 2000. *Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology*. Dordrecht: Kluwer Academic.

[8] Mistra. 2005. 3 *Jurus Melawan Diabetes Mellitus*. Jakarta : Puspa Swara.

[9] Purnana, Parida. Dan Supriyatno, Catur.2013 *Deteksi Penyakit DiabetesType II denganNaive Bayes Berbasis Particle Swarm Optimization.* Jurnal Teknologi Informasi Volume 9 No.2 ISSN 1414-9999.

[10] Sunjana, 2010, *Aplikasi Mining Data Mahasiswa dengan Metode Klasifikasi Decision Tree*, Seminar Nasional Aplikasi Teknologi Informasi, Vol 7 pp. 24-29.

[11] Tandra, Hans. 2009. *Osteoporosis Mengenal, Mengatasi, dan Mencegah Tulang Keropos*. Jakarta: Gramedia Pustaka Utama.

[12] Yanita, Selly, ridho, ahmad. & lailil. 2013. *Perbandingan K-Nearest Neighbor dan Fuzzy K-Nearest Neighbor pada Diagnosis Penyakit Diabetes Melitus.*Jurnal Doro Volume 2 no.10

[13] Han, J. and Kamber, M., 2006. Data Mining: Concepts and Techniques, University of Illinois at Urbana-Champaign