CLASSIFICATION OF DIABETES DISEASE USING *NAIVE BAYES* Case Study : SITI KHADIJAH HOSPITAL

[1]Ida Lailatul Qurnia, [2]Eko Prasetyo, [3]Rifki Fahrial Zainal

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Bhayangkara
Jl. A. Yani 114 Wonocolo Surabaya 60231
Email:[1]IdaLailatulQurnia@ubhara.ac.id, [2]eko@ubhara.ac.id, [3]rifki@ubhara.ac.id

ABSTRACT

Less knowledge about symptoms and how to treat the disease of diabetes mellitus as well as a number of specialist diabetes mellitus which is still limited is one of the causes of the growing number of people affected by the disease. Diabetes disease classification system development aims to predict the type of diabetes patient or user who already suffer from diabetes mellitus. Therefore this system is made to diagnose the type of diabetes through laboratory test results, namely in the form of gender, age, disease history, family history, systolic, diastolic tensi tensi, temperature, pulse, blood sugar, fasting blood sugar JPP and Random blood sugar. That is by using the method of naive bayes as a method to process data on the patient's diagnosis. Test results of this system indicates that the system is able to predict the type of diabetes in patients, from the amount of data as much as 200 patient data, with an output that is the form of Diabetes Without Complications, Diabetes Type II and Normal but obtained the lowest accuracy rating of 39% and the value of the highest accuracy of 80%.

Keyword: *Classification, Naive Bayes, Diabetes Mellitus, Random Blood Sugar, A History Of The Disease In The Past*

1. INTRODUCTION

Diabetes is a disease in which the body cannot produce insulin (blood sugar balance hormones) or the insulin produced is insufficient insulin or don't work well. Therefore will cause increased blood sugars while review. Therefore, this study aims to help patients in order to know the problems of early diagnosis of the disease diabetes mellitus, so patients can find out his condition is being affected by the disease of diabetes, that is immediately checked himself into the hospital to get the handling medically. The methods used for classification of diabetes is naive bayes method that is one of the algorithms found in the technique of classification. Naive Bayes classification is by the method of probability and statistics. This method aims to conduct classification data on a particular class. Based on fact, the naive bayes algorithm will be applied in this study to determine the type of diabetes in patients who are already affected by diabetes namely diabetic without complications or type II diabetes (insulin), using a predefined data include gender, i.e. the value of blood sugar (limited to the value of blood sugar blood sugar values, JPP fasting and random blood sugar value), age, systolic tensi tensi, diastolic, temperature, pulse, a history of the disease in the past and family history as input variables of the system.

2. RESEARCH METHODS

This research uses Bayes algorithm is naive. Naive Bayes algorithm is one of the algorithms found in the technique of classification. Naive Bayes classification is by the method of probability and statistics. Naive Bayes algorithm is a classification algorithm techniques that are easy to apply and quick process

3. ANALYSIS AND SYSTEM DESIGN

As for the several stages in conducting classification with naive bayes algorithm diabetes i.e. enter data for calculating each training class. In table 4.1 there are eleven feature to be diagnosed to know types of diabetes in patients, namely gender, age, disease history, family history, systolic, diastolic tensi tensi, temperature, pulse, blood sugar, fasting blood sugar JPP and Random blood sugar. And there are three classes of output that would result in mind i.e. Diabetes Without Complications, Diabetes Type II (insulin) and Normal. In this study, the patients against the predictions do not yet know the diabetes class has namely criteria gender male, age 30 years, a history of the disease in the past instead of DM, no family history, tensi tensi 100 systolic, diastolic 70, temperature of 36.5 ° C, 67x/min pulse, blood sugar blood sugar 120 jpp, fasting and random blood sugar is 80 to 90, with training data data 15 hospital patients Siti Khadijah.

1. patients at the hospital Siti Khadijah

no	name	gender	age	history of the disease in the past	family history	systolic	diastolic	the temperature	nadi	blood sugar	fasting blood sugar	random blood sugar	description
1	Kartinah	P	38	DM	no	130	100	37	96	302	226	256	Diabetes Without Comp lications
2	Nur Hajari	P	62	DM	yes	150	90	37,5	100	202	106	110	Diabetes Type II
3	Moch Abadah	L	59	Tidak Ada	no	220	100	37,5	98	133	267	118	Diabetes Type II
4	Sugamin	L	61	DM	yes	90	50	37	117	190	128	166	Diabetes Type II
5	Sugiyati	P	66	DM	no	160	100	35	86	257	222	262	Diabetes Without Comp lications
6	Iyem Suhartini	P	59	Bukan DM	yes	100	60	36	60	162	125	106	Diabetes Without Comp lications
7	Sukarsih	P	67	DM	no	110	70	36	100	221	163	288	Diabetes Without Comp lications
8	Juliah	P	65	Tidak Ada	yes	120	80	36	80	209	122	101	Diabetes Without Comp lications
9	Sukirwadi	L	61	Bukan DM	no	100	72	37,5	65	130	100	70	Normal
10	Nur Charanah	P	62	Tidak Ada	yes	110	70	37	86	110	85	125	Normal
11	Amat Saman	L	75	Bukan DM	no	90	60	36	70	105	95	100	Normal
12	Aminah	P	67	Tidak Ada	yes	105	65	36	96	120	110	92	Normal
13	Salamah	P	65	DM	no	160	90	37,5	120	278	262	272	Diabetes Type II
14	Maria Ulfa	P	27	Tidak Ada	yes	110	70	36	96	180	110	110	Diabetes Type II
15	Rusdi	L	65	Tidak Ada	no	115	75	36,5	76	135	100	85	Normal
16	Zahrah	P	30	Bukan DM	yes	100	70	36,5	67	120	80	90	7

1) Perform calculations of the number of each class
   P (H) = (Number Of Each Class)/(Number Of Overall Grade)
   P (Diabetes Without Complications) = 0.3333
   P (Type II Diabetes) = 0.3333
   P (Normal) = 0.3333
2) Perform calculations and probability for each feature class
   Next calculate the number of features and the probability for each class, for only kategorikal data calculated based on how much the same data on the features in one class and then divided by the number of classes as for numerical data needed for the calculation of the mean average value – knowing the median, probability calculation variants and calculation features.
   Formula Gaussian distribution:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

   Gaussian distribution, which is used to calculate the probability of numerical data

Table 2. The Probability Of Each Feature

No.	class	Gender	age	History of Disease	Family history	Tensi/Mmhg		Tempe rature/ C	Nadi x/mt	Results Laboratorium		
						systolic	diastolic			Blood sugar	Fasting blood sugar	Rando m blood sugar
1	Dwc	1	0,0043	0,2	1	0,0071	0,0178	0,4434	0,0141	0,0008	0,0018	0,0019
2	DT II	0,6	0,0101	0	1	0,0052	0,019	0,4007	0,00009	0,003	0,0027	0,0027
3	Normal	0,4	0,00000000006	0,4	1	0,0381	0,0636	0,6049	0,0208	0,031	0,0062	0,0192

Next calculate the initial probability that is the multiplication of the values of the probabilities of each feature in each class, whereas to calculate the probability of the end that is the calculation of the probability multiplied by the beginning of class.

Table 3. Results of Probability beginning and end

No.	Class	early Probability	Final Probabiility
1	Diabetes Without Complications	1,99766E-18	6,65888E-19
2	Diabetes Type II	0	0
3	Normal	1,06103-21	3,53677E-22

The latter compares the results from each class of probability that is looking for the greatest value among the class of Diabetes Without Complications, Diabetes Type II class or classes Normal, due to the value found in the largest class of Diabetes Without Complications so the output is "Diabetes Without Complications".

4. RESULTS AND DISCUSSION

System testing is testing in entering data into the form – form that has been provided. At this stage of testing is done by randomly:

a. testing with training data and test data.
b. testing with training data 45 and test data45.
c. testing with training data 90 and test data 90.
d. testing with 150 training data and test data 60

Based on the test results of 60 test data obtained the results that there are 24 data corresponding to the actual class.

Table 4. The Results Of The Testing System

testing	Number Of correct preditictons			Number Of incorrect preditictons			average
	DWC	DT II	NORMAL	DWC	DT II	NORMAL	
Test I	5	5	2	0	0	3	15
Test II	13	9	0	2	6	15	45
Test III	15	20	0	15	10	30	90
Test IV	16	8	0	4	12	20	60

Testing I:
From training data and 15 ,15 test data obtained as a result of a correct prediction value of as many as 12 of the overall data IE 5 data from the class of Diabetes Without Complications, 5 data from class of Diabetes Type II and 2 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 80% while the rate of error of 20%.

Testing II:
From training data and 45, 45 test data obtained as a result of a correct prediction value of as much as 22 of the total data, namely data from the 13th class of DiabetesWithout Complications, 9 data from class of Diabetes Type II and 0 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 49% and 51% error rate.

Testing III:
From training data and the 90-90 test data obtained as a result of a correct prediction value of as much as 35 of the overall data i.e. 15 data from classes of Diabetes Without Complications, 20 data from the class of Diabetes Type II and 0 data from Normal classes. From the calculations it can be concluded that the system has an accuracy of 39% and the rate of error of 61%.

Testing IV:
From training data and 150 60 test data obtained as a result of a correct predictionvalue of as much as 24 of the overall data i.e. 16 data from classes of Diabetes Without Complications, 8 data from classroom Diabetes Type II and 0 data from Normalclasses. From the calculations it can be concluded that the system has an accuracy of 40% and the rate of errors by 60%.

For performance assessment process is carried out by the concept of precision that is the metric for measuring system performance in obtaining relevant data and metrics to measure i.e. recall system performance in obtaining relevant data unreadable (E.Prasetyo, 2014).

Table 5. Performance Appraisal System

	perfomance	**DWC**	**DT II**	**Normal**
Testing I	Precision	1	0,625	1
	Recall	1	1	0,4
Testing II	Precision	0,52	0,45	0
	Recall	0,8667	0,6	0
Testing III	Precision	0,375	0,4	0
	Recall	0,5	0,6667	0
Testing IV	Precision	0,3333	0,6667	0
	Recall	0,8	0,4	0

Process calculation of precision do with TP (True Positive) divided by the TP (True Positive) plus FP (False Positive). Whereas the calculation of the recall is done by means of TP (True Positive) divided by the TP (True Positive) plus FN (False Negative).

4. CONCLUSION

Of such research gives the conclusion that the system can help to classify the type of diabetes based on the results of the laboratory examinations on patients affected by diabetes mellitus. That is by using the method of naive bayes classification, but in a working model built have insufficient performance when dujikan with larger training data, because of

some performance testing shows the lowest accuracy of 39% and the highest accuracy of 80%, then need to add the features – features more in order to get better results

REFERENCES

[1] Andriani, Anik.(2013), Sistem *Prediksi Penyakit Diabetes Berbasis Decision Tree*, AMIK BSI. Jakarta.
[2] Budiansyah, Ari.(2010), *Sistem Pakar Diagnosa Penyakit Melalui Telapak Tangan Dan Lidah Berbasis Web Di Klinik BRC Bandung*, Universitas Komputer Indonesia. Bandung
[3] Bustami.(2014). *Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi.*Universitas Malikussaleh. Aceh Utara.
[4] Hermawati, Fajar Astuti.(2009). *DATA MINING*, Andi : Yogyakarta.
[5] Kusumadewi, Sri.(2009), *Klasifikasi Status Gizi menggunakan Naive Bayesian Classification*, Universitas Islam Indonesia. Yogyakarta.
[6] Nurkhozin, Agus.(2011), *Komparasi Hasil Klasifikasi Penyakit Diabetes Mellitus Menggunakan Jaringan Syaraf Tiruan Backpropagation dan Learning Vector Quantization*, Universitas Negeri Yogyakarta. Yogyakarta
[7] Otok, Bambang W.(2012), *Boosting Neural Network dan Boosting Cart Pada Klasifikasi Diabetes Militus Tipe I*, Institut Teknologi Sepuluh November. Surabaya.
[8] Prasetyo, E.(2012). *DATA MINING – Konsep dan Aplikasi Menggunakan Matlab*, Andi : Yogyakarta.
[9] Prasetyo, E.(2014). *DATA MINING – Mengolah Data Menjadi Informasi Menggunakan Matlab*, Andi : Yogyakarta.
[10] Rozaq, Abdur.(2011), *Klasifikasi Dokumen Berrita Pada Portal Berbahasa Inggris Dengan Menggunakan Algoritma Naïve Bayes*, Institut Teknologi Sepuluh November. Surabaya
[11] Soeparman.(1991). *ILMU PENYAKIT DALAM – Jilid I Edisi Kedua*, Balai Penerbit FKUI : Jakarta.
[12] Sugiharto, Grawas.(2008), *Klasifikasi Dokumen teks Berbahasa Arab Menggunakan Algoritma Naive Bayes*, Institut Teknologi Sepuluh November. Surabaya.
[13] Tampubolon, Mariani Valentina. (2010), *Sistem Pendukung Keputusan Penentuan Penyakit Diabetes Mellitus Dengan Metode Sugeno*, Universitas Sumatera Utara. Medan
[14] WHO.(2010). *Global Status Report on noncommunicable Disease*. Switzerland: WHO Press.