IMPLEMENTATION OF NAIVE BAYES METHOD IN CLASSIFICATION OF BREAST CANCER DISEASE

[1]ALAMSYAH, [2]EKO PRASETYO, [3]R. DIMAS ADTYO

[123]Informatics Engineering Program, Faculty of Engineering, Bhayangkara University

Jl. A. Yani 114 Surabaya 60231

Email: [1]acahster@gmail.com, [2]eko@ubhara.ac.id, [3]dimas@ubhara.ac.id

ABSTRACT

Less knowledge of early symptoms of breast cancer and how to deal with it early and the number of specialist doctors who are still limited is one factor contributors because of the increasing number of people affected by breast cancer disease. The development of breast cancer disease classification system aims to predict the early diagnosis of breast cancer disease in users or patients into two categories of malignant or benign. The initial diagnoses of this system prediction variable include Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size (Single Epithelial Cell) Size), Bare nuclei, Bland Chromatin, Normal nucleoli, Mitosis Using the naive bayes method to process diagnostic data in patients, the results of this system test show that the system is able to predict and classify breast cancer disease into two categories (malignant or benign) with the amount of data testing of 500 data. With the output of malignant or YA and benign or NO, the system is able to predict with an accuracy value of 98%.

Keywords: *Classification, Naive Bayes, Breast Cancer.*

1. INTRODUCTION

Breast cancer or often referred to as Breast Cancer is a malignant tumor derived from cells found in the breast. The breasts consist of lobules, ducts, fat and connective tissues, blood vessels and lymphs. In general, Breast Cancer comes from cells in the ducts, some of which come from the lobules and other tissues. Statistically the risk of Breast Cancer is increased in women nullipara, early menarche, late menopause and in women who have first child pregnancy over the age of 30 years. Less than 1% of breast cancer occurs at the age of less than 25 years, after age over 39 years incident increases rapidly, the highest incidence is found at the age of 45-50 years. Various studies have been done related to Breast Cancer case, among others, research using binary logistic regression method in breast cancer case by Darsyah, 2013 on breast cancer based on mammography result using SVM, the accuracy of classification obtained from SVM model is 99%. Sivaramakrisma's research, et al 2000 compared the performance of mammographic improvement algorithms. For microcalsification, adaptive neighborhood contrast improvement algorithm is best by 49%.

2. RESEARCH METHODS

This study uses Naive Bayes Algorithm. Naive Bayes algorithm is one of the algorithms found in the classification technique. Naive Bayes is a classification with probability and statistic methods. The Naive Bayes algorithm is also one of the classification algorithms that is easy to implement and has fast processing.

3. ANALYSIS AND DESIGN SYSTEM

As for several stages in the classification of breast cancer with naive bayes algorithm is to enter training data to perform the calculation of each class. There are 9 variables that will be diagnosed to know the type of breast cancer in patients, namely Clump Thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single Epithelial cell size, Bare nuclei, Bland chromatin, Normal nucleoli, Mitoses and there are 2 classes The outputs to be identified are Malignant / Dangerous and Benign. Predicted patients with unknown class of breast cancer who have variables include Clump Thickness 1, Uniformity of cell size 3, Uniformity of cell shape 2, Marginal Adhesion 5, Single Epithelial cell size 7, Bare nuclei 1, Bland chromatin 4 , Normal nucleoli 2, Mitoses 7 with train data 15 patient data.

Tabel 1. Data Training

NO	DATA	CLUMP THICKNESS	CELL SIZE UNIFORMITY	CELL SHAPE UNIFORMITY	MARGINAL ADHESION	SINGLE EPI CELL SIZE	BARE NUCLEI	BLAND CHROMATIN	NORMAL NUCLEOLI	MITOSES	CLASS
1	data_61	5	3	5	5	3	3	4	10	1	malignant
2	data_62	1	1	1	1	2	2	2	1	1	benign
3	data_63	9	10	10	1	10	8	3	3	1	malignant
4	data_64	6	3	4	1	5	2	3	9	1	malignant
5	data_65	1	1	1	1	2	1	2	1	1	benign
6	data_66	10	4	2	1	3	2	4	3	10	malignant
7	data_67	4	1	1	1	2	1	3	1	1	benign
8	data_68	5	3	4	1	8	10	4	9	1	malignant
9	data_69	8	3	8	3	4	9	8	9	8	malignant
10	data_70	1	1	1	1	2	1	3	2	1	benign
11	data_71	5	1	3	1	2	1	2	1	1	benign
12	data_72	6	10	2	8	10	2	7	8	10	malignant
13	data_73	1	3	3	2	2	1	7	2	1	benign
14	data_74	9	4	5	10	6	10	4	8	1	malignant
15	data_75	10	6	4	1	3	4	3	2	3	malignant
16	data_xx	1	3	2	5	7	1	4	2	7	?

Calculating the number of each class

P (H) = (Number of Classes) / (Total Class)

P (Classification Description | Benign): 6/15 = 0.400

P (Class Description Malignant): 9/15 = 0.600

Perform feature and probability calculations on each class

Next calculate the number of features and probabilities in each class, where the categorical data is calculated based on how much the same amount of data in features within a class then divided by the number of classes.

Gaussian Distribution Formula:

$$P(X_i = x_i | H = h_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

The Gaussian distribution formula is used to calculate the probability of data in numerical form.

Table 2. Probability of each feature

NO	CLASS	CLUMP THICKNESS	CELL SIZE UNIFORMITY	CELL SHAPE UNIFORMITY	MARGINAL ADHESION	SINGLE EPI CELL SIZE	BARE NUCLEI	BLAND CHROMATIN	NORMAL NUCLEOLI	MITOSES
1	Benign	0.177	0.06	0.367	5.023	0	0.9	0.187	0.336	0
2	Malignant	0.001	5.676	3.57	0.104	0.125	0.05	0.107	1.04	0.074

Next calculate the initial probability of multiplying the probability value of each feature in each class, while to calculate the final probability of class calculation times the initial probability.

Table 3. Preliminary and Final Probability Results

NO	CLASS	PROBABILITY INITIAL	PROBABILITY END
1	Benign	0	0
2	Malignant	0.000000108	0.0000000648

Finally compare the result of each probability class that is looking for the biggest value between Malignant or Benign class, because the biggest value is in Malignant class then its output is "Malignant".

4. RESULT AND DISCUSSION

System testing is a test in entering data into the form - the form provided. At this stage the test is done by using 140 training data on the system. Based on the test results of 140 training data obtained results that there are 137 data in accordance with the actual class.

Table 4. System Test Results

Testing 140 Data		**Predicted System Results**	
		Malignant	**Benign**
Original Class	**Malignant**	47	1
	Benign	2	90
Total		140	

From 140 tested training data got correct predicted value as much as 137 from all data that is 47 data from Malignant class and 90 data from Benign class. From the calculation, it can be concluded that the system has accuracy of 98% with the error rate of 2%.

Table 5. Performance Assessment System

Testing	Performance	Benign	Malignant
140 Data	Precision	0.989	0.960
	Recall	0.978	0.979

The calculation process is done by using the formula:

Precision : $\frac{True\ Positive}{True\ Positive+False\ Positive}$

Recall : $\frac{True\ Positive}{True\ Positive+False\ Negstive}$

5. CONCLUSION

From the study concluded that this system can help the classification of breast cancer type based on the results of variable examination in patients affected by breast cancer disease by using the method of naive bayes in the process of classification. Based on the result of interconnection system test and functional testing of the program, it can be

concluded that the program is feasible to be overall and has diagnostic result with accuracy level on 96% precision malignant, 99% precision benign, malignant recall 98% and recall benign equal to 98%.

REFERENCES

[1] Bustami. (2014), *Penerapan Algoritma Naive Bayes Untuk Mengklasifikasii Data Nasabah Asuransi*. Universitas Malikussaleh. Aceh Utara.

[2] Fitriani, I. R. (2014), *Peningkatan Metode Naive Bayes Clasification Untuk Penentuan Tingkat Keganasan Kanker Payudara Menggunakan Particle Swarm Optimization.* Universitas Dian Nuswantoro. Jawa Tengah.

[3] Jatmika, W. (2009), *Deteksi Kanker Payudara Menggunakan Ekstraksi Fitur Statistical Pada Citra Mammogram Berbasis Jaringan Syaraf Tiruan.* Universitas Muria. Kudus.

[4] Karina, E. N dan Yamasari, Y. (2013), *Aplikasi Diagnosa Kanker Kandungan Dengan Menggunakan Metode Naive Bayes (Study Kasus : Rumah Sakit Islam Surabaya).*Universitas Negeri Surabaya. Surabaya

[5] Kusumadewi, Sri. (2009), *Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification.* Universitas Islam Indonesia. Yogyakarta.

[6] Kardinah. (2007), DEPKES – RI. *Kanker Payudara : Bagaimana menghindari berbagai ancaman*. Diakses tanggal 21 April 2016. Dari : http://www.depkes.go.id/kanker.html.

[7] Maulana, I. U dan Santoso, A. (2014), *Klasifikasi Kanker Payudara Menggunakan Decision Tree Dengan Algoritma Interative Dichotomizer-3.* Universitas Brawijaya. Malang.

[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] Sholihin, M. (2011), Klasifikasi Kanker Pada Citra Mammogram. Universitas Islam Lamongan. Lamongan.

[11] Wikipedia. (2011), KANKER PAYUDARA. *Definisi dan detil tentang kanker payudara*. Diakses tanggal 20 April 2016. Dari : http://id.wikipedia.org/wiki/Kanker_Payudara.html.

[12] Wirawan, I. M. A. (2014), *Sistem Fuzzy Pendukung Keputusan Untuk Diagnosa Kanker Payudara.*Universitas Pendidikan Ganesha. Singaraja.