Implementation Data Mining using Decision Tree Method-Algorithm C4.5 for Postpartum Depression Diagnosis

Aris Supriyanto[1,*], *Suryono* Suryono[2], and *Jatmiko* Endro Susesno[2]

[1]Magister Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang – Indonesia.
[2]Departement of Physics, Science and Mathematics Faculty, Diponegoro University, Semarang – Indonesia.

Abstract. Postpartum depression is a serious problem that needs to be addressed because it has negative effects on family, child welfare, cognitive, and mother child interactions. Diagnosis is done based on psychological condition, blood pressure, respiration, body temperature, and classification data extract by decision tree C4.5 algorithm method. Results of this study in the form of an online information system that can identify the level of depression more quickly and precisely. The results showed the greatest gain on the psychological variables of 0.57 node 1, blood pressure 0.54 node 2, body temperature 0,54 node 3, means that the three variables are more influential on the condition of depressed patients, and should be given priority treatment. Test results from 50 patients with 50 examinations showed 62% prevalence, 65.62% sensitivity, specificity 77.77%, negative predictive value of 56%, and positive predictive value 84%, and

Keywords: postpartum depression; decision tree; algorithm C4.5; online information system.

1 Introduction

Mining data becomes popular data and some technology information for large data. The challenge in the world today is, how to utilize the data on the surface, become new information and add to it. Web-based information systems will reduce the use of paper in the world of medicine and decisions in forest conservation to create an environment with good climate and oxygen. Therefore an algorithm is used that is able to find information needed in making the necessary decisions [1].

The classification method used in this study uses a decision tree algorithm. In this study made an information system that implements the decision tree classification method to diagnose the level of postpartum depression or also called postpartum depression. Diagnosis is done by detecting the general condition of the patient such as heart rate, pulse, breathing, body temperature and psychological condition of the patient. Based question of Edinburgh Postnatal Depression Scale (EPDS), 369 women 46.1% experienced an increase in depression score. Postpartum depression has a negative effect on a family, child welfare, cognitive interaction, socio-emotional mother, and child. Depression is increasingly recognized as a risky disease, disruption followed by chronic stress and this case requires special attention to be handled by health, psychologists and developed using information technology [2].

Base on World Health Organization (WHO), Depression disorder is ranked fourth and is expected to rise in second class by 2020. During the postpartum period, it currently affects about 10 to 20% of women susceptible to depression in the first year of childbirth. However, only 50% experience depression [3].

Women who experience postpartum depression are more likely to turn to professional help, they feel attracted to make women feel more comfortable discussing symptoms of depression [4]. Based on existing data, the pattern of a disease and the factors that influence the diagnosis can be identified with the help of data mining. To determine the classification of patients' depressive rate decisions, the examination data is tailored to the acquisition of expert knowledge to generate conclusions from the patient's health condition.

This study aims to apply data mining with the C4.5 algorithm to determine the level of postpartum depression in accordance with the categories and factors associated with postpartum and the appropriate resolution using the information system. This research is to improve the service of post-natal patients using online information systems. The use of information systems can improve organizational performance because the process can be done automatically by increasing economic benefits [5].

2 Methods

This information system is built using patient medical record data that is the result of examination of vital signs of the heartbeat, pulse, breathing, body temperature and patient psychological data. Psychological data were obtained from the EPDS questionnaire. Stages are done by determining decision variables, categories of data and testing, knowledge formation, mining with algorithm C.4.5, determine the results and evaluate the data to

* Corresponding author: arissupriyanto67@gmail.com

determine the performance results. In this study used the following methods.

2.1 Data Mining

Data mining is a technique for finding hidden patterns of dataset using statistical method approaches [6]. Data mining is done by extracting information from the dataset and converting it into understandable structures and then finding patterns in the data set. some of the techniques used to extract information are classification, estimation, grouping, prediction, and association [7]. Data mining is in the process of using certain techniques and calculations to produce data that has become new information that is useful and can produce added value.

The method used in this study uses a decision tree with C4.5 algorithm. C4.5 algorithm is a method to classify data, creates a threshold and then divides the list into attribute values that are above the Algortihm C4.5 can able to create a decision with a shape resembling the physical condition of the tree from the root until leaf. Once created, the calculation process is done by removing the useless part and replacing it with a leaf node [8].

Data mining in the process using certain techniques and calculations to produce data that has become useful new information. Data mining has a number of important techniques such as preprocessing by dividing data into data trending and data testing and then labeling the results of classification data. Classification is one grouping technique based on data classes or labels used for the analysis of category data. In this study, the object of research is the patient's pregnant mother. In this study will also discuss the algorithm approach C.4.5 data mining is used for the diagnosis of Postpartum Depression.

2.2. Decision Tree

Decision tree is a technique used to classify objects that have many types. This technique consists of a set of decision nodes that are drawn from each root tip until leaf [9]. Decision trees are able to eliminate unnecessary calculations, since the sample is tested only on the basis of certain criteria or classes. Flexible to select features from different internal nodes, the selected feature distinguishes a criterion over other criteria within the same node thus improving the quality of the resulting decision when compared to the more conventional one-stage counting method.

The process on the decision tree is to change the shape of the data table into a model tree. By using decision tree problems that are multicriteria can be simplified making it easier to make the decision. Decision tree process is done by changing the data table into a model tree. The structure of the decision tree generally starts from the root until leaf.

2.3 Algorithm C4.5

C4.5 algorithm uses entropy calculation to get the biggest weight. Suppose that variable X with probability, mean per symbol, is needed to send a value that represents the entropy value X. The average number of information needed has $-Log_2\ (p)$. For outcome variables, only use the number $-Log_2\ (pj)$ has the same weight as the probability of results, resulting in the following formula [10]:

$$H(X) = -\sum\ pj\ Log_2\ (ph) \quad (1)$$

H (X) is a set of cases, ph is a proportion. Suppose the candidate split *H*, which partitioned the training data set into several subsets, *T1, T2, T3, T4 to Tk*. The average information need can then be calculated as the sum of entropy for individual subset, as follows:

$$H_x\ (T) = \sum_{i=1}^{k} Pi\ Hs(Tx) \quad (2)$$

$H_S\ (T)$ is a feature, k number of partitions attribute A, Ti is proportion to T. The next stage calculates the value of gain and the result of the highest gain is used as the tree root. To calculate the gain value the formula is used:

$$Gain\ (S) = H(X) - Hx(T) \quad (3)$$

Gain (S) is the result of the calculation of H (X) minus Hx (T. At each decision node, C4.5 algorithm chooses optimal gain to split, which is the biggest gain.

2.4 Evaluation of Classification Performance

After getting the calculation results, we evaluate the results to measure the performance results. The probability of a patient who is tested positive has a disorder, whether it depends on the prevalence of the disorder. Sensitivity and specificity are presented in a more modern and clinically relevant concept of positive and negative test prediction values [11] with the following formula:

- Prevalence using formula $T_{disease}/\ Total \times 100$. The higher the prevalence of disease in the population, the higher the predictive positive value. Thus the main way to increase results in a screening is to target tests on a group of people at high risk of developing the disease.
- Sensitifity using formula *A/(A+C).* Sensitivity is able to show which individuals suffer from the pain of a whole sickly population.
- Predictive Positive Value using formula *A/(A+B) × 100.* The proportion of patients who test positive and actually suffer pain. If a person tests positive, this test able to show the probability that the patient is actually suffering from the disease.
- Predictive Negative Value using formula *D/(D+C) × 100.* The proportion of patients who tested was negative and did not really suffer pain. If a person's test is negative, this test is able to show the probability that the patient is really not suffering from the disease.

  Where:

  T number of sick patients
  A number of positive depression
  B number of positive not depression

C number of negative depression
D number of negative not depression

3 Implementation

The information system in this study, using patient medical record form the patient vital signs data and patient psychological data, psychological data from the EPDS questionnaire. In fig 1, the process that is carried out in determining the decision variables, data categories, mining with the C4.5 algorithm, analysis of training and testing data, tree formation, decision making, and classification evaluation. The final result is a website interface in the form of 3 important parts, the results of diagnosis, results of treatment decisions, and results of performance evaluation.

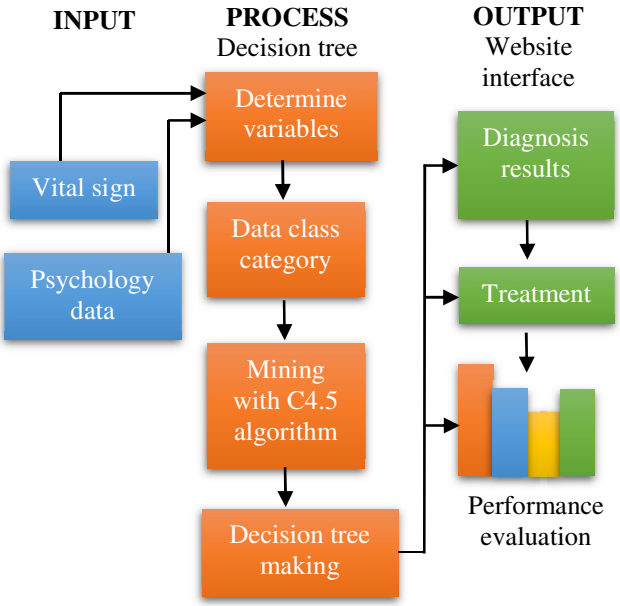

Fig. 1. Automatic diagnosis system using decision tree

This study uses medical record data from blood pressure, pulse, breathing, temperature, and psychology of maternity patients. From the results of the examination in table 1, there are 50 patients with a total of 50 examinations, there were 21 positive patients experiencing depression, and 4 patients were not depressed, then at a negative value, there were 11 depressed patients and 14 patients were not depressed. Overall, there were 31 depressed patients and 19 non-depressed patient history of patient.

Table 1. Positive and negative depression

Condition	Depression	Not Depression	Total
Positive	21	4	25
Negative	11	14	25
Total	31	19	50

Table 2 shows the number of depressed patients more than non-depressed patients. The next important step is to determine the effects of blood pressure, pulse, breathing, body temperature or psychological condition of the patient on depression. At node 1 the biggest gain on psychological variables with a value of 0.57. Node 2 the largest gain in the blood pressure variable with a value of 0.57. Node 3 gets the greatest gain at temperature, with a value of 0.54. Node 4 is the biggest gain on the blood pressure variable with a value of 0.54 from the value of the node that has been obtained at temperature, with a value of 0.54. Node 4 is the biggest gain on the blood pressure variable with a value of 0.54 from the value of the node that has been obtained.

Table 2. History of patient examination patient

Variable Node 1	Condition	Case	Depression	No
Total		50	32	18
blood pressure (mmHg)	low	8	3	5
	normal	27	14	13
	prahypertension	12	12	0
	hypertension (1)	3	3	0
	hyipertensi (2)	0	0	0
pulse (s/m)	low	1	0	1
	normal	47	30	17
	high	2	2	0
respiration (s/m)	low	0	0	0
	normal	28	12	16
	fast	13	12	1
	very fast	8	8	0
temperature	low	12	8	4
	normal	31	17	14
	high	6	6	0
	very high	1	1	0
psichology	normal	16	2	14
	medium	10	6	4
	heavy	20	20	0
	very heavy	4	4	0

This is a mathematical calculation with algorithm C4.5.
gain [pulse]=

$$0{,}94 - \left(\left(\frac{1}{50}\right) * 0\right) + \left(\frac{47}{50}\right) * 0{,}94 + \left(\frac{2}{50}\right) * 0 = 0{,}05$$

gain [respiration]=

$$0{,}9 - \left(\left(\frac{0}{50}\right) * 0\right) + \left(\frac{28}{50}\right) * 0{,}39 + \left(\frac{13}{50}\right) * 0{,}98 = 0{,}28$$

gain [temperature]=

$$0{,}95 - \left(\left(\frac{12}{50}\right) * 0{,}91\right) + \left(\frac{31}{50}\right) * 0{,}99 + \left(\frac{6}{50}\right) * 0 + \left(\frac{1}{50}\right) * 0 = 0{,}1$$

gain [psikology]=

$$0{,}95 - \left(\left(\frac{16}{50}\right) * 0{,}91\right) + \left(\frac{10}{50}\right) * 0{,}99 + \left(\frac{20}{50}\right) * 0 + \left(\frac{4}{50}\right) * 0 = 0{,}57$$

gain [blood preasure]= $0{,}54 - 0 = 0{,}54$
node 1 entropy [Total blood pessure prahipertensi] =

$$\left(\left(\frac{12}{12}\right) * \left(log_2\left(\frac{12}{12}\right)\right)\right) + \left(\left(\frac{0}{12}\right) * \left(log_2\left(\frac{0}{12}\right)\right)\right) = 0$$

gain [pulse] $= 0{,}54 - 0 = 0{,}54$
gain [respiration] $= 0{,}54 - 0 = 0{,}54$
gain [temperature]$= 0{,}54 - 0 = 0{,}54$
node 4 entropy [temperature- high] =

$$\left(\left(\frac{6}{6}\right) * \left(log_2\left(\frac{6}{6}\right)\right)\right) + \left(\left(\frac{0}{6}\right) * \left(log_2\left(\frac{0}{6}\right)\right)\right) = 0$$

gain [pulse] $= 0{,}54 - 0 = 0{,}54$
gain [respiration] $= 0{,}54 - 0 = 0{,}54$
gain [temperature]$= 0{,}54 - 0 = 0{,}54$
gain [blood preasure]$= 0{,}54 - 0 = 0{,}54$

These numbers realy heavly on patient data, the calculation are no longer done conventionally but whenthere are new data and new entities, the tree root will change automatically on the system information dashboard. When compared to the examination of data mining without the help of information system, detection of the priority status of public patients is no longer a priority but only random.

At node 1 after calculated using C 4.5 algorithm, the highest gain is in the psychological variable, so psychology is used as the first root node. This means that the patient's psychological condition should be the main concern. The root of the next tree is made to examine other influencing variables and the analysis is continued at node 2 of the normal psychological variable.

At node 2 after calculated using the C4.5 algorithm, the biggest gain is on the blood pressure variable. Blood pressure becomes node 2 in the examination. The root of the tree is still evolving to test other variables that affect the patient's depression.

At node 3 after calculated using the C4.5 algorithm, the gain value is the same, so one variable can be selected as the next root. The tree roots continue to check for other affecting variables and the analysis continues at node 4 of the high-temperature variable.

In fig 2, at nodes 3 and 4 have got the root tip so the calculation is stopped. So when the largest gain group from node 1 to end there is a psychology variable with a value of 0.57 unity of blood pressure with a gain value of 0.54 of the temperature variable with a value of 0.54 and a blood pressure variable value. with a gain value of 0.54. Fig 3 is the result of computing the entropy result and the acquisition of each variable in the same way until node 4 Decision trees were obtained from C4.5 calculations with a total of 50 patients. The entire examination of all patients is 50 times, resulting in the following decision tree scheme.

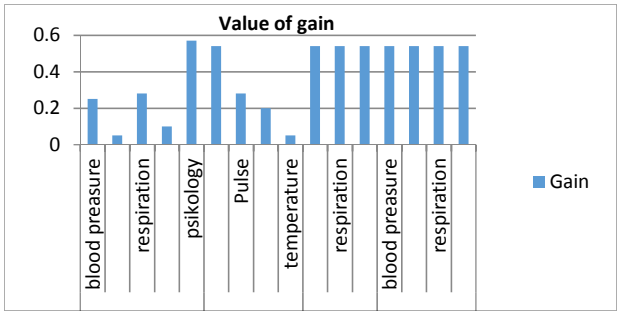

Fig. 2. The highest value of gain

This is a root calculating using Algortihm C4.5
IF Psycologi == Medium

- If Blood Preassure == Hypertension (1)
  return Depression
- if Blood Preassure == Normal
  return No
- if Blood Preassure == Low
  return No

if Psycology == BERAT return Depression
Array ([Blood Preassure] => Array ([Hypertensy (1)] => Array ([case] => Hypertensy(1) [Value] => Depression) [Normal] => Array ([case] => Normal [value] => No) [Low] => Array ([case] => Low [nilai] =>)))
if Psycology == NORMAL

- if Blood Preassure == Hypertension (1)
  return Depression
- if Blood Preassure == Normal
  return No
- if Blood Preassure == Low
  return No

if Psycology == Very heavy return Depression

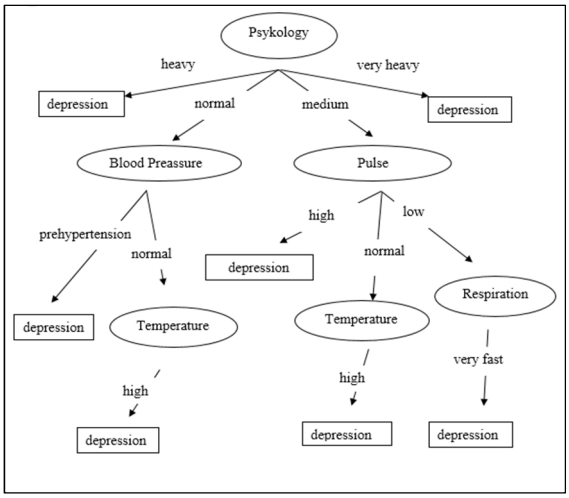

Fig. 3. Result of decession tree

Trees originate from psychological variables due to the largest gain value at node 1. Then the second node is the root of blood pressure and pulse because they have the greatest gain in nodes 2 and 3. The root ends at node 4 at the root of temperature and breath. That the tree describes the result of the variable with the largest gain on each node that has the most influence on the patient's depression. Table 6 is a condition from the calculation result, evaluation with the result of performance as follows.

If considering the results of calculations C4.5 examination of psychological conditions, blood pressure, temperature more get an action to reduce the probability of depression level. Predictive Positive Value of 84% and Predictive Negative value 56% indicates the number of patients who are more depressed than non-depressed, meaning that to improve the accuracy of the examination, 84% of patients with positive depression should be

examined, priority on psychological conditions, blood pressure, and body.

Table 6. Result of Performance Evaluation

Indicator	Calculation	Result
Prevalence	$\frac{31}{50}$ x 100	62%
Sensitifity	$\frac{21}{21+11}$ x 100	65.62 %
Spesifisity	$\frac{14}{14+4}$ x 100	77.77 %
Predictive Positive Value	$\frac{21}{21+4}$ x 100	84 %
Predictive Negative Value	$\frac{14}{14+11}$ x 100	56 %

4 Conclusion

The results showed a prevalence value of 62 %. Of the 50 people examined, 32 were depressed, 18 were not depressed. This figure shows the number of depressed patients more than those who are not depressed. The weights that have an effect on the patient based on the obtained gain are a psychological condition, blood pressure, and body temperature.

Based on the results of sensitivity, the test accuracy of 65.62%. The greater the sensitivity, the greater the influence of psychological variables, blood pressure, and body temperature to increase patient depression. Specificity shows the value of 77.77% in other words 14 people out of 18 people with negative results are completely negative and 4 positive people affected by depression. From negative patients is shown depression opportunities as many as 4 people.

Then among apositive patients, only 84% were completely depressed and for those who test negative 56% were not depressed. Based on prevalence percentage values, sensitivity, specificity, positive predictive value and positive predictive value addressed more depressed patients than non-depressed patients, even from patients with depressive depression suggesting depression. The results of this calculation will be the appropriate treatment performed on depression patients with priority over the psychology condition, blood pressure, and body temperature of the patient.

References

[1] A. Anguera, J. M. Barreiro, J. A. Lara, and D. Lizcano, Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry, *Comput. Struct. Biotechnol. J.*, vol. **14**, pp. 185–199, 2016.

[2] Hashima-E-Nasreen, M. Edhborg, M. Petzold, Y. Forsell, and Z. N. Kabir, Incidence and Risk Factor of Postpartum Depressive Symptoms in Women: A Population Based Prospective Cohort Study in a Rural District in Bangladesh, *J. Depress. Anxiety*, vol. **4**, no. 2, pp. 4–11, 2015.

[3] C. El-Hachem *et al.*, Early identification of women at risk of postpartum depression using the Edinburgh Postnatal Depression Scale (EPDS) in a sample of Lebanese women, *BMC Psychiatry*, vol. **14**, pp. 242–247, 2014.

[4] H. Woolhouse, S. Brown, A. Krastev, S. Perlen, and J. Gunn, Seeking help for anxiety and depression after childbirth: Results of the Maternal Health Study, *Arch. Womens. Ment. Health*, vol. **12**, no. 2, pp. 75–83, 2009.

[5] S. Suryono, J.E.Suseno, C.Mashuri, A. D. Sabila, J.A.M. Nugraha, M.H. Primasiwi, RFID Sensor for Automated Prediction of Reorder Point (ROP) Values in a Vendor Management Inventory (VMI) System Using Fuzzy Time Series, *American Scientific Publishers*., Vol. **23**, 2398–2400, 2017.

[6] C. Colak, E. Karaman, and M. G. Turtay, Application of knowledge discovery process on the prediction of stroke, *Comput Methods Programs Biomed*, vol. **119**, no. 3, pp. 181–185, 2015.

[7] V. Karthikeyani, I. P. Begum, K. Tajudin, and I. S. Begam, Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction, *Int. J. Comput. Appl.*, vol. **60**, no. 12, pp. 26–31, 2012.

[8] S. Sathyadevan and R. R. Nair, Comparative analysis of decision tree algorithms: Id3, c4.5 and random forest, *Smart Innov. Syst. Technol.*, vol. **31**, pp. 549–562, 2015.

[9] F. S. Khan, R. M. Anwer, O. Torgersson, and G. Falkman, Data mining in oral medicine using decision trees, *World Acad. Sci. Eng. Technol.*, vol. **37**, pp. 225–230, 2008.

[10] D. T. Larose, *Data Mining Methods and Models*. 2006.

[11] H. Faller, Sensitivity, specificity, positive and negative predictive value, *Rehabilitation (Stuttg).*, vol. **44**, no. 1, pp. 44–9, 2005.