Classification of diabetes disease using decision tree algorithm (C4.5)

B A C Permana¹*, R Ahmad², H Bahtiar³, A Sudianto¹ and I Gunawan¹

¹ Informatics Engineering Department, Faculty of Engineering, Hamzanwadi University, Lombok, Indonesia
² Information Management Department, Faculty of College of informatics, Chaoyang University of Technology, Taichung, Taiwan
³ Information Systems Department, Faculty of Engineering, Hamzanwadi University, Lombok, Indonesia

*andriska.cp@gmail.com

Abstract. Diabetes is one of the most common health problems in the world. Diabetes is also known as "the silent killer" because according to WHO (2016) diabetes increased from 108 million in 1980 to about 422 million adults had diabetes in 2014. If not handled properly, diabetes can become chronic and damage other organs and can cause death. This disease has several symptoms in the patient but evaluating the different factors or symptom variables required to determine which variables are more dominant. This research aims to establish the most influential variable of the many variables causing diabetes mess. We suggest using a data mining decision tree (C4.5) in this paper to forecast diabetes to help doctors analyse the disease sooner. Data mining has carried out various approaches to predict a disease, one of them is the use of c4.5. In this research, produce a decision tree and the result shown that polydipsia play a role in diabetes with accuracy 90.38 %. One of the most dominant signs of diabetics is the sign of polydipsia.

1. Introduction

According to data obtained from the WHO official website, there are around 422 million people around the world who have diabetes. Generally, people with diabetes are from low- and middle-income countries. However, in the last 3 decades, it has been found that the increasing number of diabetics is evenly distributed throughout the world in both low-income and high-income countries. Each year the death associated with diabetes reaches 1.6 million people. Diabetes is a condition in which the body has problems producing the hormone insulin which can cause damage to other organs in the body [1–3].

Research related to classification methods has been done before for disease using certain classification algorithms. According to the decision tree is the most powerful classification technique in a study conducted to classify diabetic patients in a population in Canada [4]. According to getting the best classification technique with best precision value equal to 0.770 and recall of 0.775 using the Hoeffding Tree algorithm [5]. According to diabetes detection using a deep learning algorithm because early detection of diabetes is believed to be very important, the use of a support vector...
machine (SVM) can improve performance by 0.03% and 0.06% by using the Convolutional Neural Network (CNN) [6].

Decision trees are well-established methods for classification tasks because of their fast construction time and good interpretability [7]. This study performs a classification technique using the C4.5 algorithm to obtain a decision tree [8], that will show the parameters that have the greatest influence on the occurrence of diabetes, by knowing the early symptoms diabetes can be treated early with the hope that later it can reduce the death rate for sufferers caused by diabetes.

We propose early detection and prevention of diabetes that will reduce the degree of complexity and high diabetes-related costs, and we use the decision tree (C4.5) classifier to apply it for knowledge analysis. As a result, well-timed diagnosis and avoidance lead to a decrease in mortality and risk of diabetes complications, resulting in improved quality of life.

2. Methods
The dataset used in this study is derived from secondary data obtained from the early-stage diabetes dataset which can be accessed through https://www.kaggle.com/singhakash/early-stage-diabetes-risk-prediction-datasets. The description of dataset is belonging to: This dataset has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The attribute of this dataset is: Age with range (1.20-65), Sex with (Male and Female), (Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Obesity) with condition (Yes and No), Class with (Positive and Negative) result, for more details, see Table 1.

The data contained in the source consists of 520 record with several variables (age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity) [9]. Decision tree (C4.5) is an algorithm that can generate a decision tree and classify an object. C4.5 is a better form of Iterative Dichotomised 3 (ID 3) algorithm [10]. This study performs a classification technique using the c4.5 algorithm to obtain a decision tree that will show the parameters that have the greatest role in diabetes [11].

The work steps as contained in Figure 1. The first stage of the approach we suggest is to collect data sets of symptoms of diabetes symptoms, public data at this stage. Furthermore, in the second stage, we pick the data by performing feature selection, taking only clinical symptoms that are considered diabetes indicators, which will then be used as attributes when implemented in the model. In the third stage, data that passed the feature selection is split into two data, namely training data and test data, where training data is used to test the model and testing data to test the model accuracy. In the next step, the training data will be processed using the C4.5 model; a model will be obtained from this stage which can show in Figure 2. where a detailed description of the C4.5 model results will be clarified when the C4.5 algorithm stage is implemented.
Fig. 1. Classification technique using the C4.5 algorithm to obtain a decision tree.

3. Results and discussion

3.1. Feature selection
Feature selection has been the subject of concern for a long time and much work have been done. With the development of large databases and consequent criteria for useful machine learning techniques, new problems emerge, and new approaches to selecting features are in demand [12]. The selection feature is a technique used to select the parameters that are considered the most influential on the class [13]. In this study, from 16 attributes, 13 attributes were selected by eliminating three attributes, namely age, gender and irritability.

3.2. The implementation of C4.5 algorithm
The following Table 1. Containing selected data sets and used as the basis for processing using the C4.5 algorithm:
Table 1. Containing selected data sets.

| No | A  | B  | C  | D  | E  | F  | G  | H  | I  | J  | K  | L  | M  | Class |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-------|
| 1  | No | Yes| No | Yes| No | Yes| No | Yes| No | Yes| Yes| Yes| Yes| Positive |
| 2  | No | No | No | Yes| No | No | Yes| No | Yes| No | Yes| No | No | Positive |
| 3  | Yes| No | Yes| Yes| No | No | Yes| No | Yes| No | Yes| No | No | Positive |
| 4  | No | No | Yes| Yes| Yes| No | No | Yes| No | No | No | No | No | Positive |
| 5  | Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Positive |
| 6  | Yes| Yes| No | Yes| Yes| Yes| Yes| No | Yes| Yes| No | No | No | Positive |
| 7  | Yes| Yes| Yes| Yes| Yes| Yes| Yes| No | Yes| Yes| No | No | No | Positive |
| 8  | Yes| Yes| Yes| Yes| Yes| Yes| Yes| No | No | No | No | No | No | Positive |
| 9  | Yes| Yes| Yes| Yes| Yes| Yes| Yes| No | No | No | No | No | No | Negative |
| 10 | Yes| Yes| Yes| No | Yes| Yes| Yes| No | Yes| Yes| No | No | No | Negative |
| 11 | Yes| Yes| Yes| Yes| Yes| Yes| Yes| No | No | No | No | No | No | Negative |
| 12 | ...| ..| ..| ..| ..| ..| ..| ..| ..| ..| ..| ..| ..| ..|
| 520| No | No | No | No | No | No | No | No | No | No | No | No | No | Negative |

Information: A= Polyuria; B= Polydipsia; C= Sudden weight loss; D= Weakness; E= Polyphagia; F= Genital thrush; G= Visual blurring; H= Itching; I= Delayed healing; J= Partial paresis; K= Muscle stiffness; L= Alopecia; M= Obesity

To get the root value in the decision tree, it is necessary to calculate the entropy value and the gain value of each parameter using the formula for finding the entropy value as listed in (1) and the gain value as contained in (2)[14]:

\[ \text{Entropy}(S) = \sum_{i=0}^{n} -p_i \log_2(p_i) \]  

\[ \text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \]  

The following is the calculation for the entropy value based on formula (1):

\[ \left( -\frac{320}{520} \log_2 \frac{320}{520} + \left( \frac{200}{520} \log_2 \frac{200}{520} \right) \right) = 0.96123660 \]

Furthermore, after getting the entropy value, the entropy value is used as the basis for calculating the gain value for each parameter. The calculation for the polyuria parameter values is as follows:

- Polyuria Positive:

  \[ \frac{\log \left( \frac{243}{320} \right) - \log \left( \frac{3}{64} \right) }{\log(2)} \]

  \[ \left( -\frac{243}{320} \log_2 \frac{243}{320} + \left( \frac{15}{320} \log_2 \frac{15}{320} \right) \right) = 0.50851453 \]

- Polyuria Negative:

  \[ \left( -\frac{77}{200} \log_2 \frac{77}{200} + \left( \frac{195}{200} \log_2 \frac{195}{200} \right) \right) = 0.63421093 \]

- Polyuria gain:

  \[ \left( 0.96123660 - \left( \frac{320}{520} \times 0.50851453 \right) + \left( \frac{200}{520} \times 0.63421093 \right) \right) = 0.404 \]

The same technique is applied to the other parameters in order to obtain a table of the results of the entropy value and gain value used to determine the first root of the decision tree as shown in Table 2:
Table 2. Entropy value and gain value.

| Parameter                  | All Case | Positive | Negative | Entropy | Gain  |
|----------------------------|----------|----------|----------|---------|-------|
| Total                      | 520      | 320      | 200      | 0.9612  |       |
| POLYURIA                   |          |          |          |         |       |
| YES                        | 243      | 15       | 228      | 0.509   | 0.404 |
| NO                         | 77       | 185      | 112      | 0.634   |       |
| GENITAL THRUSH             |          |          |          |         |       |
| YES                        | 83       | 33       | 50       | 0.843   | 0.415 |
| NO                         | 237      | 167      | 70       | 0.073   |       |
| VISUAL BLURRING            |          |          |          |         |       |
| YES                        | 175      | 58       | 117      | 0.923   | 0.128 |
| NO                         | 145      | 142      | 3    | 0.687   |       |
| ITCHING                    |          |          |          |         |       |
| YES                        | 154      | 99       | 55       | 1.031   | 0.050 |
| NO                         | 166      | 101      | 65       | 0.721   |       |
| DELAYED HEALING            |          |          |          |         |       |
| YES                        | 153      | 86       | 67       | 1.018   | 0.074 |
| NO                         | 167      | 114      | 53       | 0.679   |       |
| PARTIAL PARESIS            |          |          |          |         |       |
| YES                        | 192      | 32       | 160      | 0.774   | 0.246 |
| NO                         | 128      | 168      | 37       | 0.623   |       |
| MUSCLE STIFFNESS           |          |          |          |         |       |
| YES                        | 135      | 60       | 75       | 0.978   | 0.220 |
| NO                         | 185      | 140      | 45       | 0.360   |       |
| ALOPECIA                   |          |          |          |         |       |
| YES                        | 78       | 101      | 23       | 1.022   | 0.267 |
| NO                         | 242      | 99       | 144      | 0.169   |       |
| POLYDIPSIA                 |          |          |          |         |       |
| YES                        | 225      | 8        | 217      | 0.490   | 0.440 |
| NO                         | 95       | 192      | 97       | 0.567   |       |
| SUDDEN WEIGH               |          |          |          |         |       |
| YES                        | 188      | 29       | 159      | 0.765   | 0.264 |
| NO                         | 132      | 171      | 61       | 0.589   |       |
| POLYPHAGIA                 |          |          |          |         |       |
| YES                        | 189      | 48       | 141      | 0.859   | 0.163 |
| NO                         | 131      | 152      | 79       | 0.700   |       |
| WEAKNESS                   |          |          |          |         |       |
| YES                        | 218      | 87       | 131      | 0.888   | 0.046 |
| NO                         | 102      | 113      | 57       | 0.960   |       |
| OBESITY                    |          |          |          |         |       |
| YES                        | 61       | 27       | 34       | 0.757   | 0.379 |
| NO                         | 259      | 173      | 86       | 0.302   |       |

Based on the results table, it can be seen that the highest gain value is in the polydipsia parameter, which is 0.440. Therefore, it is certain that the first root will be occupied by the polydipsia parameter.

The instrument used to display the decision tree in this study is the rapid miner, for previously processed data the results of the decision tree are as follows:
From the results of the decision tree model processing (Figure 2), it can be explained that polydipsia has the most influence on whether a patient has diabetes or not. For more details, it can be explained as follows:

- If the polydipsia is no, polyuria is also no, then the patient is declared negative.
- If the polydipsia is no, polyuria is yes, delayed healing no, then the patient is declared positive.
- If the polydipsia is no, polyuria is yes, delayed healing is yes, itching is no, then the patient is declared positive.
- If the polydipsia is no, polyuria is yes, delayed healing is yes, itching is yes, alopecia is no, partial is no, obesity no, then the patient is declared positive.
- If the polydipsia is yes, polyuria is no, polyphagia is yes, then the patient is declared positive.
- If the polydipsia is yes, polyuria is no, polyphagia is no, itching is no, then the patient is declared positive.
• If the polydipsia is yes, polyuria is no, polyphagia is no, itching is yes, delayed healing is yes, then the patient is declared positive.
• If the polydipsia is yes, polyuria is no, polyphagia is no, itching is yes, delayed healing is no, muscle stiffness is no, then the patient is declared positive.
• If the polydipsia is yes, polyuria is no, polyphagia is no, itching is yes, delayed healing is no, muscle stiffness is yes, then the patient is declared negative.

The results of the model performance using the tools are shown in Table 3.

|                     | true Positive | true Negative | class precision |
|---------------------|---------------|---------------|-----------------|
| pred. Positive      | 78            | 1             | 98.73%          |
| pred. Negative      | 14            | 63            | 81.82%          |
| class recall        | 84.78%        | 98.44%        |

To get the accuracy value, the formula (3) is below:

\[
\text{Accuracy} = \frac{TP + TN}{\text{total instances}}
\]  

or it can be said that the accuracy value for this algorithm model is 90.38% [15].

4. Conclusion
This disease has several symptoms in the patient but evaluating the different factors or symptom variables required to determine which variables are more dominant. This research aims to establish the most influential variable of the many variables causing diabetes mess. We have using a data mining decision tree (C4.5) in this paper to forecast diabetes to help doctors analyse the disease sooner. Data mining has carried out various approaches to predicting disease; one of them is c4.5. The experimental results show that the parameter that has the greatest influence on diabetes is polydipsia, the performance results show a fairly good accuracy value, namely 90.38% so that this algorithm model can be concluded as quite good. Therefore, someone who has symptoms of polydipsia can check diabetes early. We prefer to refine the model in the future by incorporating more data from various sources and considering other variables such as eating patterns, lifestyle, etc.

Acknowledgment
Thanks to Hamzanwadi University for providing grants for this research.

References
[1] Tenenbaum-Gavish K, Sharabi-Nov A, Binyamin D, Møller H J, Danon D, Rothman L, Hadar E, Idelson A, Vogel I, Koren O, Nicolaides K H, Gronbaek H and Meiri H 2020 First trimester biomarkers for prediction of gestational diabetes mellitus Placenta 101 80–9
[2] Tigga N P and Garg S 2020 Prediction of Type 2 Diabetes using Machine Learning Classification Methods Procedia Comput. Sci. 167 706–16
[3] Edla D R and Cheruku R 2017 Diabetes-Finder: A Bat Optimized Classification System for Type-2 Diabetes Procedia Comput. Sci. 115 235–42
[4] Perveen S, Shahbaz M, Guergachi A and Keshavjee K 2016 Performance Analysis of Data Mining Classification Techniques to Predict Diabetes Procedia Comput. Sci. 82 115–21
[5] Mercaldo F, Nardone V and Santone A 2017 Diabetes Mellitus Affected Patients Classification
and Diagnosis through Machine Learning Techniques *Procedia Comput. Sci.* **112** 2519–28
[6] Swapna G, Vinayakumar R and Soman K P 2018 Diabetes detection using deep learning algorithms *ICT Express* **4** 243–6
[7] Hopner F 2020 Multidimensional Decision Tree Splits to Improve Interpretability *Procedia Comput. Sci.* **176** 156–65
[8] Budiman E, Haviluddin, Dengan N, Kridalaksana A H, Wati M and Purnawansyah 2018 Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation *Lect. Notes Electr. Eng.* **488** 380–9
[9] Islam F, Ferdousi R, Rahman S and Bushra H Y 2019 *Computer Vision and Machine Intelligence in Medical Image Analysis*
[10] HESSINA B, MERBOUHA A, EZZIKOURI H and ERRITALI M 2014 A comparative study of decision tree ID3 and C4.5 *Int. J. Adv. Comput. Sci. Appl.* **4** 13–9
[11] Sharma S, Agrawal J and Sharma S 2013 Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies *Int. J. Comput. Appl.* **82** 28–32
[12] Dash M and Liu H 1997 Feature selection for classification *Intell. Data Anal.* **1** 131–56
[13] Salappa a., Doumpos M and Zopounidis C 2007 Feature selection algorithms in classification problems: an experimental evaluation *Optim. Methods Softw.*
[14] Kirshners A, Parshutin S and Gorskis H 2016 Entropy-Based Classifier Enhancement to Handle Imbalanced Class Problem *Procedia Comput. Sci.* **104** 586–91
[15] Nellore S B 2015 Various performance measures in Binary classification-An Overview of ROC study *IJISET-International J. Innov. Sci. Eng. Technol.* **2** 596–605