Classification of Diabetes Disease Using Decision Tree Algorithm (C4.5)

**Baiq andriska candra permana1, Aris Sudianto2,** **Hariman Bahtiar3,** **Ramli Ahmad4,** **Indra Gunawan5**

1,2,5Informatics Engineering Department, Faculty of Engineering, universitas hamzanwadi, Indonesia

3Information Systems Department, Faculty of Engineering, universitas hamzanwadi, Indonesia

4Computer Engineering Department, Faculty of Engineering, universitas hamzanwadi, Indonesia

email : 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. 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 polydhipsia play a role in diabetes with 90.38 % accuracy..

**Keyword :** Prediction, decision tree,c4.5,diabetes

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 method for classification task 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 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.

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 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, deleyed healing, partial paresis, muscle stiffness, alopecia, obesity). 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 Dischotomiser 3 (ID 3) algorithm[8]. 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. The work steps as contained in Figure 1:



**Figure 1.** Classification Technique Using The C4.5 Algorithm To Obtain A Decision Tree

Results and Discussion

3.1 fitur selection

The selection feature is a technique used to select the parameters that are considered the most influential on the class. 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 is a table 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 | No | No | Yes | Yes | No | Yes | Yes | Yes | Positive |
| 2 | No | No | No | Yes | No | No | Yes | No | No | Yes | No | Yes | No | Positive |
| 3 | Yes | No | No | Yes | Yes | No | No | Yes | Yes | No | Yes | Yes | No | Positive |
| 4 | No | No | Yes | Yes | Yes | Yes | No | Yes | Yes | No | No | No | No | Positive |
| 5 | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Positive |
| 6 | Yes | Yes | No | Yes | Yes | No | Yes | Yes | Yes | No | Yes | Yes | Yes | Positive |
| 7 | Yes | Yes | No | Yes | Yes | Yes | No | No | Yes | Yes | No | No | No | Positive |
| 8 | Yes | Yes | Yes | Yes | No | No | Yes | Yes | No | Yes | Yes | No | No | Positive |
| 9 | Yes | Yes | No | Yes | Yes | Yes | No | Yes | No | Yes | Yes | No | Yes | Positive |
| 10 | No | Yes | Yes | Yes | Yes | No | Yes | Yes | No | No | No | Yes | No | Positive |
| 11 | Yes | Yes | No | Yes | No | Yes | No | No | Yes | No | Yes | Yes | No | Positive |
| 12 | … | ….. | …. | …. | ….. | ….. | ….. | …… | ….. | ….. | ….. | ….. | ….. | …… |
| 520 | No | No | No | No | No | No | No | No | No | No | No | No | No | Negative |

|  |  |
| --- | --- |
| Keterangan : 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 gian value of each parameter using the formula for finding the entropy value as listed in (1) and the gain value as contained in (2)[9] :

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

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:

Poliyuria Positive :

Poliyuria Negative :

Poliyuria gain :

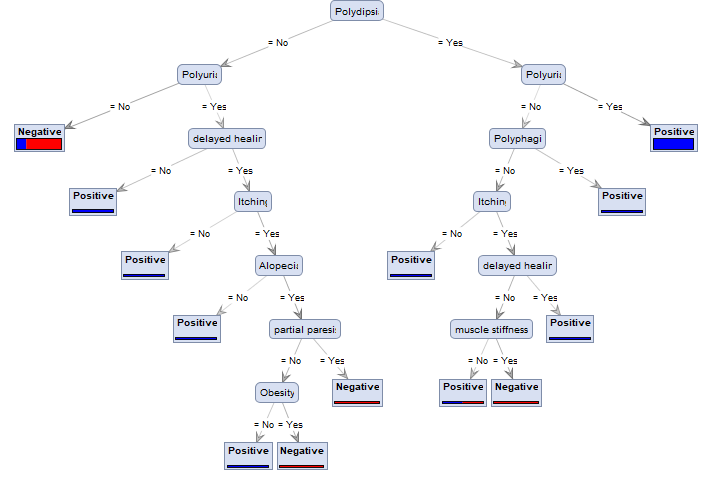
The same technique is applied to the other parameters in order to obtain a table of the results of the entrophy value and gain value used to determine the first root of the decision tree as shown in table 2:

**Table 2**. Entrophy Value And Gain Value

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



**Figure 2**. Results Of The Decision Tree

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

**Table 3.** Model Performance

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 :

0,9038

or it can be said that the accuracy value for this algorithm model is 90.38%.[10]

Conclusion

The experimental results show that the parameter that has the greatest influence on diabetes is polydhipsia, 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 polydhipsia can check diabetes early.

**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] HSSINA 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

[9] Kirshners A, Parshutin S and Gorskis H 2016 Entropy-Based Classifier Enhancement to Handle Imbalanced Class Problem *Procedia Comput. Sci.* 104 586–91

[10] 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