Application of Data Mining In Predicting Herpes Disease Using the C4.5 Algorithm

Lisna Wulian Urfa
Department of Informatics
Universitas Jenderal Achmad Yani
West Java, Cimahi, Indonesia
Lisnawulianw18@if.unjani.ac.id

Tacbir Hendro Pudjiantoro, Fajri Rakhmat Umbara
Universitas Jenderal Achmad Yani
West Java, Cimahi, Indonesia
tacbirpudjiantoro@gmail.com, fajri.rakhmat@lecture.unjani.ac.id

Abstract

Herpes is a type of skin disease characterized by unilateral radicular pain and the appearance of vesicular lesions limited to the skin area. This c4.5 algorithm method is included in the decision tree algorithm to build computational models to predict accurately. According to statistical research ever released by the World Health Organization, one in six people has herpes. Estimates from this study show that about 67% of people worldwide have the herpes virus. Estimates from this study show that about 67% of people worldwide have the herpes virus. Prediction and early detection to reduce the risk of exposure to other people so that medical action can be taken for healing. This prediction is carried out using the Decision tree c4.5 method. This research predicts herpes using eight variables, namely Chlamydia, sugar, alcohol, hepatitis, pregnancy, HIV, Ethylparaben, and Butylparaben. Based on the test results, it was found that the resulting model used the decision tree method, which was 90%. We also recommend using other ways to increase or as a comparison of.

Keywords

Prediction, Herpes, C4.5, Decision Tree, Algorithm.

Introduction

Herpes is a type of skin disease characterized by unilateral radicular pain and the appearance of vesicular lesions limited to the skin area. The characteristic feature of herpes is usually the appearance of tiny red bubbles or groups on the skin's surface accompanied by itching and burning. Herpes can also be spread through skin and intercourse. [1] This herpes disease can be spread to other people through skin-to-skin contact. So to avoid the spread of herpes, this research was carried out to predict people with the possibility of being able to spread or be infected by this herpes virus to reduce the risk of possible exposure to expect people with the case of exposure so that treatment can be carried out in infected people.[2]

On the other hand, in today's digital era, technological developments provide many benefits in various fields, one of which is the health sector. One of the technologies needed in the health and medical fields is a chronic disease risk prediction system in Indonesia with good decision-making and high accuracy, using this system can help individuals to improve the quality of their health and avoid dangerous health situations. before it's too late [6]. This prediction system will produce high and low-risk values for a disease, which is based on medical records and variable assessment data.

In the health and medical fields, Data Mining has been widely used. It has been widely recognized for its potential in the process of discovering valuable information in large volumes of data. To analyze large amounts of data in the database, one of the techniques in Data Mining is used. Namely, the Prediction technique, where the prediction results are class labeled, for pattern formation extracted from historical data, and the data used for training includes descriptive attributes (supervised data)[3].

Prediction is an attempt to systematically seek and obtain answers that are closest to what might happen in the future according to factual information, not necessarily providing clear results or answers. One of the algorithms in this Prediction method is Decision Tree C4.5 which has been widely used in many cases to deal with prediction and classification problems[4].

The previous study entitled "Requirements engineering of a Herpes Simplex Virus patient registry: Alpha phase" that is predicting the Herpes virus using the Random Forest algorithm. Where this model can predict, train, and test models to predict the risk of herpes virus infection. However, this study has limitations, namely only using the Random Forest algorithm and not using other algorithms[5]. To find out using different methods, this research uses the C4.5 classification method.

This C4.5 algorithm method is included in the Decision Tree algorithm; besides previous research no one has used the C4.5 algorithm[6]. The advantages of the C4.5 algorithm are that it can predict by providing high accuracy. This decision tree can be used in various fields, such as medical disease analysis, text classification, smartphone user classification, images and many more. Decision trees are a powerful method used in fields such as machine learning, image processing, and pattern identification. Decision trees consist of several types, namely Iterative Dichotomies 3 (ID3), Successor of ID3 (C4.5), Classification And Regression Tree (CART) and so on [7].

1. Method

In this study, there are several stages carried out in achieving the goal, namely data collection, pre-processing,The first stage is an important stage in this research because data mining has original data that has been obtained through data collection or data collection. After the data is obtained then, pre-process the data. This stage is the stage for cleaning inappropriate or missing data and changing the data to be according to the needs of the method used. Then implement the c4.5 algorithm technique using a tool to create the model, namely Python, then proceed with testing and evaluation. Testing and evaluation need to be done for the final performance results with the suitability of the software as a guarantee of product quality. This stage is also the final stage and the conclusion of the results of this research that has been carried out from each of the stages.

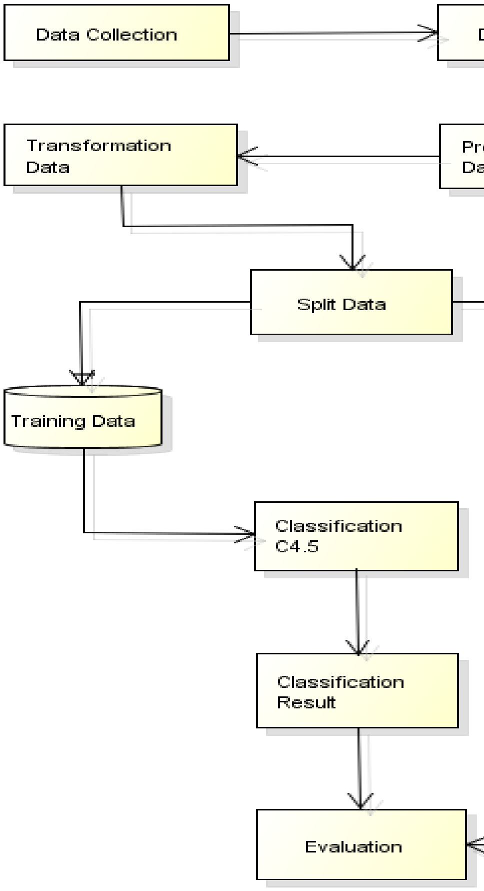


Figure 1. Research Methods

Data Collection

In this study, medical records of herpes patients were obtained from www.kaggle.com .The data used in this study were patient data with the attributes Alcohol consumption, hepatitis A, hepatitis B, hepatitis B Core Anti Body, hepatitis B surface antigen, hepatitis D (Anti-HDV), hepatitis E, and HIV.

Pre-Processing Data

At this stage, it has a function to change patient data whose data is already available and then processed into data that is ready to be processed into objects of research. This research has several pre-process stages, including data cleaning and data selection.

Data Cleaning

Data cleaning or data cleaning is a process carried out to clean patient data that does not have complete or missing medical record data to maintain data quality. the data cleaning process carried out in this study was to improve data and remove incomplete data for large numbers of data such as variables which initially had 15 variables before data cleaning including :

Chlamydia, butyl paraben, ethyl paraben, methyl paraben, propyl paraben, consumption of coffee, tea, sugar with cream or sugar, consumption of alcohol, hepatitis A, Hepatitis B, Hepatitis B Core antibodies, hepatitis B surface antigen, hepatitis D (Anti-HDV), Hepatitis E, HIV, Pregnancy.
Becomes 8 variables after cleaning the data among them:
Alcohol consumption, hepatitis A, hepatitis B, hepatitis B Core Anti Body, hepatitis B surface antigen, hepatitis D (Anti-HDV), hepatitis E, HIV.

Becomes 8 variables after cleaning the data among them:
Alcohol consumption, hepatitis A, hepatitis B, hepatitis B Core Anti Body, hepatitis B surface antigen, hepatitis D (Anti-HDV), hepatitis E, and HIV.

Data Selection

The second stage after data cleaning is data selection, where data that already has complete information is then selected according to the information needed. Selection data of the 9 required attributes were obtained from medical records based on 1 data source.

Decision Tree C4.5 Implementation

The next stage is the Algoritma C4.5 Decision Tree if in the previous stage, where the data has been processed and transformed. The data is ready to be used as a decision tree model for Algorithm C4.5. Then it is classified using Algorithm C4.5 to find rule patterns to produce a decision tree. So that it can be known which class the data belongs to so that it can be seen the level of accuracy of the C4.5 Algorithm classification using this model.This algorithm has the function of studying existing data to produce a pattern or model, following are the steps for building a decision tree with the c4.5 [8] decision tree algorithm.
Select the attribute as the root by calculating the value of entropy, information gain, split information and gain ratio. The attribute that has the highest gain ratio value will be selected as the root node of the tree.
Create a branch for each value.
Divide cases into branches.
Repeat the process for each branch until all cases on the branch have the same class.
There are several stages in making a decision tree with the C4.5 algorithm :
Prepare training data. Training data is usually taken from historical data that has occurred before and has been grouped into certain classes.[9]
Determine the roots of the tree. The roots will be taken from the selected attribute by calculating the gain value of each attribute, and the highest gain value will be the first root. Before calculating the gain value of the attribute, first, calculate the entropy value. To calculate the entropy value, the formula is used [10]

1.

$$Entropy(S) = \sum_{i=1}^{n} -pi\ x\ log_2(pi)$$

$$\text{GainRatio}(S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

$$\text{SplitInfo}(S, A) = \sum_{i=1}^{n} \frac{|Si|}{|S|} \log_2 \frac{|Si|}{|S|}$$

2. Repeat entropy calculation, find the next highest gain ratio and set it into the next node
3. Check if all the attributes have formed the tree. Otherwise, repeat the previous step to all branches to form a simpler tree.[11]

Information:
- S describes the number of cases
- n is the total partition of attribute A
- pi describes the probability value gained from class, and it is divided by the total number of cases (Proportion of Si to S)
- A shows the attribute
- SI is total cases from partition i

2. Result and Discussion

1. Implementation prediction system

   Herpes disease information data obtained from the website. The information displayed through this data is in the form of variable characteristics of herpes disease.

2. Construction of decision tree model

   In this study, the data was divided into training data and testing data as much as 6,870 or 70% and partitioned for training data and 2,943 or 30% for testing data. The training data is used to build a decision tree, and the results can be seen in table 1.

Atribut	Entropy	Gain
Alcohol Consuming	0,918295834 0	0,144484
Hepatitis A	0,918295834 0,985228136	0,005802
Hepatitis B	0,970950594 0,970950594	0
Hepatitis B Core Anti Body	0,985228136 0	0,281291
Hepatitis B Surface Antigen	1 0	0,170951
Hepatitis D (Anti-HDV)	0,99107606 0	0,078982
Hepatitis E	0,918295834 0	0,144484
HIV	0,918295834 0	0,144484

Table 1. Algorithm C4.5 calculation result

3. Decision Tree Model Analysis

Based on the model construction that has been made, we construct a tree model that describes the relationship between attributes. The tree model can be seen in figure 2.

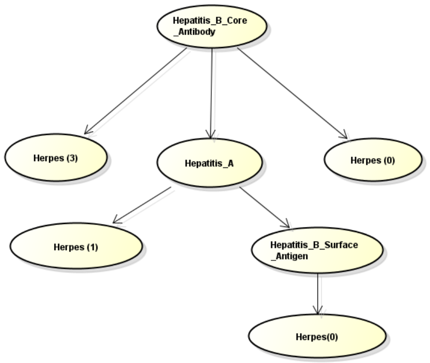
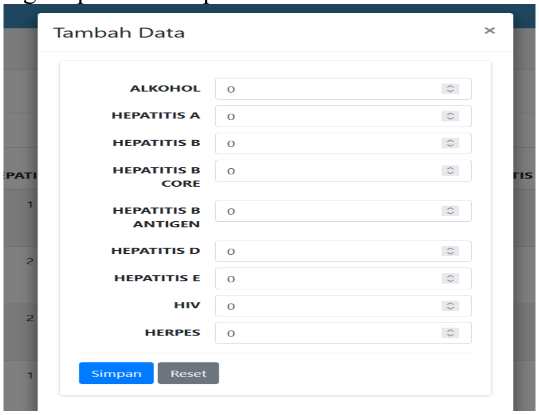


Figure 2. Decision tree models

The results of the c4.5 algorithm decision tree model to identify herpes disease based on medical record data with 8 attributes.

4. Implementation of Prediction Herpes System Models

Figure 3 Shows the results of implementing herpes disease prediction into a web-based software model.



Tambah Data

ALKOHOL 0
HEPATITIS A 0
HEPATITIS B 0
HEPATITIS B CORE 0
HEPATITIS B ANTIGEN 0
HEPATITIS D 0
HEPATITIS E 0
HIV 0
HERPES 0

Figure 3. implementation of prediction systems

In Figure 2 the system shows a data-added menu for predictive processing using the C4.5 algorithm by entering data into the form of several attribute data from herpes, with the results shown in Figure 3.

Testing

To measure the performance of the c4.5 model algorithm created, the confusion matrix is used with several variable combinations of predicted values and actual values. The confusion matrix is shown in table 2 to determine the accuracy.

TP 4417	FP 1571
FN 1274	TN 2552

Table 2. Confusion matrix results

The test data result are then used to calculate Accuracy, Precision, and Recall. The calculation results are shown in Table 3.

Recall	77,7%
precision	73,7%
accurancy	77,6%

Table 3. Evalution results

Based on the results of test that have been carried out using the confusional matrix, it displays the results of Recall by 77,7%, Precision by 73,7%, and accuracy by 77,6%, so the result of algorithm C4.5

modeling are quite good and can be used to predict the potential of an indvidual's risk of having the disease.

3. Conclusion

This study can conclude that predicting the risk of herpes disease using the c4.5 algorithm method can be used quite well in predicting this study. The results of testing the system that has been carried out get an accuracy of 71%. This prediction system can help to determine the type of herpes disease. Future research is expected to complement and develop this research in terms of the methods to be used, and the addition of other variable data in the hope that the results created will be maximized and also accurate.

References

O. Taupiqurrohman, A. Noviyanti, M. Yusuf, and T. Subroto, "Analisis In Silico Capsid Scaffold Protein Virus Herpes Simpleks-1 Untuk Pengembangan Vaksin Herpes," *Chim. Nat. Acta*, vol. 5, no. 1, p. 21, 2017, doi: 10.24198/cna.v5.n1.12817.

J. I. Cohen, D. S. Davenport, J. A. Stewart, S. Deitchman, J. K. Hilliard, and L. E. Chapman, "Recommendations for Prevention of and Therapy for Exposure to B Virus," vol. 20892, pp. 1191–1203, 2002.

N. L. Fitriyani and M. Syafrudin, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," vol. 7, 2019.

B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4 . 5," no. 2, pp. 13–19.

S. Surodina, C. Lam, S. Grbich, M. Milne-Ives, M. van Velthoven, and E. Meinert, "Requirements Engineering of a Herpes Simplex Virus Patient Registry: Alpha Phase," no. Grant 18654, pp. 152–160, 2020, doi: 10.21203/rs.3.rs-38387/v1.

H. Jantan, "Human Talent Prediction in HRM using C4 . 5 Classification Algorithm," no. November, 2010.

D. T. Induction, "Performance Analysis of Attribute Selection Methods Decision Tree Induction Decision Tree Induction," 2018.

A. A. Aldino and H. Sulistiani, "DECISION TREE C4 . 5 ALGORITHM FOR TUITION AID GRANT PROGRAM CLASSIFICATION (CASE STUDY : DEPARTMENT OF INFORMATION SYSTEM , UNIVERSITAS TEKNOKRAT INDONESIA)," vol. 7, no. 1, pp. 40–50, 2020.

A. Rohman *et al.*, "IMPLEMENTASI DATA MINING DENGAN ALGORITMA DECISION TREE C4 . 5 UNTUK PREDIKSI KELULUSAN MAHASISWA DI UNIVERSITAS," pp. 134–139, 2019.

Z. Azmi, M. Dahria, P. Studi, S. Komputer, P. Studi, and S. Informasi, "DECISION TREE BERBASIS ALGORITMA UNTUK," pp. 157–164, 1978.

E. P. Cynthia, "MENGKLASIFIKASI DATA PENJUALAN BISNIS GERAI," no. July, 2018, doi: 10.30645/jurasik.v3i0.60.

Biographies

Lisna Wulian Urfa is a final year undergraduate student in the department of informatics, Universitas Jenderal Achmad Yani, Indonesia. Her primary interests are systems analysis, web service technology, data and software engineering.

Tacbir Hendro Pudjiantoro is an Associate Professor. Doctoral Candidate from the Indonesian University of Education. Researcher in the field of Knowledge Management and handles several Information Systems projects

Fajri Rakhmat Umbara is a lecturer in the Department of Informatics, Faculty of Science and Information, Universitas Jenderal Achmad Yani, Indonesia. His research includes data mining and software engineering.