JCHR

Journal of Community Health Research 2022; 11(3): 158-164.

Comparison of Different Decision Tree Algorithms for Classification of Retinopathy Patients in Yazd City, Central Part of Iran

Mohsen Askarishahi [1], Amin Karami[1], Nasim Namiranian [2]

1. Dpartments of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, yazd, Iran
2. PhD Shahid Sadoughi University of Medical Sciences and Health Services

ARTICLE INFO

Original Article

Received: 25 June 2022
Accepted: 16 August 2022

OPEN ACCESS

Corresponding Author:
Amin Karami
amn.krm.rbt@gmail.com

ABSTRACT

Introduction: Diabetes is one of the most common diseases caused by metabolic disorders. It is the result of impaired secretion or function of insulin. The prevalence of diabetes is increasing rapidly. The aim of this study is to investigate the performance of different decision tree algorithms in the diagnosis of diabetic retinopathy. It was done using a database regarding diabetic patients. They were referred to Yazd Diabetes Research Center.

Method: This study was analytical cross-sectional. 2613 patients visited Yazd City's research and treatment center. Their demographic information was received in the first stage. Then, they were tested by the nursing team, and the patient's information form was completed by the respective nurse. After that, the descriptive indicators of mean, mode, median, variance, frequency, and percentage of missing data were observed. Four diagnostic models (Chadi), classification tree and regression (C and R), (Quest) and C 5.0 were compared. Authors evaluated the performance of these four models using three statistical criteria: accuracy, sensitivity, and specificity. Gains chart was used for more accurate comparison of models. SPSS MODELER V 18.0 software was used for data processing and modeling. The significance level was considered 5%.

Result: In this study, among the demographic and clinical variables, BMI, duration of disease, type of drug used, age, hypertension, gender, cholesterol, and hemoglobin A1c were entered in the final model. The dependent variable of retinopathy was investigated. It was based on the obtained criteria of accuracy (71.75), sensitivity (75.60), specificity (57.14) in the CART model; accuracy (65.84), sensitivity (65.86), specificity (65.76) of the Quest model; accuracy (69.33), sensitivity (67.35), specificity (76.81) of Chaid model; and accuracy (73.27), sensitivity (79.65), specificity (49.05) of Chaid model.

Conclusion: Based on the criteria of accuracy, sensitivity, specificity, and comparison of Gain Chart for four algorithms, Chaid algorithm showed better performance. Therefore, for further research, the authors suggest this algorithm.

Keywords: Retinopathy, Diabetes, Decision Tree, Yazd, Data Mining

How to cite this paper:

Askarishahi M, Karami A, Namiranian N. Comparison of Different Decision Tree Algorithms for Classification of Retinopathy Patients in Yazd City, Central Part of Iran. J Community Health Research 2022; 11(3): 158-164.

[DOI: 10.18502/jchr.v11i3.10867]

Introduction

Diabetes is one of the most common diseases caused by metabolic disorders. It is through defects in the secretion or function of insulin. The prevalence of diabetes is increasing rapidly, and the total number of cases worldwide is expected to increase. From 171 million people in 2000 with a prevalence of 2.8 percent in all age groups, it increases to 366 million people with a prevalence of 4.4 percent by 2030. (1, 2)

On the other hand, the cost of treating diabetes and its complications is a major concern, especially in developing countries. The prevalence of type 2 diabetes there is high (3).

Statistics indicate that 14 to 23 percent of Iranians over 30 have diabetes. Seven million Iranians are suffering from diabetes, and in general, one percent is added to the number of diabetics every year. In the age group over 30, the estimate is 16.3 percent (7)

Diabetic retinopathy (DR) is one of the most common causes of diabetes blindness. It is also a part of cardiovascular side effects and the leading cause of blindness among middle-aged people worldwide.Millions of diabetic patients have DR.(8, 9).

The DiabCare Asia 2008 study reports that 42% of people with type 2 diabetes also experience complications of retinopathy. The risk of developing DR is proportional to the amount of time a person develops diabetes. (10-13) The reported prevalence of DR in the diabetic population in Yazd province is 29.6% (14, 15).

Diagnosing diabetic retinopathy in the early stages is necessary to prevent it completely. Many physical tests such as optical coherence tomography can be used to diagnose diabetic retinopathy. But, they take a lot of time and have a negative impact on patients. (16) DR, not only disrupts the patient's vision, but also burdens It brings a person to his family and society. (17-19)

Vision problems have brought an estimated cost of $3 trillion in 2010. This, directly and indirectly, in individuals, health care systems, and the global economy will reach $3.6 trillion in 2020 (15). One solution to this problem is to develop a simple and inexpensive tool to identify and screen individuals who are at risk of developing retinopathy. (1)

Recently, data mining, including decision trees, has become popular in medical research. For example, the medical use of decision trees is in the diagnosis of a medical condition from the set of disease symptoms. In it, the classes defined by the decision tree can be composed of different clinical sub-branches regarding diseases and different treatments. (20, 21, 22)

Materials and methods:

Data set description

The data set of Yazd Diabetes Center used in the first face-to-face visit was from 2613 diabetic people visiting this center. The file was then tested and measured.

In this study, variables of sex, age, duration of disease, type of the drug used, blood pressure, BMI, hemoglobin A1C, cholesterol, triglyceride, blood creatinine, blood sugar level as independent variables and dependent variable of retinopathy were examined.

Decision tree model algorithms

Decision trees

The decision tree (DT) in data mining is a model used to display classifiers and regressions. As the name implies, this tree is made up of a number of nodes and branches. In the decision tree that performs the classification operation. (23,24)

DT algorithms commonly used in academia and industry can be classified into four types: C5.0, CART, Chaid, and Quest.

[DOI: 10.18502/jchr.v11i3.10867]

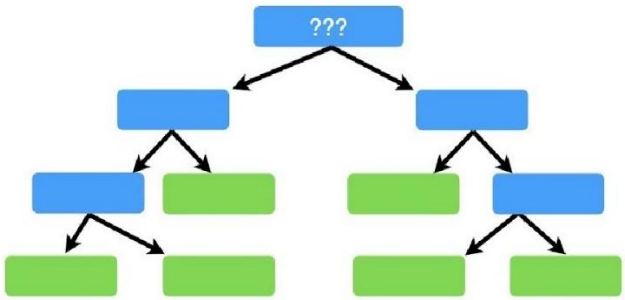

Chart 1. An example of a decision tree diagram

1. CHAID

The Chaid algorithm was originally designed for nominal variables. This algorithm uses different statistical tests depending on the type of class label. This algorithm stops when it reaches a defined maximum depth, or the number of instances in the current node is less than a defined value. The Chaid algorithm does not run any pruning methods and can also control incomplete values.

The Chaid tree model relies on Chi-square test to determine the best division at each stage. (26)

2. Classification and regression tree (C and R)

The C and R tree algorithm creates a regression model or a classification model, depending on whether the target variable is continuous or classified.

C and R tree models using the Gini index to minimize variance increasingly split data to find homogeneous subsets. The purpose of tree planting is to create subgroups with similar output values. Gini impurity measurement is as follows:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Equation: 1

Pi is the probability of classifying an object in a particular class.

3. Tree (Quest)

The classification tree obtained from this algorithm, such as the CART model, has binary divisions. The criterion for deciding the variables is using P-value of the F-statistic of the ANOVA test for quantitative variables and the P-value of the chi-square statistic of the agreement tables for the variables qualitatively.

4. Algorithm C 5.0

C5.0 is an improved version of the C4.5 and ID3 algorithms. It is a commercial product designed to analyze large data sets. C5.0 uses the concept of entropy to measure purity.

Measurement indicators

The performance of each classification model is evaluated using three statistical criteria: accuracy, sensitivity and specificity. These measures are defined using true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

True positive (TP) = the number of cases the patient is correctly diagnosed.

False positive (FP) = the number of times a patient is misdiagnosed.

True negative (TN) = the number of cases that have been correctly diagnosed as healthy.

False negative (FN) = the number of cases that have been incorrectly identified as healthy.

Sensitivity:

Sensitivity is a test regarding the ability to correctly determine a patient's case. To estimate it, researchers have to calculate the positive ratio in the patient's case.

Basically, a test is used for diagnosis. Therefore, the possibility that the test can correctly diagnose the disease should be well understood. It does not provide the sensitivity and specificity of such information.

Mathematically, this can be said as follows:

$$Sensitivity = \frac{TP}{TP+FN}$$

Equation: (2)

Specificity

[DOI: 10.18502/jchr.v11i3.10867]

A test is characterized by the ability to correctly determine healthy cases. To estimate this, the researchers have to calculate the real negative ratio in healthy cases. Mathematically, this can be as the following: (27)

$$Specificity = \frac{TN}{TN+FP}$$

Equation: (3)

Statistical analysis steps

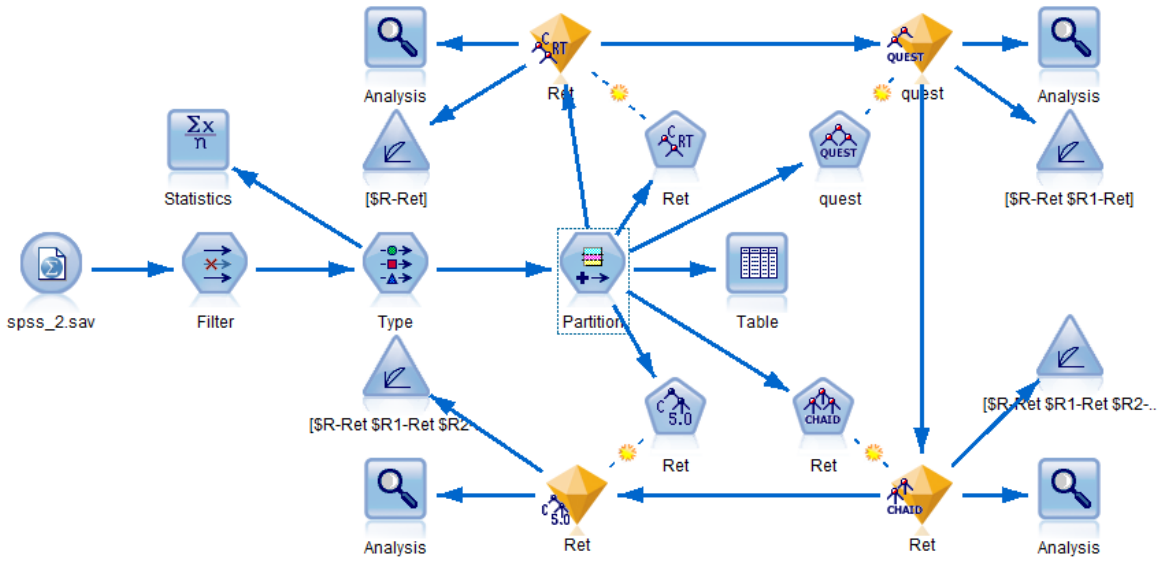

Chart 2. Model generation steps in spss modeler software

SPSS MODELER software, one of the softwares in the field of data mining and text analysis from IBM, was used to build prediction models and perform other analytical tasks. For each model, the authors performed two activities to test accuracy and compare the models with each other:

1- Entering and correcting the data structure

2- Determining the dependent and independent variables and the type of variables

3-Partitioning the data into test and trade categories

4- Making statistical models

5- Calculation of accuracy, sensitivity and specificity rates

6- Creating Gain Charts to compare models

Results

Draw a Gain Chart

The structure of the Gain Chart diagram is like drawing a diagram. The closer the diagrams of each algorithm are to the BEST-RET axis, the better evidence of the algorithm's performance is the green diagram. This characterizes the Chaid algorithm.

It is higher in the test category (Test) and in the training category (Train) than the diagrams of other algorithms.

[DOI: 10.18502/jchr.v11i3.10867]

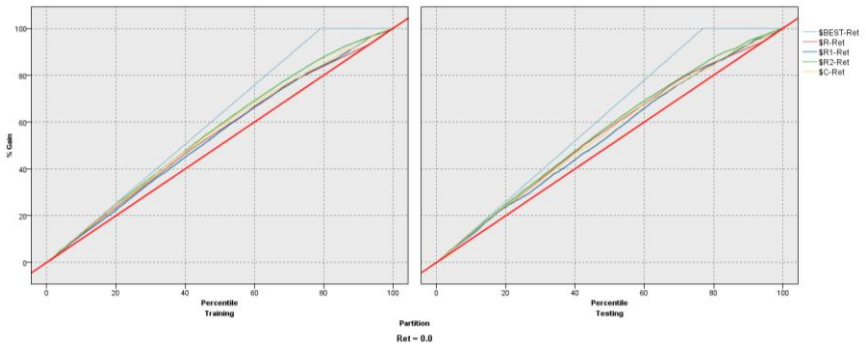

Chart 3. Gain Charts to compare models

Calculate accuracy, sensitivity and specificity indicators

Predictions of all models are compared with the main classes to identify true positive, true negative, false positive, and false negative values. Each cell contains the raw number of items classified to combine with the desired and actual classifier outputs. The values of statistical parameters (sensitivity, specificity and accuracy of the total classification) are calculated from four models and are presented in Table 2. Accuracy, sensitivity and specificity approximate the probability of positive and negative labels being correct. They evaluate the usefulness of the algorithm in a single model.

Moreover, the percentage and number of missing data were added to all the data from each variable to analyze the most suitable algorithm with missing data.

According to the results of Chart(4), the variables related to retinopathy according to the Chaid algorithm in the order of the most influencing variables was as follows:

BMI had the greatest impact in terms of this algorithm. Duration of illness, medication, patient's age, hypertension, gender, cholesterol, hemoglobin ,A1C were next in terms of importance.

Table 1. Table of values of sensitivity, specificity and accuracy for each model

Model	Test or train	Specifity	Sensitivity	Accuracy
Cart	test	55.95	78.69	**73.44**
	train	57.14	75.60	71.75
Quest	test	59.58	66.71	65.07
	train	65.76	65.86	65.84
Chaid	test	72.02	67.96	68.9
	train	76.81	67.35	69.33
C5.0	test	46.63	79.16	71.65
model	train	49.05	79.65	73.27

Table 2. Table of percentage of missing data for each variable

Variable	Number of data	Number of missing data	Percentage of missing data
Gender	2613	0	%0
Age	2374	239	9.15 %
Duration	934	1679	64.26 %
Drug	2613	0	%0
Hypertension	2613	0	%0
Retinopathy	2613	0	%0
BMI	1701	912	34.91 %
Blood sugar level	2360	253	9.69 %
H A1C	2564	49	1.88 %
Creatinine	2546	67	2.57 %
Cholesterol	1289	1324	50.67 %
Cholesterol	1290	1323	50.63 %

[DOI: 10.18502/jchr.v11i3.10867]

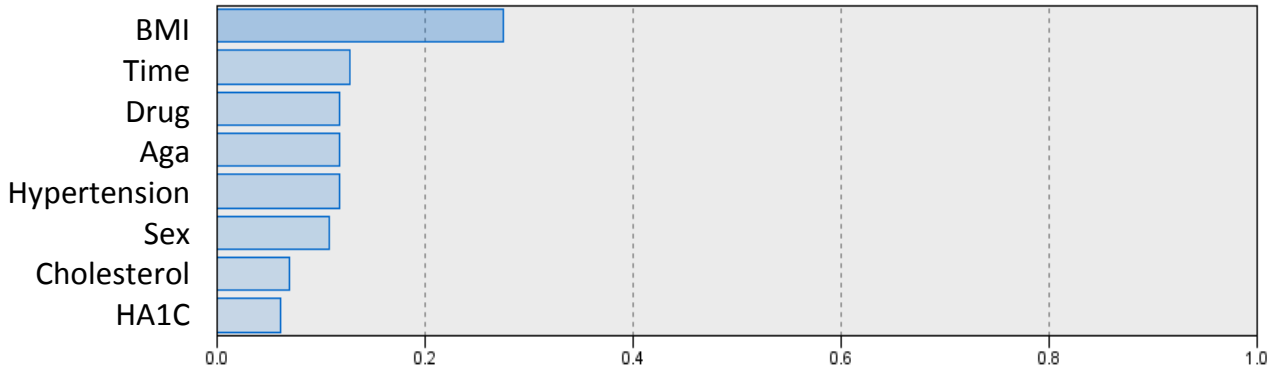

Chart 4. Variables related to Chaid's algorithm in the order of the most influential variables

Discussion

The significance level of 0.05 was for splitting and the significance level of 0.05 for merging was for output of the software. The results of Table 2 show that the classification accuracy in the Cart model test sample with 73.44% is the best accuracy. Model C5.0 and Chaid with rates of 71.65 and 68.9 percent are at a short distance from the approximate value of the Cart model. After them is the Quest model with an approximate value of 65.07 percent. Sensitivity analysis is often used to determine the degree to which each predictive feature contributes to the identification of output class values (27).

The practical meaning of the specificity rate generally represents the percentage of times that the test in this study correctly diagnosed cases as positive. For this reason, it is significant for authors that the higher the percentage of specificity error, the higher is the percentage of those who are ill and wrongly diagnosed as healthy. This causes lack of trust in the existing test. Furthermore, sensitivity represents the percentage of times that the test correctly places healthy people in the category of healthy people. Its error causes healthy people to be wrongly classified in the category of ill people. The importance of sensitivity is, thus, lower than the characteristic case. The high feature rate is of double importance for the researchers.

According to the sensitivity and specificity rates in the test table above, the authors conclude that the test sample in the Chaid model has the highest specificity rate with a value of 72.2%. This is followed by the Quest model with 59.58%, the Cart model with 57.14%, and the C5.0 model with 49.05%. This suggests that the Chaid model has the best results in the table. By vareful observation, the researchers found that Gains Chart in the example of testing the validity of this claim shows the Chaid model is better than other models.

Compared to other models, the Chaid chart has the ability to react to missing data and create splitting on the missing data due to the large amount of missing data. In some data variables seen in Table (2), the authors can prove better performance of this model compared to other models.

In the evaluation paper of Cart, Chaid and Quest algorithms, the purpose of this study was to discover the capability of three types of decision tree algorithms: Cart, Chaid, and Quest. It showed the results regarding predicting the construction of the algorithm. Among the three algorithms, Chaid created the highest classification rules and displayed the greatest prediction accuracy (29).

In the study of breast cancer diagnosis using decision tree models and SVM, the authors evaluated classification performance of four different decision tree models of Chaid, C and R, Quest, and C5. Then, they compared the results with SVM in breast cancer diagnosis. Significance analysis has shown that the feature of "cell size uniformity" in Chaid and Quest was the most important feature in differentiating cancer from healthy samples (30).

[DOI: 10.18502/jchr.v11i3.10867]

In the study comparing different decision tree algorithms to evaluate the severity and type of collisions in the urban network, this research compared different decision tree algorithms to evaluate the severity and type of collisions in the urban network (case study: Mashhad city). Comparing the models, 4 decision tree algorithms including Quest, Chaid, Cart, and C5 algorithms have been used. It has been to other models made.

Conclusion

In this study, the classification performance of four different decision tree algorithms of Chaid, Cart, Quest, and C5.0 in the diagnosis of retinopathy is evaluated. This is done using data from diabetic patients referred to the Diabetes Center of Yazd

The results showed that the best model in terms of performance accuracy is decision tree model using Chaid algorithm.

Acknowledgment

Author like to thank healthcare Data modeling center of Shahid sadougi university of medial science. This study was part of MSc thesis with ethical code IR.SSU.SPH.REC.1399.063.

Author Contribution

All authors contributed to data collection and modeling

References

1. Elsalamony HA, Elsayad AM. Bank direct marketing based on neural network and C5. 0Models. Int J Eng Adv Technol IJEAT. 2013;.(6)2
2. Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications: Academic Press; .2009
3. Floares A, Birlutiu A, editors. Decision tree models for developing molecular classifiers for cancer diagnosis. The 2012International Joint Conference on Neural Networks (IJCNN); 2012: IEEE.
4. Kass GV. An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1980;29(2):.27-119
5. Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural networks. 1990;3(5):.60-551
6. Breiman L. Fried man JH, Olshen RA, Stone CJ. Classification and regression trees Belmont, California: Wadsworth International Group. 1984.
7. Lin C-L, Fan C-L. Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. Journal of Asian Architecture and Building Engineering. 2019;18(6):539-53.
8. Elsayad AM, Elsalamony H. Diagnosis of breast cancer using decision tree models and SVM. International Journal of Computer Applications. 2013;83(5).

[DOI: 10.18502/jchr.v11i3.10867]