TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

Diabetic and Kidney Disease Prediction in Human based on Their Age Group using C4.5 Decision Tree Algorithm in Python

Sibi R, Research Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore – 641 008, sibir17mft008@skasc.ac.in

Valarmathi V, Assistant Professor, Department of Information Technology, Sri Krishna Arts and Science College, Coimbatore – 641 008, valarmathiv@skasc.ac.in

Kavitha S K, Assistant Professor, Department of Information Technology, Sri Krishna Arts and Science College, Coimbatore – 641 008, kavithask@skasc.ac.in

Article Info
Volume 82
Page Number: 8335 - 8342
Publication Issue:
January-February 2020

Article History
Article Received: *5 April 2019*
Revised: *18 Jun 2019*
Accepted: *24 October 2019*
Publication: *07 February 2020*

Abstract:

An immense measure of information is produced inside the fields of tending and prescription; specialists need to make an on the spot contact with patients to work out the wounds and sicknesses. This analysis highlights the appliance of classifying and predicting a selected disease by implementing the operations on medical information generated within the field of medical and healthcare. In this study an efficient c4.5 algorithmic rule is employed for prediction of a particular disease by training it on a group of information before implementation. Wrong clinical decisions taken by medical practitioners will cause any hurt or lead to serious loss of life of a patient that is tough to afford by any hospital. To accumulate an explicit and value effective treatment, technology based mostly data mining systems can be created to make worthy decisions. The main aim of this analysis is to make a basic decision network which can determine and extract previously unseen patterns, relations and concepts connected with multiple disease from a historical information records of specified multiple diseases. The decisions taken by medical practitioners with the help of technology can result in effective and low value treatments. There is an insufficiency of technology and analysis system and methods to get connections, concepts and patterns in the medical information. Data mining is an engineering study of extracting previously undiscovered patterns from a specific set of information. In this study, data mining methods namely, C4.5 algorithms are used for detecting the disease in human based on their age group.

Keywords: *Decision Tree, C4.5, Prediction, Medical Dataset, Python.*

I. INTRODUCTION

Data mining refers to methods of extracting or mining information from ample amounts of information. It is the method of searching available patterns by scanning the massive quantity of information. Storing sizable amount of information is helpful to extract valuable knowledge. To seek out constructive patterns within the information, there are totally different varieties of algorithms which may categorize the information either mechanically or semi mechanically. These patterns are used to get

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

the sets of rules. The patterns discovered should be significant such that they'll cause several advantages like decisions making, market research, financial growth, business intelligence, healthcare etc., to urge such meaningful patterns, considerably great deal of information is needed. Besides predicting future observation, data mining is additionally helpful for summarizing the underlying relationship in information. Medicinal services information handling has decent imminent for investigating the concealed examples among the information sets of the restorative space. These patterns will be used for clinical diagnosing. This analysis aims to research the various predictive/ descriptive data mining techniques introduced in recent years for diabetic, kidney and pressure disease diagnosing. Healthcare diagnosing is taken into account as a major yet complicated task that must be carried out exactly and efficiently. The equipment of a similar would be extremely helpful. Prediction of human disease based on age group using c4.5 algorithmic rule in data mining.

II. Literature Survey

ID3 algorithmic guideline is utilized in light of the fact that the instructing calculation to bring up rank of cardiovascular breakdown with the decision tree to utilize and investigate very surprising information handling systems for the forecast of cardiovascular ailment with KNN [1]. Examined the expectation of urinary organ ailments by exploitation Support Vector Machine (SVM) and Artificial Neural Network [9]. Examined k-overlap cross approval, characterization strategy, classification astute K-Nearest Neighbour[CKNN], Support Vector Machine [SVM], LDA Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, applied arithmetic standardization and Back engendering systems for diabetic diagnosis. The C4.5, CART, k-NN and SVM depicted the exactness pace of eighty six, 85%, seventy eight and seventy four for anticipating diabetic and cardiovascular malady [15]. Forecast is wide used in all parts like Weather, Sports, Marketing, Education, Business, Politics and Agriculture in this way on.

Pradhan et al [8] presents polygenic issue location framework that depicts it's inadequate of the body made inferable the lack of hypoglycaemic operator and has as of late increased quality, all around. The twenty beginning century with its inert way among suburbia and a fast, social, urban way all jeopardize Associate in Nursing person's life and advance him towards polygenic issue. However' specialists analyze polygenic issue utilizing an aldohexose investigate, we will in general can't obviously characterize the individual as diabetic or not bolstered these side effects. by and large a pre diabetic half will caution the specialists and to boot the patient with respect to the depreciatory wellbeing and will mindful the patient concerning the concerned measures.

III. Proposed System

The proposed framework applying information mining strategies in unmistakable infection in human. To pass judgment if applying information mining methods to forecast will give as dependable execution as accomplished. The proposed

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

methodology is intended to foresee the human infection like glucose, pee salt and weight dependent on the age gathering. This forecast is performed utilizing C4.5 Decision tree.

3.1 C4.5

C4.5 is partner in nursing recipe used to create a choice tree created by Ross Quinlan. C4.5 is Associate in nursing augmentation of Quinlan's prior ID3 recipe. The choice trees created by C4.5 is utilized for order and thus, C4.5 is typically decided as an applied science classifier. Creators of the AI code outline the C4.5 recipe as a milestone choice tree program that is no doubt the AI workhorse most by and large utilized in pursue to the present reason.

C4.5 constructs choice trees from a gathering of instructing information inside a similar way as ID3, exploitation the origination of data entropy. The instructing information could be a set S=s_1, s_2, … .s_p of effectively ordered examples. Each example s_i comprises of a p-dimensional vector (x_(1,i), x_(2,i), … , x_(p,i)), where the x_i speak to characteristic esteems or decisions of the example, also because of the class inside that s_i falls.

At every hub of the tree, C4.5 picks the property of the information that practically all successfully parts its arrangement of tests into subsets improved in one class or the other option. The gravelly rule is that the standardized data gain (contrast in entropy). The quality with the ideal standardized data gain is picked to settle on the decision. The C4.5 equation then recurses on the parceled sub records.

This equation choices two or three base cases. Every one of the models inside the rundown have a place with a comparable class. When this occurs, it only brings a leaf hub for the choice tree spoken correspondence to choose that class.

Formulas used to compute C4.5 is given below

As a rule, Let P is a likelihood circulation P = (p1, p2, pn) and an example S then the Information passed by this appropriation, additionally called the entropy of P is giving by

$$E = \frac{\sum P_i + N_i}{P+N} I(P_i + N_i) \quad (1)$$

The capacities data gain enables us to quantify the level of blending of classes for all examples and along these lines any situation of the tree in development. It stays to characterize a capacity to pick the test that must name the present hub.

$$I(P,N) = -\frac{P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right) \quad (2)$$

where I = INFORMATION GAIN

E = ENTROPHY

G = GAIN

P = YES

N = NO

At long last the addition is processed by the accompanying formula

$$G = I - E \quad (3)$$

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

IV. IMPLEMENTATION

In this segment, the exhibition of the proposed work is estimated. The examination is done utilizing python language. The python Idle is utilized for advancement and the PostgreSQL is utilized for database get to.

Steps to Implement C4.5 Algorithm for Disease Prediction

1. Take an informational collection as contribution from backend Postgre SQL, database availability is made through Python
2. On the off chance that that set has more highlights, at that point apply the element determination procedure as pre-handling system.
3. Rehash the emphasis from stage 4 to stage 6.
4. Assess the entropy worth and data gain proportion of every one of the four entropies (age_group, blood_sugar, urine_salt and pressure).

For P = 1, N=2

$$I(1,2) = -\frac{1}{1+2}\log_2\left(\frac{1}{1+2}\right) - \frac{2}{1+2}\log_2\left(\frac{2}{1+2}\right)$$

$$E = \frac{\sum 1_i + 2_i}{1+2} I(1_i + 2_i) \quad , \ (i = 1, 2, n)$$

iteration-1 $G = I - E$

5. Build the models utilizing c4.5 calculation dependent on different entropies.
6. Discover the precision and execution time of each model, store the incentive in an exhibit.
7. Locate a model that has greatest Accuracy to discover root component.
8. On the off chance that two models have most extreme exactness, at that point
9. Notice a base execution time of the model that has most exactness.
10. Characterize the model which has least execution time.
11. Else characterization done by the model that has most precision.
12. End.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Data sets of few diseases were downloaded from the UCI repository whereas few of them were been created by consulting doctors and researching concerning the disease. The data sets of disease designated were blood sugar and urine salt as a serious field with 490 rows. Attributes like Patient_id, Age, Sex, Address, phone number were common in all the target data set. Different attributes that are considered for prediction are case history and fat.

The classification algorithmic program that is running at the rear end on the data sets compare the testing data, additionally referred to as previously unseen information with the training data, additionally referred to as the information within the data set. The input information whose properties match with the training data is returned as class label of that instance.

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

TABLE I

Gain of the root element

MAJOR ATTRIBUTES	GAIN VALUES
Age_Grp	0.25
Blood_Sugar	0.047
Urine_salt	0.151
Pressure	0.033

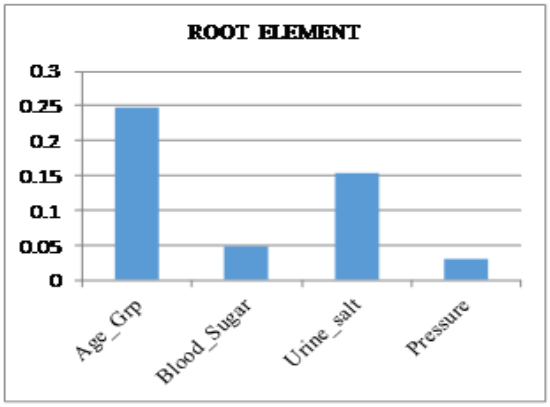


Fig 1. Gain of the root element

The gain of the root element is described in Table I. Gain is computed for Age_Group, Blood_Sugar, Urine_Salt and Pressure attributes. The determined qualities are 0.250, 0.047, 0.151 and 0.033 graphical portrayal of root component in Fig 1.

TABLE II

Gain value for young people

Young People	Values
Gain(B_S)	0.02
Gain(U_S)	0.97
Gain(Pres)	0.52

From the output it is clear that the young people mostly affected by urine salt compared to other disease such as blood sugar and pressure. The graphical represented of these result is shown in Fig 2.

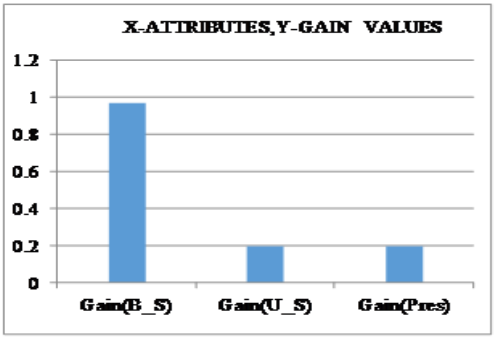


Fig 2. Gain value for young people

TABLE II

Gain value for Middle age people

Middle People	Values
Gain(B_S)	0.97
Gain(U_S)	0.2
Gain(Pres)	0.2

From the output it is clear that the Middle age people mostly affected by Blood sugar compared to other disease such as Urine salt and pressure. The graphical portrayal of these outcome is appeared in Fig 3.

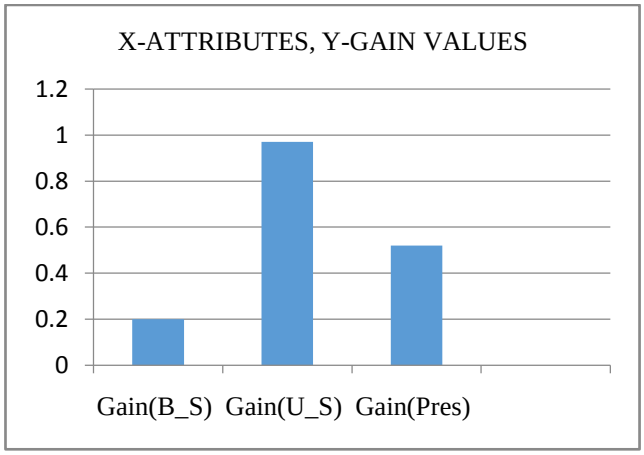
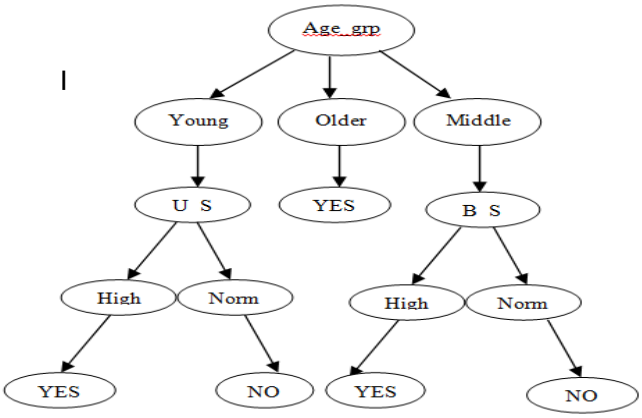


Fig 3. Gain values of middle age people

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

In this study the C4.5 decision tree used to find the disease based on the age group. It's concluded that the young and middle age group people largely affected by urine salt that ends up in kidney disease and blood glucose ends up in diabetics severally.

Fig 4. Decision Tree



VI. Conclusion

The current advancement, development and improvement within the data mining algorithmic rule have assured an ease in getting insight and precise prediction result. The analysis enlightens all the challenges, problems and obstacles sweet-faced by medical practitioners whereas decisive the disease of a patient while not considering the medical data of the patient.

In this work, a model for locating information from clinical data sets that aids a doctor in clinical decision making was proposed. The proposed model had been tailored and evaluated with medical data sets. The medical data set with 490 records are taken for the proposed work. The conclusion comes from the analysis results of C4.5.

The potency of the algorithms is investigated in several medical data sets. From the experimental analysis, it's determined that the proposed ways of research work address the identification of disease chance supported specialists knowledge domain. This produces higher accuracy compared to the present algorithms. Furthermore, this research work is useful for the future researchers to create a research on the medical domain.

VII. Future Work

The present work specialize in two diseases one is blood sugar and another one is urine salt. In future, additional disease has to be detected and predict using the latest approach. Improving the efficiency of the model by implementing the agents with a dynamic choice of information mining techniques counting on the case. It'll be effective because the agent has the power to determine that algorithmic rule to use supported its belief. Also, this could be extended with the prediction of treatments needed and prescribed medicines for the various diseases.

VIII. References

1. Aishwarya, R., and P. Gayathri. "A Method for Classification Using Machine Learning Technique for Diabetes." (2013).
2. Milovic, Boris. "Prediction and decision making in health care using data mining." *Kuwait chapter of arabian journal of business and management review* 33.848 (2012): 1-11.
3. Rafe, Vahid, and Roghayeh Hashemi Farhoud. "A Survey on Data Mining

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

Approaches in Medicine." *International Research Journal of Applied and Basic Science* 4.1 (2013): 196-202.

4. Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." *Global journal of computer science and technology* 10.10 (2010): 38-43.
5. Koutsojannis, Constantinos, and Ioannis Hatzilygeroudis. "Using a neurofuzzy approach in a medical application." *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, Berlin, Heidelberg, 2007.
6. Chen, Dechang, et al. "Developing prognostic systems of cancer patients by ensemble clustering." *BioMed Research International* 2009 (2009).
7. Karla, RN, 'Cardiovascular diseases in Women and Children', *The Hindu open page Chennai* viewed 29 September, 2013 pp. 16.
8. Pradhan, M. A., et al. "A genetic programming approach for detection of diabetes." *Int J Comput Eng Res (ijceronline. com)* 2.6 (2012): 91.
9. Patel, Shamsher Bahadur, Pramod Kumar Yadav, and D. P. Shukla. "Predict the diagnosis of heart disease patients using classification mining techniques." *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)* 4.2 (2013): 61-64.
10. Sa, S. "Intelligent heart disease prediction system using data mining techniques." *International Journal of healthcare & biomedical Research* 1 (2013): 94-101.
11. Jabbar, M. Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. "Heart disease prediction system using associative classification and genetic algorithm." *arXiv preprint arXiv:1303.5919* (2013).
12. Chowdhury, Dilip Roy, Mridula Chatterjee, and R. K. Samanta. "An artificial neural network model for neonatal disease diagnosis." *International Journal of Artificial Intelligence and Expert Systems (IJAE)* 2.3 (2011): 96-106.
13. Wang, Tinghua, et al. "Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels." *Expert Systems with Applications* 37.9 (2010): 6663-6668.
14. Aslandogan, Y. Alp, and Gauri A. Mahajani. "Evidence combination in medical data mining." *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*. Vol. 2. IEEE, 2004.
15. Karthikeyani, V., et al. "Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction." *International Journal of Computer Applications* 60.12 (2012).
16. Gupta, Shelly, Dharminder Kumar, and Anand Sharma. "Performance analysis of various data mining classification techniques on healthcare data." *International journal of computer science & Information Technology (IJCSIT)* 3.4 (2011): 155-169.

January-February 2020
ISSN: 0193-4120 Page No. 8335 - 8342

17. Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." *International journal on recent and innovation trends in computing and communication* 2.10 (2014): 3003-3008.
18. Amin, Syed Umar, Kavita Agarwal, and Rizwan Beg. "Genetic neural network based data mining in prediction of heart disease using risk factors." *2013 IEEE Conference on Information & Communication Technologies*. IEEE, 2013.
19. Lakshmi, K. S., and G. Vadivu. "A novel approach for disease comorbidity prediction using weighted association rule mining." *Journal of Ambient Intelligence and Humanized Computing* (2019): 1-8