RESEARCH ARTICLE OPEN ACCESS

COMPREHENSIVE STUDY ON EFFICIENT DIABETES DISEASE PREDICTION WITH USING VARIOUS ADVANCE DECISION TREE MODELS ALGORITHMS

[1]Amireddy Srinish Reddy, Sanjay Pachouri[2]

[1]RESEARCH SCHOLAR, DEPT OF CSE, SUNRISE UNIVERSITY, ALWAR, RAJASTHAN, INDIA
[2]RESEARCH SUPERVISOR, DEPT OF CSE, SUNRISE UNIVERSITY, ALWAR, RAJASTHAN,INDIA

Abstract:-

Data mining has from long been individuals companion and deliverer from numerous points of view and one of the techniques is through decision making. With expanding wellbeing concerns diabetes has a cutting edge scourge with millions around the globe influenced. Data mining is developing in significance to taking care of such true disease issues through its devices. The accompanying study proposes to utilize the UCI vault Diabetes dataset and create DECISION TREE models for classification using LAD tree, NB tree, and a Genetic J48 tree. The DECISION TREE based classifier models study incorporates various parameters like computational overheads consumed, highlights, productivity, and exactness and gives the outcomes. This hereditary J48 display precisely arranges the dataset when contrasted and the other two models as far as exactness and speed.

***Keywords*:- Diabetes, M-TREE, J48, C4.5, FB Tree, Classification.**

I. INTRODUCTION

Diabetes is a gathering of metabolic issue in which there are high glucose levels over a prolonged period. Diabetes mellitus is a mind boggling gathering of diseases caused by various reasons. People experiencing diabetes have hyperglycemia (high glucose) either on the grounds that there is the low production of insulin or body cells don't utilize the created insulin. There is three fundamental sort of diabetes. These are typ1, type2, gestational diabetes. Common reasons for diabetics incorporate expanded recurrence of urination, particularly around evening time, much of the time feeling parched, shortcoming and weakness, unexplained loss of weight, genital tingling or thrush, obscured vision, increment in recuperating time of cuts and wounds.

This paper utilizes data mining and measurable investigation procedures to distinguish the overwhelming elements causing diabetes in people. As of now factors that are believed to be critical like age, BMI, High Cholesterol, Hyperthyroid, Hypertension, Sleeplessness, Arthritis, Vision issues, Skin issues, Kidney issues, Amputation because of unhealed injuries, Numbness/Tingling/Irritation are only considered. Among these, the most critical ones prompting diabetes are recognized. Qualities of each noteworthy factor are examined in diabetic

and non-diabetic people prompting learning disclosure of very critical reasons for diabetes as a rule. The whole data set is likewise subject to classification using two diverse DECISION TREEinduction techniques and a near study of the strategies is additionally attempted. An endeavor is additionally made to foresee diabetics in people using the learning increased through DECISION TREEinduction and this is utilized to construct a product show for the equivalent. Association decides that administer diabetes are additionally created using Association rule mining. Grouping is utilized to perform distinct data mining.

Data mining is progressively connected to varying backgrounds as the information consequently produced is connected to take care of the issue. All things considered, it is significant in medicinal services where the attention is on diabetes. India as a nation is touted to be the worldwide capital of the sugar – diabetes by having the biggest number of individuals influenced by this disease.

Subsequently the focal point of this study is turned towards the application of the data mining instrument specifically DECISION TREEclassification to the diabetes dataset. A DECISION TREEis a wonderful classification display. The PIMA Indian database is considered here which is taken from the UCI vault. The decision classifiers utilized here for the object are LAD [Least Absolute Deviation] M-TREE, NB [Navies Bayes] DECISION TREEand the Genetic J48 M-TREE, where using the dataset the highlights are separated from the dataset to give decisions. In the mean time, parameters are broke down where the quantity of highlights produced, precision, computational burdens and productivity in tree formation are recorded. Consequently this study centers around the significance of DECISION TREEdata mining structure in diabetes.

II. LITERATURE SURVEY

The diabetes chance score demonstrate considering Age, BMI, midriff outline, history of antihypertensive medication treatment, high blood glucose, physical action, and day by day consumption of organic products, berries, or vegetables as clear cut factors [1]. The creators show chance elements, in the last model, including circulatory strain, cholesterol, back torment, greasy nourishment, weight list or liquor list [2]. They consider the clinical factors, for example, BMI, circulatory strain, glycemia, cholesterol, or cardiovascular hazard in the model [3]. There are various data mining procedures and calculation utilized for discovering diabetes. Neural Network, Artificial neural fluffy obstruction framework, K-Nearest-Neighbor (KNN), Genetic Algorithm, Back Propagation calculation etc[4]. These procedures and the algorithms give the better outcome to the general population and the specialists in regards to the conclusion of diabetes. From these outcomes, the general population can anticipate he is influenced with the diabetes or not.

Anticipate the human utilize the contributions from complex tests conducted in labs and furthermore foresee the disease dependent on hazard factors, for example, tobacco smoking, liquor consumption, age, family ancestry, diabetes, hypertension, elevated cholesterol, physical

inertia, weight. In this study, we are using three various types of grouping procedures named as Hierarchical bunching; Density-based bunching, and Simple K-Means grouping. [5]Weka is utilized as an apparatus. They registered another variable age new as an ostensible variable, separating into three gathering's young age, middle age and maturity and the objective variable diabetes_diag_binary is a double factor. They discovered 34% of the population whose age was underneath 20 years was not influenced by diabetes. [6]33.9% of the population whose age was over 20 and underneath 45 years was not influenced by diabetes. 26.8% of the population whose age was over 45 years was not diabetic.

To dodge the perilous complications of diabetes, patients should control a blood glucose level as the HbA1c (collective blood glucose level for 3 months) ought to be under 7%. [7] In this paper another anticipated model has been produced by using data mining procedures [8]. The model plans to characterize diabetic patients into two classes which are: under control (HbA1c < 7%) and crazy (HbA1c > 7%). The medicines gets ready for 10061 diabetic patients were utilized to fabricate the model. After a comprehensive study for classification systems, three algorithms have been chosen which were Naive Bayes, Logistic and J48 [9]. By using WEKA application, the model has been actualized.

Considering the predominance of diabetes among people the study is gone for discovering the attributes that decide the nearness of diabetes and to track the most extreme number of people experiencing diabetes with 249 population using WEKA instrument [10]. Classification is a strategy used to separate models portraying vital data classes or to foresee the future data.

Jianchao Han [5] utilized WEKA DECISION TREEto fabricate and anticipate type 2 diabetes dataset which considered only the Plasma Insulin quality as the primary trait while disregarding alternate properties given in the dataset. Asma B.M. Patil [7] performed distinctive classification algorithms with changing correctnesses and recommended enhanced prediction exactness using weighted minimum squares SVM.. A.Aljarullah [6] additionally utilized WEKA DECISION TREEclassifier on the diabetes dataset with association rule being executed to produce a combination of qualities. E.G.Yildirim [8] proposed two models to be specific Adaptive Neuro-Fuzzy Inference System – 1-Rough Set 2-ANFIS models. Parthiban et al. [9] in his exploration work proposed diabetic patient getting heart assault disease using Naive Bayes data mining classifier method by using a base preparing dataset. Huang Y.et al. [10] in his work proposed decision support, prediction and estimation by separating designs from vast data sets. Huang, Feixiang; Wang, Shengyong[11] contemplated a diabetic person having vein nerves harm, eye retinopathy, coronary illness, kidney disease and so forth. Gaganjot Kaur anticipated a changed J48 Classification Algorithm for the Prediction of Diabetes [12].

Hussein Asmaa S, Wail M [13] et al expressed that by and by 246 million individuals are having diabetes or its related variations and which will twofold by 2025 coming to 500 million soon contacting 1billion. DECISION TREEAlgorithm is to discover the manner in which the properties and highlights removed for a settled dataset. Via preparing datasets, the classes for

recently created cases are being discovered [20], which thusly produces a prediction for test data inputs.

III. Related work

A. Data Sources

There are various elements causing diabetics in people. Test population consisting of 337 patients who are getting treated in an I.S. nursing home research focus in Trichy are taken. Physical and environmental history of guardians and kin are considered. Data relating to 202 diabetics and 135 non-diabetics were gathered for similar properties. A questionnaire was made consisting of various parameters with respect to the elements affecting diabetics. Those questionnaires were appropriated to the patients who are visiting the middle for week by week/monthly registration. The response was fulfilling. Out of a few autonomous characteristics gathered from outpatients, plainly only a portion of the elements truly assume an indispensable job in causing diabetics in people**.**

B. Measurable Analysis

Weka is a workbench that contains a collection of visualization instruments and algorithms for data examination and prescient demonstrating, together with graphical UIs for simple access to these functions. Weka bolsters a few standard data mining assignments, all the more particularly, data preprocessing, bunching, classification, regression, visualization, and highlight selection. WEKA is a prevalent data mining instrument. It is utilized to break down the most critical elements causing diabetics. It is additionally used to play out a measurable investigation of every individual characteristic.

C. Data Mining

Data Mining might be characterized as the composite of systems utilized to identify designs in vast data sets to extricate shrouded snippets of information. It is a genuinely new strategy used to find concealed examples in the conduct of data. While analysts have for quite a while been performing Data Mining physically, late advances in measurable programming, processing force, and capacity abilities have empowered us to effortlessly and precisely extricate concealed examples from databases.

1) Use of classification methods

Classification is the most commonly connected data mining method, which utilizes an arrangement of pre-ordered precedents to build up a model that can characterize the population of records on the loose. The data classification process includes learning and classification. In taking in the preparation data are investigated by the classification calculation. In classification, test data are utilized to gauge the precision of the classification rules. In the event that the

precision is satisfactory the principles can be connected to the new data tuples. DECISION TREEinduction is a mainstream strategy utilized for classification and prediction.

2) Use of k-implies bunching

Bunching can be said as identification of comparative classes of articles. By using grouping strategies we can additionally recognize thick and inadequate regions in question space and can find generally speaking distribution example and correlations among data properties. Classification approach can likewise be utilized for successful methods for recognizing gatherings or classes of the question however it turns out to be exorbitant so grouping can be utilized as a pre-preparing approach for quality subset selection and classification.

3) Use of associations rule mining

Association Rule Mining is a prevalent and very much examined strategy for finding fascinating relations between factors in expansive databases. To choose fascinating guidelines from the arrangement of every single conceivable standard, constraints on various proportions of criticalness and intrigue can be utilized**.**

IV. M-TREE

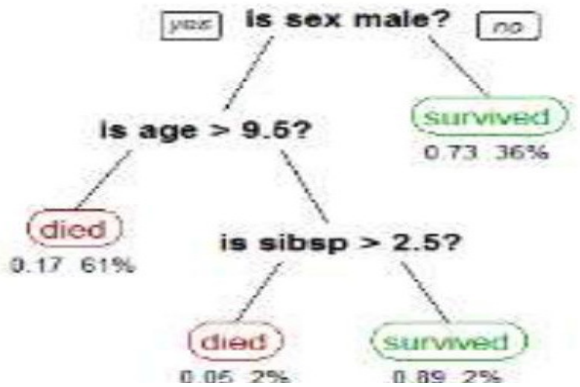

Figure 1 A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.Proposed work

A DECISION TREEis a straightforward depiction for characterizing tests. Here the Solitary point is classification.In this philosophy, accept that the majority of the info highlights have limited unmistakable areas. Every component of the space of the classification is known as a class. In a DECISION TREEwhich otherwise called classification tree, each inside (non-leaf) hub is ordered with an information highlight. The circular segments are named with the majority of the conceivable estimations of the yield highlight or the bend prompts a subordinate decision hub on a different information include. Each leaf of the tree is named with a class or a probability scrambling over the classes [12].

4.1 AlgorithmS for DECISION TREEInduction

The DECISION TREEinduction calculation chips away at the concept of recursively, by picking the best quality to partition the data and growing the leaf hubs of the tree until the point that the completion condition is met. The decision of the best part test condition is dictated by likening the pollution of kid hubs. A short time later construction of the M-TREE, a tree-pruning stage can be expert to lessen the measure of the M-TREE. M-TREEs that are excessively immense are slanted, making it impossible to a phenomenon recognized as overfitting. Pruning underpins by trimming the parts of the underlying tree in a way that expands the interpretation capacity of the M-TREE. [13]

The data are gathered from constant UCI store and it conforms to Type II diabetes dependent on the given properties. The dataset has ten characteristics which anticipate the onset of diabetes in grown-ups. The properties are given dependent on data types. The data set depends on both numerical and ostensible data types. Here the Patient Id, Plasma insulin-glucose data are given in the numeric data type and BMI, Blood Pressure, sexual orientation data are given in ostensible kind.

DIABETES DATASET

The factors being examined is whether the tolerant shows diabetes as indicated by WHO criteria

Results: The parameters utilized are genuine esteemed somewhere in the range of 0 and 1, changed into a paired decision using a cutoff of 0.448. There are 57 preparing cases in the PIMA Indian dataset, there are 768 instances and 9 Attributes like Number of times pregnant, Plasma glucose concentration, oral glucose test, a 2-Hour serum insulin (mu U/ml), Diastolic circulatory strain (mm Hg), the Triceps [skin fold] thickness estimated in mm, Diabetes family function, patients Age in years lastly the Class [whether tried positive or tried negative] and Body mass index[BMI] which is the weight in kg isolated by the stature in m .Class Distribution: Class esteem 1 having diabetes and 0 implies negative diabetes.

Class Value Number of examples

500

268

LOAD DATA SET

In this venture, the PIMA Indians Diabetes dataset is contribution from the UCI storehouse to the algorithms.

Superfluous FEATURE REMOVAL First takes out the unimportant highlights in the data set using the component expulsion calculation and after that discover pertinence between each element and the objective element which is to compute the separation among every single changing element using Euclidian and Manhattan models. In the event that the separation is observed to be more noteworthy than the edge esteem, it is important else it is an immaterial component. Therefore the immaterial highlights are evacuated and the significant highlights are acquired.

Evacuating REDUNDANT FEATURE There are three stages.

i. Minimum Spanning Tree Construction

ii. Tree Partition - Clustering

iii. Representative component selection

V. RESULTS AND DISCUSSION

It has been discovered that of the three DECISION TREE classification algorithms, (I) diverse qualities result in various classification exactnesses; (ii) there is where corresponding classification precision; and (iii) the qualities, in which the best classification highlights are got, are distinctive for both the data sets, the altered – hereditary J48 DECISION TREE display is observed to be the best. The outcomes and discoveries are classified underneath with proper diagrams.

PERFORMANCE STUDY OF ALGORITHM

Algorithms	**Accuracy**	**Accuracy**
LAD **DECISION TREE**	69.685	0.33
ADT **DECISION TREE**	70.866	0.34
J48 **DECISION TREE**	64.173	0.29
M-**TREE**	67.176	0.28

Table 1 Accuracy of different **DECISION TREE***Algorithms*

CONCLUSION

The study in this way effectively demonstrates the comparison of the three DECISION TREEclassification models for the UCI archive diabetes dataset and demonstrates the tree structure framed empowering clients to settle on precise decisions dependent on the information parameters. Further, the hereditary J48 show is observed to be the most efficient and exact when contrasted and the other two decision models as far as time, exactness and highlights. In future, the models may incorporate other decision emotionally supportive networks with parameters from clinical tests helping prediction of diabetes.

REFERENCES

1. Han J. Kamber. M, "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers.
2. Margaret H. Dunham, "Data Mining Techniques and Algorithms", Prentice Hall Publishers. S.Priya, "An improved data mining model to predict the occurrence of Type 2 diabetes" ICON3C 2012, Proceedings published in IJCA.
3. T.Mitchell, "Machine Learning", McGraw -Hill, New York- 2 edition, 2010
4. JIanchao Han, Juan C.Rodriguze, Mohsen Beheshti, "Diabetes Data Analysis and Prediction
5. model discovery" IEEE, Second International conference on future generation communication and networking, pp 96-99,2011.
6. K. Meena, N. Vijayalakshmi, “An Analysis of Risk Factor for Diabetes using Data Mining Approach”, Indian Journal of Public Health Research and Development, Vol. 6, Issue No. 2, pp 112-117, April-June 2015.
7. [Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, “Diagnosis of Diabetes Mellitus based on Risk Factors”, International Journal of Computer Applications, Vol.10, Issue No.4, November.2010.
8. [P.Thangaraju, B.Deepa, T.Karthikeyan, “Comparison of Data mining Techniques for Forecasting Diabetes Mellitus”,
9. J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," Diabetes Care,
10. P. Radha, Dr. B. Srinivasan, “ Predicting Diabetes by consequencing the various Data mining Classification Techniques”

11. Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.

12. Varsha Kavi and Divyesh Joshi , "A Survey on Enhancing Data Processing of Positive and Negative Association Rule Mining", International Journal of Computer Sciences and Engineering, Volume-02,Issue-03, Page No (139-143), Mar -2014.

13. Sharma, Trilok Chand, and Manoj Jain. "WEKA Approach for Comparative Study of Classification Algorithm."

14. P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool". International Journal of Scientific & Engineering Research, Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool.

15. K. R. Lakshmi and S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013.