International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P) Volume-9, Issue-6, June 2019

Analysis Of Decision Tree For Diabetes Prediction

Karnika Dwivedi, Dr.Hari Om Sharan, Vinod Vishwakarma

Abstract— **Data mining has from long been human beings friend and savior in numerous ways and one of the methods is through decision making. With increasing health issues polygenic disorder incorporates a modern-day scourge with millions round the world affected. Data mining is growing in connection to finding such globe unwellness issues through its tools. The following study proposes to use the UCI repository polygenic disorder dataset and generate call tree models for classification mistreatment LAD tree, NB tree and a Genetic J48 tree. The decision tree based classifier models study includes various parameters like computational overheads consumed, features, efficiency and accuracy and provides the results. This genetic J48 model accurately classifies the dataset compared with the opposite 2 models in terms of accuracy and speed.**

Index Terms— **Diabetes, Decision Tree, J48, C4.5, FB Tree, Classification**

I. INTRODUCTION

Data mining is increasingly applied to all walks of life as the information thus generated is applied to solve the problem. Even so it is profound in health care where the focus is on diabetes. India as a country is touted to be the global capital of the sugar – diabetes by having the largest number of people affected by this disease. Hence the main target of this study is turned towards the appliance of the information mining tool specifically call tree classification to the polygenic disease dataset. A decision tree is a wonderful classification model. The PIMA Indian info is taken into account here that is taken from the UCI repository. The decision classifiers used here for the aim square measure LAD [Least Absolute Deviation] call tree, NB [Navies Bayes] decision tree and the Genetic J48 decision tree, where using the dataset the options square measure extracted from the dataset to offer choices. Meanwhile parameters square measure analyzed wherever the quantity of options generated, accuracy, computational hundreds and potency in tree formation square measure recorded. Thus this study focuses on the importance of call tree data processing structure in polygenic disease.

II. RELATED WORK

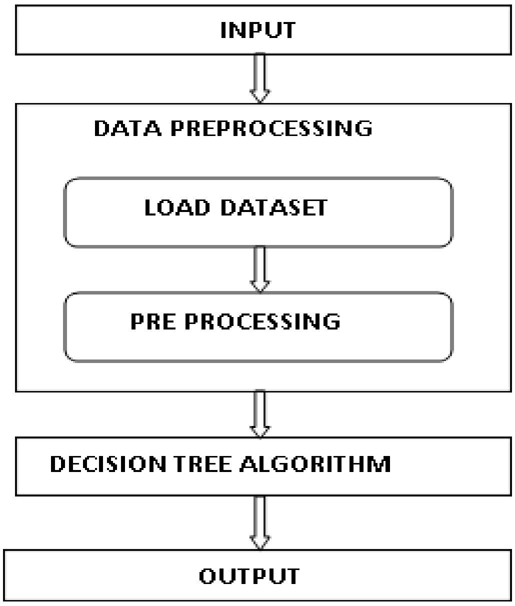
Earlier lots of works have been done in this field and to summarize a few of them have been mentioned below. Jianchao Han [5] used WEKA decision tree to build and predict type 2 diabetes data set which considered only the Plasma Insulin attribute as the main attribute while neglecting the other attributes given in the dataset. Asma B.M.Patil [7] performed completely different classification algorithms with variable accuracies and steered improved prediction accuracy exploitation weighted statistical method SVM. A.Aljarullah [6] also used WEKA decision tree classifier on the diabetes information set with association rule being enforced to get a mix of attributes. E.G.Yildirim [8] proposed two models namely Adaptive Neuro Fuzzy Inference System – 1-Rough Set 2- ANFIS models. Parthiban et al. [9] in his research work proposed diabetic patient getting heart attack disease using Naive Bayes data mining classifier technique by using a minimum training dataset. Huang Y.et al. [10] in his work projected call support, prediction and estimation by extracting patterns from large data sets. Huang, Feixiang; Wang, Shengyong[11] studied a diabetic person having blood vessel nerves damage, eye retinopathy, heart disease, kidney disease etc. Gaganjot Kaur foretold a changed J48 Classification formula for the Prediction of polygenic disorder [12]. Hussein Asmaa S,Wail M [13] et al stated that presently 246 million people are having diabetes or its related variants and which will double by 2025 coming to 500 million soon touching 1billion. Decision Tree formula is to seek out out the manner the attributes and options extracted for a hard and fast dataset. By training datasets the classes for newly generated instances are being found [20], which in turn generates prediction for test data inputs.

Karnika Dwivedi, Student of M.Tech (C.S.E), Department of computer science & engineering Faculty of Engineering at Rama University, Kanpur
Dr.Hari Om Sharan, Head of Department **,** Department of computer science & engineering Faculty of Engineering at Rama University, Kanpur
Vinod Vishwakarma, Assistant Professor at Krishna Institute of Technology, Kanpur

III. METHODOLOGY

Proposed Framework

The data are collected from real time UCI repository and it conforms to Type II diabetes based on the given attributes. The data set has 10 attributes that predict the onset of polygenic disorder in adults. The attributes descriptions are entailed below.



The attributes are given based on data types. The data set is based on both numerical and nominal data types. Here the Patient Id, Plasma insulin glucose data are given in the numeric data type and BMI, Blood Pressure, gender data are given in nominal type.

IV. DIABETES DATASET

The variables being investigated is whether the patient shows diabetes according WHO criteria Results: The parameters used are real-valued between 0 and 1, transformed into a binary decision using a cutoff of 0.448. There are 576 coaching instances within the PIMA Indian knowledge set, 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 blood pressure (mm Hg), the Triceps [skin fold] thickness measured in mm, Diabetes pedigree function, patients Age in years and finally the Class [whether tested positive or tested negative] and Body mass index [BMI] which is the weight in kg divided by the height in m .Class Distribution: Class value 1 means having diabetes and 0 means negative diabetes.

Class Value	Number of instances
0	500
1	268

GENETIC J48 TREE ALGORITHM

Procedure DecisionTreeLearner(X,Y,)
Inputs
X: set of input features, $X=\{X_1,.\ \ ,X_n\}$

Y: feature boundary

E: set of trainingexamples

Output

decision tree

if stop criteria is true **then**
return *pointEstimate(Y,E)*
else
Select feature $X_i \in X$, with domain $\{v_1,v_2\}$
let $E_1=\{e \in E: val(e,X_i)=v_1\}$
let $T_1=DecisionTreeLearner(X \setminus \{X_i\},Y,E_1)$
let $E_2=\{e \in E: val(e,X_i)=v_2\}$
let $T_2=DecisionTreeLearner(X \setminus \{X_i\},Y,E_2)$
return $\langle X_i=v_1, T_1, T_2\rangle$

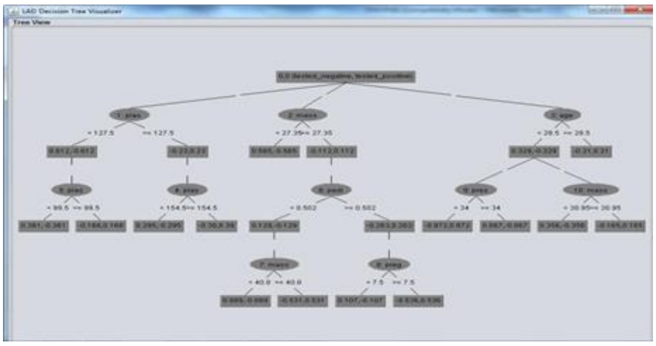
Procedure DecisionTreeClassify(e,X,Y,DT)
Inputs
X: set of input features, $X=\{X_1,.\ \ ,X_n\}$
Y: target feature

e: example to classify

DT: decision tree
Output
prediction on Y for example e

Local
S sub branch of DT
$S \leftarrow DT$
while S is an **an enclosed** node of **the shape** $\langle X_i=v,T_1,T_2\rangle$ **do**
if $val(e,X_i)=v$ **then**
$S \leftarrow T_1$
else
$S \leftarrow T_2$

return S

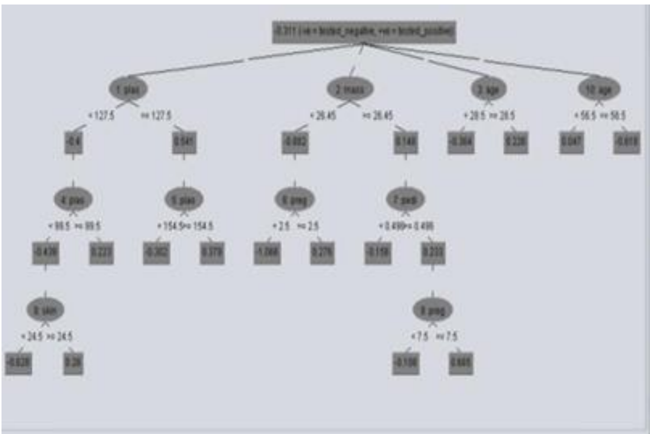


ALGORITHM STEPS

1. Check for base cases - I
2. For each attribute - *a*

Find the information gain from splitting on *a*

3. Let *a_best* be the attribute with the highest standardized gain
4. Create a decision *node* that splits on *a_best*
5. Recursive on the sublists obtained by split on the best and add those nodes as children of nodes.



LOAD DATA SET

In this project, the PIMA Indians Diabetes dataset is input from the UCI repository to the algorithms.

IRRELEVANT FEATURE REMOVAL

First eliminate the digressive options within the knowledge set mistreatment the feature removal formula and so resolve connectedness between every feature and also the target feature that is to calculate the distance between each and every varying feature using Euclidian and Manhattan models. If the distance is found to be greater than the threshold value then it is relevant else it is irrelevant feature. Thus the irrelevant features are removed and the relevant features are obtained.

REMOCING REDUNDANT FEATURE

There are three steps.

i. Minimum Spanning Tree Construction
ii. Tree Partition -- Clustering
iii. Representative feature selection

V. RESULTS AND DISCUSSION

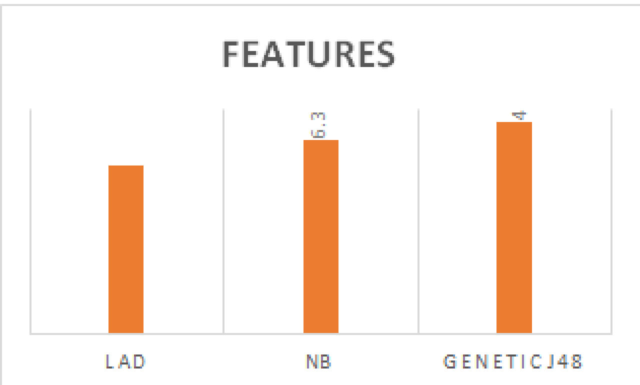
1) It has been found that of the **3 call** tree classification algorithms, (i) different values result in different classification accuracies; (ii) there is a value where corresponding classification accuracy; and (iii) the values, in which the best classification features are got, are different for both the data sets, the modified – genetic J48 Decision Tree model is found to be the best. The results and findings are tabulated below with appropriate charts

Table 1: Parameters for all three algorithms

ALGORITHM S	LAD	NB	Genetic J48
FEATURES	80.4	86.3	90
EFFICIENCY	90.4	94.8	97.2
ACCURACY	89.6	98.6	95.8

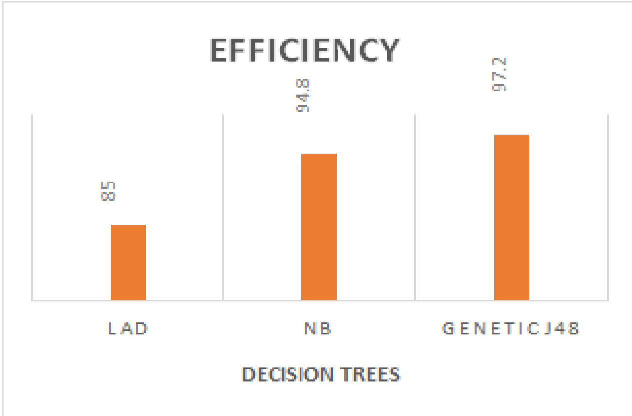
In terms of options Genetic J48 Tree shows the utmost potency with ninety four followed by NB tree and at last LAD with seventy fifth

Chart 1 showing the features generated



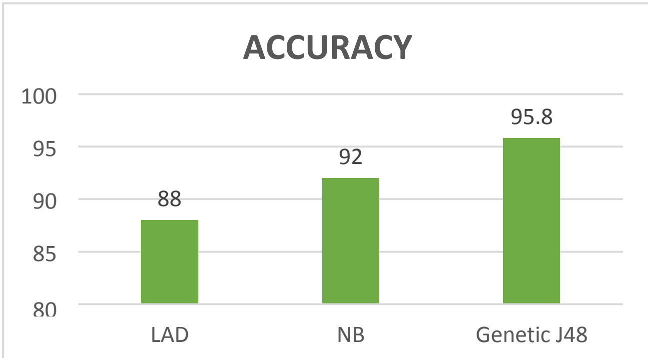
In terms of potency once more Genetic J48 Tree shows the utmost potency with ninety seven.2% followed by NB tree with 94.8% and finally LAD with 85%

Chart 2 Showing efficiency of the algorithms



Finally when it comes to accuracy the genetic J48 is most accurate as it comes with 95.8% while NB has 92% and LAD Tree has 88% accuracy.

Chart 3 Showing Accuracy of the algorithms



This means the results square measure the simplest, and the performance is optimal for the genetic J48 tree. For each of the three decision tree algorithms, although the values where the best classification accuracies are obtained are different for various parameter in the dataset, the genetic J48 is the preferred model because the classification accuracies are the best among the lot. When determining the value, besides classification accuracy, the proportion of the selected features are taken into account as well.

VI. CONCLUSION

The study thus successfully shows the comparison of the three decision tree classification models for the UCI repository diabetes dataset and shows the tree structure formed enabling users to take accurate decisions based on the input parameters. Further the genetic J48 model is found to be the foremost economical and correct in comparison with the opposite2 call models in terms of your time, accuracy and features.

In future the models could embrace alternative call support systems with parameters from clinical tests aiding prediction of the polygenic disease.

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.

[3]S.Priya, "An improved data mining model to predict the occurrence of Type 2 diabetes" ICON3C 2012, Proceedings published in IJCA.

[4]T.Mitchell, "Machine Learning", McGraw -Hill,
New York- 2 edition, 2010

[5]JIanchao Han, Juan C.Rodriguze, Mohsen Beheshti, "Diabetes Data Analysis and Prediction model discovery" IEEE, Second International conference on future generation communication and networking, pp 96-99,2011

[6]Asma A.Aljarullah, "Decision tree discovery for the diagnosis type 2 diabetes" IEEE, International conference on innovation in information technology, pp 303-307, 2011

[7]B.M Patil, R.C.Joshi, Durga Toshniwal, "Hybrid prediction model for type II diabetic patients", Expert Systems with applications, science direct, pp 8102-8108, 2012

[8]E.G.Yildirim, A.Karachoca and T.Uear, "Dosage Planning for diabetes patients using data mining methods" science direct, procedia computer science, pp. 1374-1380, 2011.

[9] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[10] Huang, Y et al. "Feature selection and classification model construction on type 2 diabetic patients", Journal of Artificial Intelligence in medicine 2012.

[11] Huang, Feixiang; Wang, Shengyong; Chan, ChienChung, "Predicting disease by using data mining based on healthcare information system," Granular Computing (GrC), 2012 IEEE International Conference on , vol., no., pp.191,194, 11-13 Aug. 2012

[12] Gaganjot Kaur “Diabetes Research” Department of Computer Science and Diabetes Federation

[13] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013, 6, 85-97