International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.8, pp : 1054-1059 1 Aug 2014

Expert System Design to Predict Heart and Diabetes Diseases

ShravanKumar Uppin, Anusuya M A
Department of Computer science and Engineering,Sjce,Mysore, India
shravan4uppin@gmail.com, anusuya_ma@yahoo.co.in

Abstract— ***The Objective of this paper is to design an expert system that predicts the heart disease and diabetes disease with reduced number of attribute using data mining technique. Classification of knowledge objects is a knowledge mining and knowledge management process used in grouping similar knowledge objects together. There are plenty of classification algorithms available in literature but decision tree is the most often used because of its ease of implementation and simpler to understand, when compared to other classification algorithms. There are many classifiers but we have used C4.5 for more accuracy and less run time. The decision tree algorithm has been applied on the knowledge of heart and diabetes disease to foretell whether diseases present or not. The Simulation result obtained from the model enables us to establish significant patterns and relationships between the medical factors and clinical factors.***

Keywords-Data Mining, Artificial Intelligence, Decision Tree, Heart Disease, Diabetes ,Classification,c4.5 Algorithm.

1. Introduction

Expert or knowledge-based systems are the commonest blazon of Artificial Intelligence systems in accepted analytic use. They accommodate medical knowledge, usually about an actual accurately authentic task, and are able to acumen with abstracts from alone patients to appear up with articular conclusions. Although there are abounding variations, the knowledge within a specialist method is usually represented in the type of a set of rules. Machine learning frameworks might be utilized to create the information bases utilized by master frameworks. Given a set of clinical cases, a machine learning framework can deliver an efficient depiction of those clinical gimmicks that interestingly describe the clinical conditions. This information could be communicated as basic controls, or regularly as a choice Tree.

Data mining techniques plays a significant function in seeing forms and extracting knowledge from a large mass of data [1]. It is very helpful to provide better patient care and effective diagnostic capabilities [2-4]. Various data mining techniques are employed in the diagnosis of heart disease such as: Genetic algorithm, classification via clustering, direct kernel self-organizing map, naïve Bayes, decision tree, neural network, core density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies [5-6].Among these classification algorithms decision tree algorithms is the most commonly used because it is easy to understand and cheap to implement. It provides a modeling technique that is easy for humans to comprehend and simplifies the classification Process [7-9].

Choice tree Algorithms are most generally utilized calculation as a result of its simplicity of usage and simpler to comprehend contrasted with other characterization calculations. The conclusion of the choice tree anticipated the amount of patients who have sickness disease or not. Decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm. The C4.5 decision tree algorithms are applied on the dataset of heart and diabetes to predict the presence or absence of disease.

2. Literature Survey

Agreeing to a recent survey by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), nearly 25 percent of deaths in the age group of 25-69 years occur because of heart diseases. In 2008, five out of the top ten reasons for mortality worldwide, other than injuries, were non-transmissible diseases; this will lead up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [10]. Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non-communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [11].

Cardiovascular disease includes coronary heart disease (CHD), cerebrovascular disease (stroke), Hypertensive heart disease, congenital heart disease, peripheral artery disease, rheumatic heart disease, inflammatory heart disease. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol [12]. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease [13].

Diabetes [14] is an illness that threatens the life of people of all nations. It occurs when the blood sugar in the body is increased above a definite level. It is an illness in which either pancreas in the body is not producing sufficient insulin or cells in the body are not using insulin properly. There's

chiefly types of diabetes, namely, type one and type2 diabetes. Type1 diabetes occurs often in children and type2 diabetes is common among adults. Another kind of diabetes called Gestational Diabetes occurs in the coursework of pregnancy. In most of the cases, it disappears after the infant is born. About 60% of the people suffer from diabetes in Asia [15]. Obesity, fat, intake of large quantity of junk food and physical inactivity are a number of the factors that contributes towards diabetes. Usually people ignore the incidence of this illness. But, in point of fact, diabetes is a serious illness and can lead to other extreme complications if not taken seriously. It can lead to lots of other complications like skin complications, foot complications, eye complications, heart illness and kidney illness etc. [16]. The increasing prevalence of diabetes is also affecting economic gains in various countries. In order predict the disease accurately with less number of parameter and run time it is necessary to identify good parameter of the disease and these are applied on c4.5 algorithm.

3. Methodology

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, which builds decision trees from a set of training data, using the concept of information entropy. The training data is a set $S=_{s1,s2,s3...}$ of already classified samples. Each sample $_{si}$ consists of a p-dimensional vector$(_{x1,i,x2,i,x3,i,.....xp,i)}$ where the $_{xj}$ represent attributes or features of the sample, as well as the class in which $_{si}$ falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain(difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublets.

Algorithm for building decision trees is

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

i. Find the normalized information gain ratio from splitting on *a*

3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recursively apply on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*.

A. Procedural Steps

- In general, if we are given a probability distribution P = (p1, p2, .., pn) then the *Information conveyed by this distribution*, also called *the Entropy of P*, is:

$$I(P) = -(p1*\log(p1) + p2*\log(p2) + .. + pn*\log(pn))---- -(1)$$

For example, if P is (0.5, 0.5) then I(P) is 1, if P is (0.67, 0.33) then I(P) is 0.92, if (1, 0) then I(P) is 0.

- Calculate information Entropy for each attribute. This is used to calculate the gain. Consider the quantity Gain(X,T) defined as

$$Gain(X,T) = Info(T) - Info(X,T)---------- (2)$$

Eqn(2) represents the difference between the *information needed to identify an element of T* and the *information needed to identify an element of T after the value of attribute X has been obtained*, that is, this is *the gain in information due to attribute X*.

- For all the attribute calculate the quantity Gain(X,T) for each Attribute. The attribute with maximum gain is used to split the decision tree. Recursively repeat the procedure.

B. C4.5 Decision Tree Builds

C4.5 Algorithm has applied for heart and diabetes data set for selected optimal attribute and the resulted the decision tree is shown in appendix 1 and appendix 2.

4. Dataset Used

A. *Heart Dataset*

The data used in this study is the Hungarian institute of cardiology. Heart disease data set is available at https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 76 row attributes. However all the published experiment only refers to 11 of them. In that we have selected optimal 7 attribute to build decision tree for the 35 trained data rows.

TABLE I. DATASET FOR HEART DISEASE

Name	Type	Description
Age	Continuous	Age in years
Fasting Blood sugar	Discrete	Fasting Blood sugar>120mg/dl 1=t; 0=f;
Chest Pain	Discrete	Chest pain type 1=asymptomatic 2=atypical angina 3=non-angina
Trestbps	Continuous	RestingBloodPressure (in mm Hg)

Restecg	Discrete	Resting Electrocardiographic result 0=normal 1=having ST-T wave abnormality 2=left ventricular
Thalach	Discrete	Maximum Heart rate Achieved

B. *Diabetes DataSet*

Diabetes data set is used uci machine learning Repository and is available at https://archive.ics.uci.edu/ml/datasets/Diabetes. All the published experiments refers to 8 attribute but we have considered 7 optimal attribute to build the decision tree for 50 trained data rows.

TABLE II. DATASET FOR DIABETES DISEASE

Name	Type	Description
Preg	Continuous	Number of times pregnant
Plas	Continuous	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Pres	Continuous	Diastolic blood pressure (mm Hg)
Skin	Continuous	Triceps skin fold thickness (mm)
Insuc	Continuous	2-Hour serum insulin (mu U/ml)
Mass	Continuous	Body mass index (weight in kg/(height in m)^2)
Pedi	Continuous	Diabetes pedigree function

5. Experimental Results

The result of the experiment is shown in below table III. We have carried out simulation result in order to evaluate the performance and usefulness of c4.5 algorithm for predicting heart disease of patient and compared with [17] of same algorithm c4.5 with reduced number of parameters and less number of trained data rows.

TABLE III. EVALUATION CRITERIA FOR HEART DISEASE

Evaluation criteria	Existing Method	Proposed Method
Timing to build in sec	0.05sec	0.025sec
Correctly classified instances	248	57
Incorrectly classified instances	46	8
Accuracy (%)	84.35%	85.96%

Similarly the result of the experiment is shown in below table IV. We have carried out simulation result in order to evaluate the performance and usefulness of c4.5 algorithm for predicting diabetes disease of patient and compared with [18] of same algorithm c4.5 with reduced number of parameters and less number of trained data rows.

TABLE IV. EVALUATION CRITERIA FOR DIABETES

Evaluation criteria	**Existing Method**	**Proposed Method**
Timing to build in sec	0.08sec	0.03sec
Accuracy (%)	73.828%	83.63%

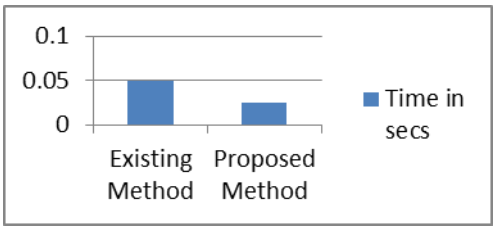

Figure 1-Time Build in sec for Heart Disease

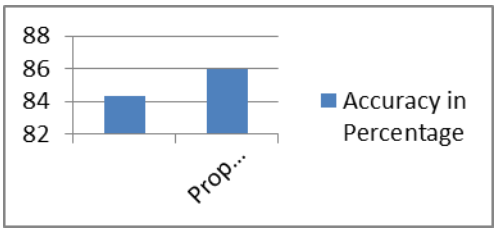

Figure 2-Accuracy in Percentage for Heart Disease

Figure 1 and Figure 2 specifies the time required to build the decision tree and the accuracy in predicting the heart disease using c4.5 algorithm.

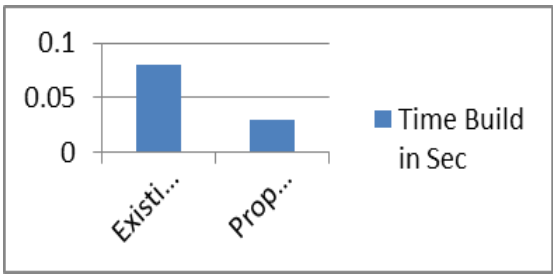

Figure 3-Time Build in sec for Diabetes disease

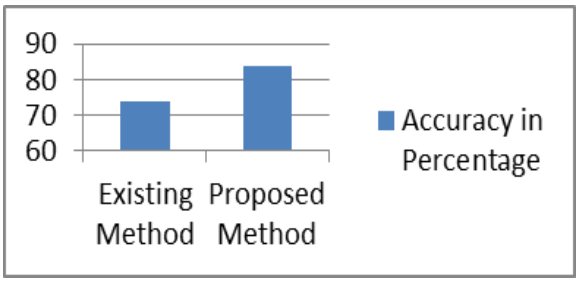

Figure 4-Accuracy in Percentage for Diabetes Disease

Figure 3and Figure 4 specifies the time required to build the decision tree and the accuracy in predicting the diabetes disease using c4.5 algorithm.

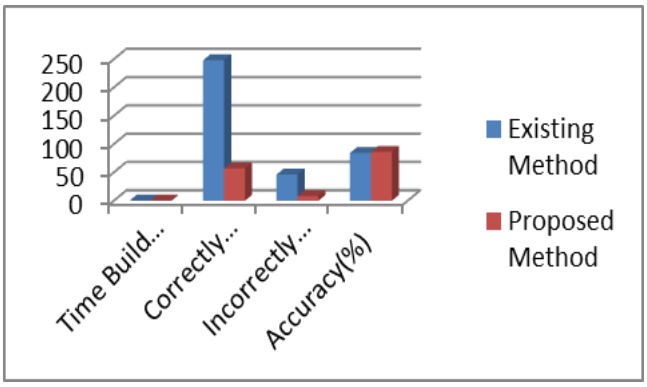

Figure 5-Comparision between Existing and Proposed Method for Heart disease

Figure 5 provides complete information about heart disease. This represent all the parameter i.e. time to build, correctly classified instances, incorrectly classified instances, accuracy over existing and proposed system.

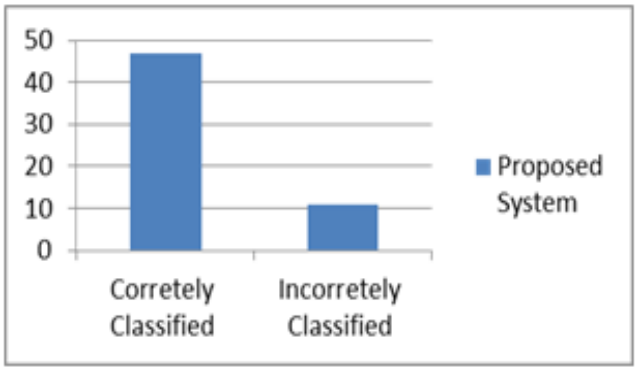

Figure 6 correctly and incorrectly classified instances for diabetes diseased

Figure 6 represents the correctly and incorrectly classified instances for proposed approach of diabetes disease.

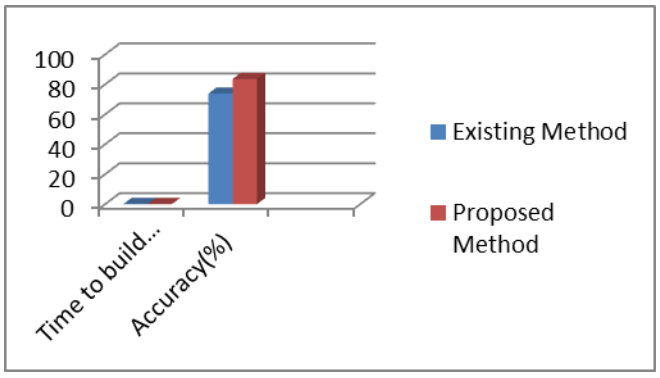

Figure 7-Comparision between Existing and Proposed Method for Diabetes disease

Figure 7 shows the time to build decision tree and accuracy of proposed system over existing system.

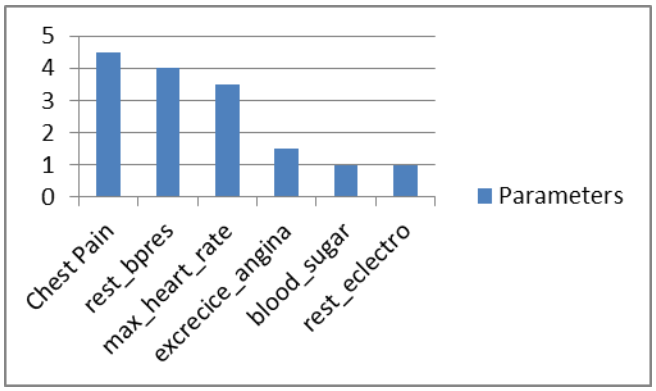

Figure 8- Comparison between importance's of attributes for heart disease

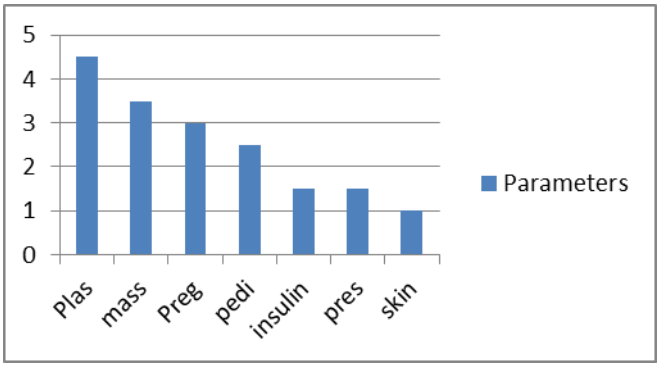

Figure 9- Comparison between importance's of attributes for Diabetes disease

Figure 8 and Figure 9 shows comparison between importance's of different attributes to build decision tree.

6. CONCLUSIONS

In medical diagnosis various data mining techniques are available. In this study, for classification of medical data we employed c4.5 algorithm because it is more accurate and takes less time and it also produce human readable classification rules which are easy to interpret. This paper investigates the optimal parameters which are more prominent in deciding heart and diabetes diseases and avoids the redundant parameters that are not important in classification. This reduced parameter selection helps us to build decision tree in very less time and performance of machine in identifying the diseases accurately. The result shows that our c4.5 algorithm accuracy is 86% and total time to build the model is .025sec in the diagnosis of heart disease patients and 83.63% accuracy and total time to build the model is 0.03 in diagnosis of diabetes.

REFERENCES

i. *Rajkumar A, Reena GS,"Diagnosis of Heart Disease Using Datamining Algorithm",Global Journal of Computer Science and Technology, 10(10):38-43,2010.*

ii. *Beule MD, Maes E, Winter ODE, Vanlaere W, Impe RV,"Artificial neural networks and risk stratification: A promising combination. Mathematical and Computer Modelling", 46(1-2):88-94,2007.*

iii. *Tantimongcolwat T, Naenna T, Ayudhya CIN, Embrechts MJ, PrachayasittikulV,"Identification of ischemic heart disease via machine learning analysis on magnetocardiograms", Computers in Biology and Medicine,38(7):817-825,2008.*

iv. *Anbarasi M, Anupriya E, Iyengar NCHSN," Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology,2(10):5370-5376,2010.*

v. *Kavitha KS, Ramakrishnan KV, Singh MK," Modeling and design of evolutionary neural network for heart disease detection", International Journal of Computer Science Issues (IJCSI),7(5):272-283,2010.*

vi. *Kaur H, Wasan SK," Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science,2(2):194-200,2006.*

vii. *Mythili T,Mukherji D, Padalia N, Naidu A,"A Heart Disease Prediction Model usingSVM-Decision Trees-Logistic Regression (SDL)",International Journal of ComputerApplications,68(16):11-15,2013.*

viii. *Pandey AK, Pandey P, Jaiswal KL, Sen AK," A Heart Disease Prediction Model usingDecision Tree", IOSR Journal of Computer Engineering (IOSR-JCE),12(6):83-86,2013.*

ix. *Anyanwu MN, Shiva SG,"Comparative Analysis of Serial Decision Tree Classification Algorithms", International Journal of Computer Science and Security, 3(3):230-240,2009.*

x. *Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report, 2005.*

xi. *Global Burden of Disease. 2004 update (2008). World Health Organization.*

xii. *Srinivas, K.," Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.*

xiii. *Yanwei Xing, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease", IEEE Transactions on Convergence Information Technology, pp(868 – 872), 21-23 Nov, 2007*

xiv. *http://en.wikipedia.org/wiki/Diabetes_mellitus*

xv. *http://care.diabetesjournals.org/content/34/6/1249.full*

xvi. *http://www.medicalnewstoday.com/info/diabetes/*

xvii. *" Data Mining Approach to Detect Heart Diseases" International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, Month Year, Page: 56-66, ISSN: 2296-1739© Helvetic Editions LTD, Switzerland www.elvedit.com.*

xviii. *"Analysis of a Population of Diabetic Patients Databases with Classifiers using c4.5 Algorithm" World Academy of Science, Engineering and Technology International Journal of Medical, Pharmaceutical Science and Engineering Vol: 7 No: 8, 2013*

Appendix 1

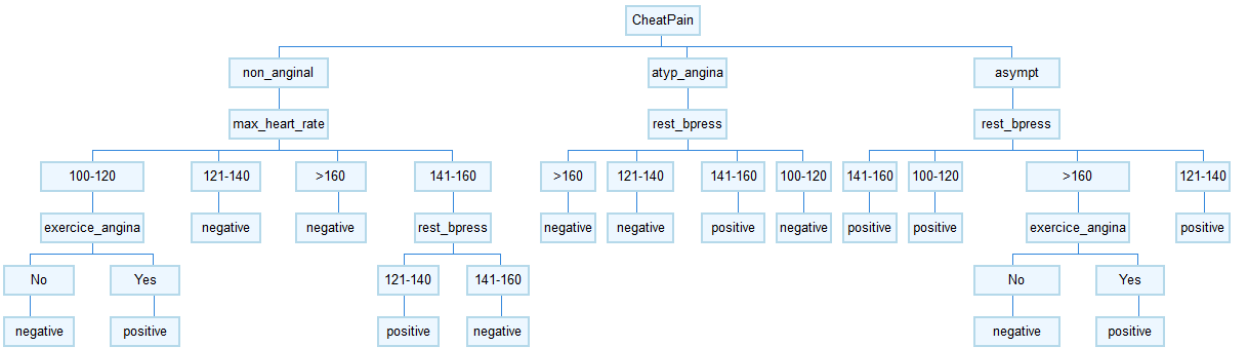

Figure 10-Decision Tree for Predicting Heart Disease

Figure 10 has two solution if the output decision is positive then we consider the heart disease is present, if the output is negative the heart disease is not present.

APPENDIX 2

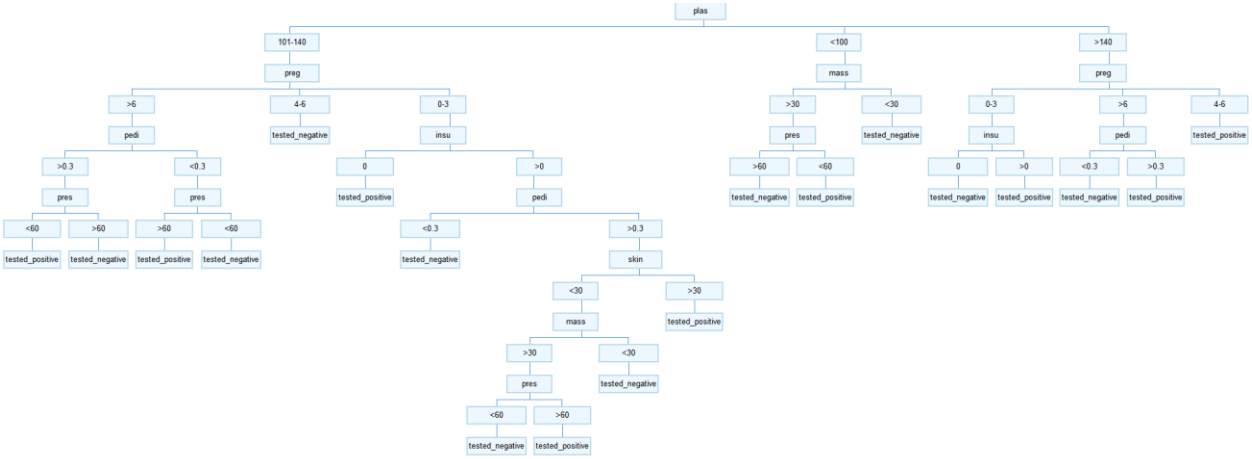

Figure 11-Decision Tree for predicting Diabetes Disease

Figure 11 has 2 solution if the output decision is tested_positive then we consider the diabetes is present, if the output decision is tested_negative then we consider the diabetes disease is not present.