International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 11, November 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

DIAGNOSE AND PREDICT DIABETIC HEART DISEASES USING DATA MINING CLASSIFICATION TECHNIQUES

Jaspreet Singh Phul[1], Sonia Vatta[2]

[1] *Research Scholar, School of CSE, Bahra University, Waknaghat.*
[2]*Assistant Professor, School of Computer Science and Engineering, Rayat Bahra University*

ABSTRACT

Data Mining is the process of extracting relevant information from the databases. According to the World Health Organization (WHO), diabetes is currently one of the biggest health concerns that the world is faced with. So, mining the diabetes data in efficient manner is ambiguous. The Pima Indians Diabetes dataset and heart disease dataset are used in this paper to collect information of patients having diabetes or not. Diabetic patient also have the option to suffer from other diseases like heart disease, eye complications, kidney disease, nerve damage, foot problems, skin complications and dental diseases. In this paper, we will predict whether diabetic patients will get chance to heart disease or not. The same will evaluate on the basis of output as Test Negative for no diabetes, Test Positive for diabetes.

Keywords: ***Data Mining, Diabetes, Heart Problem, Weka Tool, J48.***

I INTRODUCTION

Data Mining is the process of finding useful information and classification based on data patterns from a databases(dataset).The main objective of using data mining is of extraction hidden predictive data/information from the databases. In today, Datamining plays an important role in various areas like education; banking, healthcare etc.Different types of algorithms are widely used for prediction of diseases. Prediction and future rules are found by data analysis. Data mining is a multi-disciplinary area that is collection of databases technology, machine learning and artificial intelligence. Data mining is beneficial in the field of health analysis to reduce patient treatment cost as diagnosis accuracy increased. Data mining uses techniques to examine the data such as classification, clustering, association rules.

II. OBJECTIVES

The Objectives of present work is as follows:

1. To preprocess the patient data for diabetes and heart problem prediction.
2. To present a Decision Tree and Naïve Bayes Model for Diabetes and Heart Problem Prediction.
3. To Identify how decision tree and Naïve Bayes are best to diagnosis of diabetes and Heart Problem.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

III DISEASE OVERVIEW

3.1 Diabetes

Diabetes renowned as Diabetes Mellitus is a metabolic disease in which patient has high blood glucose comparative to improper insulin production in the body which is the cells of hormone that turns glucose to energy and is transported to all cells of the body. The disease is caused depending over the family, environmental factors, genetic makeup and health. In general, patient having diabetes have their blood sugar levels above normal (i.e., 4.4 to 6.1 mmol/L). There are mainly three types of Diabetes: Type 1, Type2, and Gestational Diabetes. Type 1 diabetes is insulin-dependent Diabetes. This type-1 diabetes has been raised in US. Type-2 diabetes is non-insulin dependent diabetes. The organs of the body become insulin resistant and demand for insulin increases rapidly in this case. But at this stage, pancreas does not make the required amount of insulin. Type-2 diabetes is found in people being overweight and physically inactive. This type-2 diabetes can be cured by having proper diet and regular exercise to maintain the blood circulation levels. Gestational diabetes is basically occurred during pregnancy due to high sugar levels in the blood. Regular treatment is suggested for this diabetes cure.

3.2 Heart Disease

Heart disease is the major cause of death in the world. The term "heart disease" involves narrowed or blocked blood vessels which could lead to a heart attack, chest pain (angina) or stroke. According to WHO (World Health Organization) and the CDC, heart disease is the major cause of death in the UK, USA, Canada and Australia. The number of US adults diagnosed with heart disease stands at 26.6 million (11.3% of adult population). In Punjab (India), the estimated heart disease patients are 3-4% in rural areas and 8-10% in urban areas. The heart disease is basically caused by the following factors:

- Blood sugar (diabetes).
- Depression.
- High cholesterol.
- Smoking.
- High blood pressure.
- Age.

3.3 Symptoms and diagnose

The Symptoms of patient suffering from diabetes and heart problem are explained as table 1.Also suggest test for determining whether a patient has diabetes or not, even if have diabetes, whether has a risk of heart problem.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 11, November 2016
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Table 1: Symptoms and Suggested Tests of Diabetes in case of Heart Problem

Sr. No.	Name of the disease	Symptoms	Suggested Tests	Normal range
1.	Diabetes	• Polyuria. • Polyphagia. • Polydipsia. Weight gain or strange loss. slow healing of wounds. • Blurred vision. • Fatigue. • Itchy skin.	Sugar(fasting) Sugar(PP)/(R)	60-120 Mg/dl 110-180 Mg/dl
2.	Heart disease	• Pain in the chest, arm, or below the breastbone • Indigestion. • Sweating. • Dizziness. • Rapid or irregular heartbeats	Serum Cholesterol HDL Cholesterol LDL Cholesterol VLDL Cholesterol Serum Triglycerides CPK-NAC CPK-MB LDH	130-200 mg/dl 35-65 mg/dl Up to 160 mg/dl 10-40 mg/dl 60-170 mg/dl 24-195 U/L Up to 24 U/L 230-460 L

IV RESEARCH METHODOLOGY

Methodology and techniques are used to solve research problem in described way to obtain desired results. The structural design for research methodology steps shown in fig.1.

For this, Methods and tool are used for obtaining the result. Steps are performed under the research methodologies are:

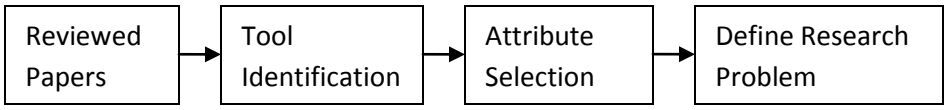

Figure 1: Structural Design for Research Methodology Process

Review Research Papers: Related information of diabetes and heart problem were gathered after reviewed papers and knowing which type of work done and which techniques have adopted.

Identify tools: The next step is performed about identification of tool for solving problem. For this, WEKA tool was selected from all.

Study attributes from database (dataset): Thoroughly studied attributes and schema of dataset to get relevant attributes.

Determine definition of research Problem: Next step is determining the definition of research problem and workflow of the problem for getting proper and accurate results.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

V CLASSIFICATION ALGORITHMS EMPLOYED

5.1. J48 Decision Trees Algorithm

J48 decision tree is easy to implement with least error. J48 is used for feature selection as well as knowledge discovery. This is useful for large dataset in order to produce more accurate results. J48 is an extension of ID3. It works on the attribute values of the available training data. Decision tree is observed by calculating the information gain for all the attributes. If a specific attribute gives an unequivocal, the attribute is terminated and the target value is assigned to it.

5.2. Naïve Bayes Algorithm

The Naïve Bayes Algorithm is a probabilistic algorithm that is sequential in nature. The Naive Bayes algorithm rely upon conditional probabilities which uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event that occurs when given the probability of another event already been occurred.

Prob (B1 given A1) = Prob(A1and B1)/Prob(A1)

Here, B1 represents the dependent event and A1 represents the prior event.

VI DATA DESCRIPTION AND PRE-PROCESSING

The classification type of data mining applied to the data collected from the UCI repository. The Table 2 shows the description of a dataset for diabetes and heart problem diseases.

Table 2: Dataset Description

Data set for Diabetes Disease	**No. of attributes**	**No. of instances**
	8	**768**
Data Set for Heart Disease	12	112

Table 3: The attributes Description of Diabetes disease

Sr. No.	**Attribute**	**Relabeled Values**
1.	Number of times pregnant	Preg
2.	Plasma glucose concentration	Plas
3.	Diastolic blood pressure(mm Hg)	Pres
4.	Triceps skin fold thickness(mm)	Skin

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

5.	2-Hour serum insulin	Insu
6.	Body mass index(kg/m^2)	mass
7.	Diabetes pedigree function	Pedi
8.	Age(years)	age
9.	Class Variable(0 or 1)	class

Table 3: The attributes Description of Heart disease

Sr. No.	Attribute	Relabeled Values
1.	Age	age
2.	Obesity	abes
3.	Heart rate	heart
4.	Chest Pain	chest
5.	Blood Pressure	pres
6.	Blood Sugar	insu
7.	Cholesterol	chols

VII ATTRIBUTE SELECTION

Using WEKA tool, we extract common attributes (age, sugar, pres) from diabetes and heart disease dataset, new dataset(DH_Disease) was created with 12 attributes. This is because of a person is checked diabetic if its outcome may be yes or no. if yes, then the person is affected by the heart disease. If not, the result will depend on the output of diabetic dataset.

Table 4: DH_Disease dataset

Sr. No.	Attribute	Relabeled Values
1.	Age	Age
2.	Obesity	Abes
3.	Heart rate	heart
4.	Chest Pain	chest
5.	Blood Pressure	Pres
6.	Blood Sugar	Insu
7.	Cholesterol	chols
8.	Body Mass Index	mass
9.	Triceps skin fold thickness	Skin
10.	Number of times pregnant	Preg
11.	Plasma glucose concentration	Plas
12.	Class Variable(0 or 1)	class

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

VIII RESULT AND PERFORMANCE

8.1 J48 Decision Tree

Apply 10 fold cross validation in training dataset. Cross-validation is used to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. The confusion matrix is obtained after using J48 decision trees as follows:

a	b	c ← classified as
40	0	4
1	50	0
1	0	14

Where a =Tested_High_affected, b = Tested_negative, c=Tested_Normal_Nonaffeected.

8.2 Naïve Bayes

The confusion matrix is obtained after using Naïve Bayes as follows:

a	b	c ← classified as
39	5	0
5	45	1
4	2	9

Where a =Tested_High_affected, b = Tested_negative, c=Tested_Normal_Nonaffeected.

Accuracy can be calculated by using formula:

$$\text{Accuracy} = \frac{\text{True Positive (TP) + True Negative (TN)}}{\text{True Positive (TP) + False Positive (FP) + True Negative (TN) + False Negative (FN)}}$$

By using this formula, the accuracy is obtained by using J48 and Naïve Bayes as shown in the table 5.

Table5: Accuracy Description and Time taken to build classifiers.

Algorithms	Accuracy	Error rate	Time taken to build the classifier
J48	95.53	4.5354	0.02
Naïve Bayes	85.34	14.4344	0.00

IX CONCLUSION

This work observed the efficiency of Naïve Bayes and J48 Classifiers for diabetes and heart disease prediction. Experimentation is accompanied using the WEKA tool. Effective comparison of both the classifiers has been done using different scales of performance evaluating measurements. Eventually, it is concluded that J48 Classifier

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

performs better than Naïve Bayes for heart disease prediction by taking measures including Classification accuracy, Error rate. The Naïve Bayes performs task in less time to build the model.

REFERENCES

1. Boshra Bahrami et.al,Prediction and diagnosis of heart disease by D.M Techniques,vol2,issue2,JMEST,feb-2015.
2. .M.A.Nishara Banu and B.Gomathy," Disease Forecasting System Using Data Mining Methods", 2014.
3. http://shodhganga.inflibnet.ac.in/handle/10603/45247
4. K.R.Ananthapadmanaban et.al,Prediction of chances-Diabetic Retinopathy using Data Mining Classification Techniques,vol7(10),IJST,October 2014.
5. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
6. 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.
7. https://weka.wikispaces.com/Exporting+Charts+from+the+Knowledge+Flow
8. http://www.openml.org/a/estimation-procedures/1
9. http://www.medicalnewstoday.com/info/diabetes
10. https://en.wikipedia.org/wiki/Diabetes_mellitus
11. P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.
12. http://www.webmd.com/heart-disease/guide/heart-disease-symptoms#1
13. http://www.medicalnewstoday.com/articles/237191.php
14. http://www.mapsofindia.com/my-india/india/prevalence-of-diabetes-in-india