HEART DISEASE AND DIABETES DIAGNOSIS USING PREDICTIVE DATA MINING

Shilpa M[1], C. Nandini[2], Anushka M[3] , Niharika R[4], Palash Singh[5], Parinitha R Raj[6]

[1]Assistant Professor, Dept. of CSE, DSATM, Bangalore, Karnataka, India
[2]Professor, Dept. of CSE, DSATM, Bangalore, Karnataka, India
[3,4,5,6]Student, Dept. of CSE, DSATM, Bangalore, Karnataka, India

***

Abstract - *Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. The hidden patterns and relationships in the data are mostly overlooked. Diagnosing cardio vascular diseases in patients is a difficult task and doctors who can accurately predict such diseases are few in number. Data Mining refers to using a variety of techniques to identify information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, predictions, forecasting and estimation. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. Optimization technique is used in predicting the risk associated with diabetes and heart diseases. The classifier such as Naïve Bayes is used for diagnosis of patients with heart disease. The classifiers are fed with data set of fixed number of attributes.*

***Key Words*: Heart disease, Diabetes Mellitus, Data Mining.**

1. INTRODUCTION

Hidden patterns and relationships can be extracted from large data sources using data mining. Data mining has been applied in several areas of medical services such has discovery of relationships among diagnosis data and stored clinical data. Modern-day medical diagnosis is a very composite process which requires precise patient data, many years of clinical experience and a good knowledge of the medical literature. The health care data collected in a huge amount are, unfortunately, not used to discover the hidden relationships. Doctors always depend on perception and experience rather than on the knowledge rich data masked in the database to take clinical decisions. The information provided by the patients may have redundant symptoms when patients suffer from multiple diseases that may have the same symptoms. The physicians may not be able to diagnose it accurately. Heart disease is a condition that affects the heart. As the "coronary arteries narrow, blood flow to the heart can slow down or stop, causing chest pain, heart attack". Diagnosing heart disease requires highly skilled and experienced physicians. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. Early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors.

Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Here, we are discussing Naïve Bayes algorithm ,association rule mining and clustering of data to improve efficiency and accuracy in predicting the risk associated with diabetes and heart diseases.

2. RELATED WORK

Various data mining techniques such as Naïve Bayes, KNN algorithm, Decision tree, Neural Network are used to predict the risk of heart disease [1]. The patient activity is monitored continuously, if there is any changes occur, then the risk level of disease is informed to the patient and doctor. This paper provides an insight about KNN data mining technique used to predict heart diseases.

Association rules [2] are implications that associate a set of potentially interacting conditions (e.g. high BMI and the presence of hypertension diagnosis) with elevated risk. The use of association rules is particularly beneficial because in addition to quantifying the diabetes risk, they also readily provide the physician with a "justification", namely the associated set of conditions. This set of conditions can be used to guide treatment towards a more personalized and targeted preventive care or diabetes management.

This study implemented a new method of ANN namely Extreme Learning Machine (ELM). ELM is a feed-forward neural network with one hidden layer, or better known as the single hidden layer feed-forward neural network [3]. ELM has advantages in learning speed, and has a better accuracy rate than conventional methods such as Moving Average and Exponential Smoothing [4]

3. METHODOLOGY

3.1 Dataset description

The data set is taken from the Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). The system is validated using data sets from Cleveland. In those datasets, totally, 12 attributes such as Age, sex, chest pain type, resting blood pressure, cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression, slope of the peak exercise ST segment and thal are presented [2].

3.2 List of attributes used

Only 12 attributes used:

1. (age) age in years

2. (sex)

3. (cp) chest pain type

– Value 1: typical angina

– Value 2: atypical angina

– Value 3: non-anginal pain

– Value 4: asymptomatic

4. (restbp) resting blood pressure

5. (chol) serum cholestrol in mg/dl

6. blood sugar >120 mg/dl) (1 = true; 0=false)

7. 19 (restecg) resting electrocardiographic results

– Value 0: normal

–Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV)

– Value 2: showing probable or definite left ventricular

hypertrophy by Estes' criteria

8. (thalach): maximum heart rate achieved

9. (exang) exercise induced angina (1 = yes; 0 = no)

10. (old peak) ST depression induced by exercise relative to rest

11. (slope) the slope of the peak exercise ST segment

– Value 1: upsloping

– Value 2: flat

– Value 3: downsloping

12. (thal) = normal; 6 = fixed defect; 7 = reversable

3.3 Data pre-processing

1. Real world data are generally

i. Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.

ii. Noisy: containing errors or outliers.

iii. Inconsistent: containing discrepancies in codes or names.

2. Tasks in data pre-processing

i. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

ii. Data integration: using multiple databases, data cubes, or files.

iii. Data transformation: normalization and aggregation.

iv. Data reduction: reducing the volume but producing the same or similar analytical results.

3.4 Algorithms and techniques

Naïve Bayesion Classifier:

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes performs consistently after reduction of number of attributes.

According to Bayesian theorem

P (A|B) =P (A)*P (B/A)/P(B),Where P (B|A)=P(A∩B)/P(A)

Based on above formula, Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression it has the excellent interpretability same as decision tree and is able to use previous data to build analysis model for future prediction.

Decision Trees:

- Decision trees are powerful and popular tools for classification and prediction.
- Decision trees represent rules, which can be understood by humans and used in knowledge system such as database.

Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable

Attribute selection. Correct selection of attributes partition the data set into distinct classes. Our work uses J48 decision tree for classification. Observations show that Decision trees outperform the other two classifiers but take more time to build the model.

Classification via clustering:

Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Observations show that Classification via clustering performs poor even after reduction of number of attributes when compared to the other two methods.

The four terms used in computing evaluation measures are used for evaluating the model and are described here. The True positives (T_Pos) refer to the positive tuples that are correctly labeled by the classifier, while True negatives (T_Neg) are the negative tuples that are correctly labeled by the classifier. False positives (F_Pos) are the negative tuples that are incorrectly labeled by the classifier, False negatives (F_Neg) are the positive tuples that were incorrectly labeled by the classifier. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Sensitivity is referred to as the true positive rate that is the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = = \frac{\text{T\_pos}}{\text{pos}}$$

Specificity is the true negative rate that is the proportion of negative tuples that are correctly identified.

$$\text{Specificity} == \frac{\text{T\_neg}}{\text{neg}}$$

Classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. It is a function of specificity and sensitivity.

$$\text{Accuracy} = = \frac{\text{T}_{\text{pos}} + \text{T\_neg}}{\text{pos} + \text{neg}}$$

Distribution association rules:

A distributional association rule is defined by an itemset I and is an implication that for a continuous outcome y, its distribution between the affected and the unaffected subpopulations is statistically significantly different. For example, the rule {htn, fibra} indicates that the patients both presenting hypertension (high blood pressure) and taking statins (cholesterol drugs) have a significantly higher chance of progression to diabetes than the patients who are either not hypertensive or do not have statins prescribed. The distributional association rules are characterized by the following statistics. For rule R, let OR denote the observed number of diabetes incidents in the subpopulation DR covered by R. Let ER denote the expected number of diabetes incidents in the subpopulation covered by R.

ER = OR −i∈DRyi,.

where yi is the martingale residual for patient.

Parameter	Weightage	Values
Male &Female	Age<30 >30to<50 Age>50&Age<70 Age>70	0.1 0.3 0.7 0.8
Smoking	Never Past Current	0.1 0.3 0.6
Overweight	Yes No	0.8 0.1
Alcohol Intake	Never Past Current	0.1 0.3 0.6

Heart Rate	Low(<60bpm)
	Normal(60to100bpm)
	High(>100bpm)
		0.9
		0.1
		0.9

Blood Sugar	High(>120&<400)
	Normal(>90&<120)
	Low(<90)
		0.5
		0.1
		0.4

Bad Cholesterol	Very High>200 High(160to200) Normal<160	0.9 0.8 0.1

Table -1: Description of the risk factors that appeared in any of the summarized rules.

Rule set and database summarization

The goal of rule set summarization is to represent a set I of rules with a smaller set A of rules such that I can be recovered from A with minimal loss of information. Since a rule is defined by a single itemset, we will use „itemset" in place of „rule" meaning the „itemset that defines the rule".

Fuzzy Clustering Means

In fuzzy clustering data elements can belong to more than one cluster. The strength of the association between the data elements and a particular cluster. In fuzzy clustering every point has a degree of belonging to as in fuzzy logic rather than belonging completely to just one cluster.

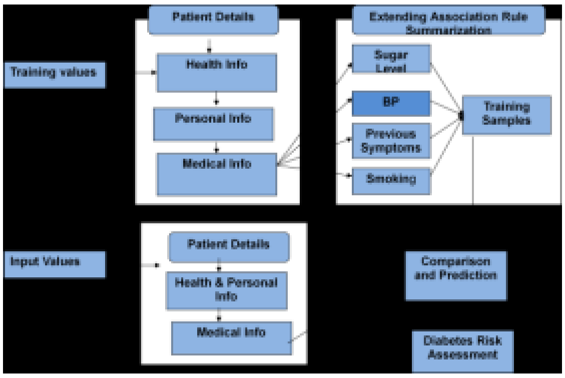
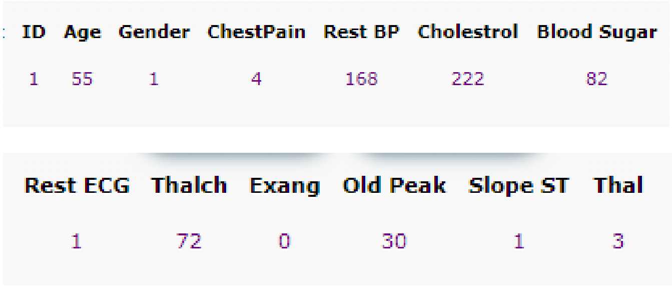


Fig -1 Overall description for risk assessment.

4. RESULT

The user is required to enter their data for the attributes shown below.

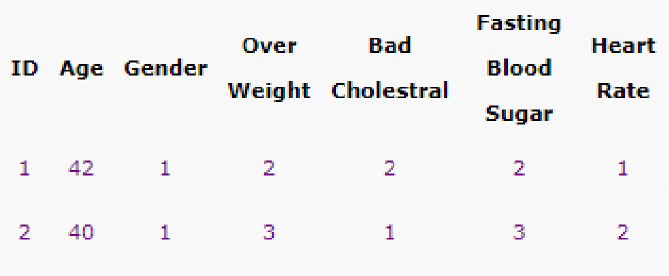
The values entered by the users are then converted into weights and clustered and the Naïve Bayes algorithm is applied to the dataset it is tested with the already classified data to improve accuracy and the result is displayed in a probabilistic manner indicating the risk of acquiring the disease.



ID	Age	Gender	ChestPain	Rest BP	Cholestrol	Blood Sugar
1	55	1	4	168	222	82

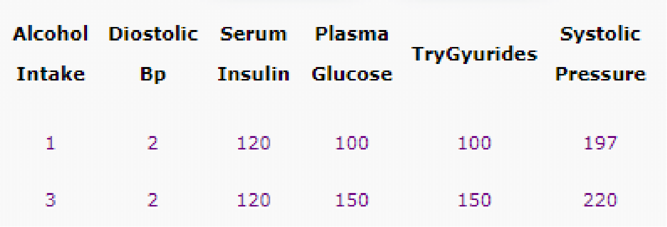
Rest ECG	Thalch	Exang	Old Peak	Slope ST	Thal
1	72	0	30	1	3

Fig -2 Patient details required for analysis of heart disease



ID	Age	Gender	Over Weight	Bad Cholestral	Fasting Blood Sugar	Heart Rate
1	42	1	2	2	2	1
2	40	1	3	1	3	2

Fig -3 Patient details required for analysis of Diabetes.



Alcohol Intake	Diostolic Bp	Serum Insulin	Plasma Glucose	TryGyurides	Systolic Pressure
1	2	120	100	100	197
3	2	120	150	150	220

Fig -4 Output window

5. CONCLUSION

Early detection is crucial to treat and reduce the complications associated with heart diseases and diabetes. Predicting disease with computer modeling and mathematical analysis are becoming more popular. Computers can be trained to predict specific diseases. Integrating the power of the human perception and the visual analytic method provides an appropriated framework for the knowledge discovery process. The main motivation of this paper is to provide an insight about detecting heart disease risk rate using data mining techniques. Various Data mining techniques and classifiers are discussed in many studies which are used for efficient diagnosis. As per the analysis mode, it is seen that many authors use various technologies and different number of attributes for their study. Hence, different technologies give different precision depending on a number of attributes considered. However, Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time.

REFERENCES

[1] Tjeresa Princy R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", IEEE, 2016.

[2] X Rexeena, Suganya Devi, S Saranya, "Risk Assessment for Diabetes Mellitus using Association Rule Mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3, Issue 2, February 2014.

[3] Jefri Junifer Pangaribuan, Shuarijto, "Diagnosis of Diabetes Mellitus Using Extreme Learning machines", IEEE, 2014.

[4] Messan Komi, Jun Li, Yongxin Zhai and Xianguo Zhang, 'Application of Data Mining Methods in Diabetes Prediction", IEEE, 2017

[5] V A Kanimozhi, Dr. T Karthikeyan, "A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease", International Journel of Advanced Research in Computer Engineering & Technology(IJARCET), Volume 5, Issue 4, April 2016.

[6] Sellappn Palaniappan, Rafiah Awang, "Inteligent Heart Disease Prediction using Data Mining Techniques", IEEE, 2008.