International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Prediction of Diabetes in Pregnant women using Machine Learning Algorithm

Sandra Treesa Tom, Sourav Sinha, Prof. Vaidhehi . V
Department of Computer Science, Christ, Deemed to be University, Bengaluru, Karnataka, India

ABSTRACT

Data Mining is a popular technology employed in the area of healthcare industry. There has been a steady up rise in use of technology in the field of health care. Predictive analytics is being seen as a more suitable approach of reducing the health issues. Hence, the study of each diseases can give new way to predict the chance of happening it again which help to prevent before it happens. Use of predictive analytics in health care would provide insight on learning patterns and factors that causes diseases. The diabetes data which consists of wide range of attributes is chosen to apply well-known classifying algorithms like J48, SMO and Naïve Bayes to find the accurate result. We compare the answers gathered from the three algorithm to fetch result that is more accurate. Data mining with Classification Algorithms plays an important role in the field of medical diagnosis to diagnose the disease. There requires a model built to analyse and extract information from the data. This paper discuss about the model, which gives an accurate result to the women about chances of getting gestational diabetes according to the data given.

Keywords: Data Mining, J48, SMO, Naïve Bayes, Classification Algorithms.

I. INTRODUCTION

The procedure of data mining deals with recognizing valuable information from vast amount of heterogeneous data and it helps to determine the patterns and rules from the data. Now days, the health organizations producing huge amount of data in the area of cancer and other health issues so it's very difficult to analyse the data. Medical data mining helps to extract hidden patterns, which thereby opens the door to an enormous source of analysis of medical data. Classification, Clustering, Regression etc., are some data mining techniques.

Diabetes is a very severe health issue when the level of blood glucose becomes unregulated. In actual, glucose acts as fuel to our body. When the body uses glucose as a fuel, insulin is essential to get the glucose into cells. Diabetics happen because of either production of insulin is insufficient or the cells don't react accurately to insulin or else both. Those who are having diabetes they suffer with polyuria (increase in thirst and hunger).

Data mining is becoming popular in health field, so, many data mining tools are used for analysing the data. For the most part information mining instruments are utilized to foresee the victories from the data collected. Developing efficient data mining tools helps to reduce the cost and time, thus increasing the efficiency.

This research is to predict the possibility of getting diabetics in pregnant and non-pregnant ladies based on the attributes like Age, BMI, Plasma concentration and the number of times being pregnant. It provides an option for the customer to choose the classifier like Naïve Bayes, J48 and SVM. When the user selects any one of the classifier it

gives the output as a message whether there is a chance of getting diabetes or not for the particular user.

II. LITERATURE REVIEW

Ashok Kumar et al [1] proposed Performance and Evaluation of Classification Data Mining Techniques in Diabetes. In this paper, they have used some selected classification algorithms like Support vector mechine, Regression, BayesNet, NaiveBayes and Decision Table for the classification of diabetic patient dataset. For accurate and proper results, classification techniques are widely used in the medical field. Data mining techniques are used in healthcare field for, Diagnosis and treatment. The result showed that Decision table had highest Accuracy than other Classification Algorithm.

Mukesh Kumari et al [2] proposed a prediction of diabetes using Bayesian Network. With this paper Bayesian Network classifier was utilized to gauge the people whether they are diabetic or not. The dataset has been accumulated from a medicinal center, which gathers the data of individuals with and without diabetes. They used Weka tool for the test investigation. Dataset contains every one of the subtle elements of a man like quick gtt(Glucose resistance test) value, casual Glucose tolerance test value,number of time pregnant,diastolic blood pressure (mmhg),triceps skin fold thickness(mm),serium insulin(μU/ml), body mass index (kg/m)diabetes pedigree function,age of individual.

V.Karthikeyani et al [3] has played out a near investigation of 10 Data Mining Classification Algorithm in Diabetes Disease Prediction .In this paper they have done a talk of C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA and PLS-LDA. They have utilized an tool called Tanagra which is a capable device that contains grouping, supervised learning, Meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The outcomes have demonstrated The CS-CRT algorithm best among tens. Chintan shah et al [4] used Data Mining Classification Algorithms for the prediction of Breast Cancer. In this research the researchers have used three different data mining classification methods for prediction of breast cancer which are decision tree algorithm, Bayes classification algorithm, K-Nearest Neighbour classification algorithm. They focus on accuracy and lowest computing time. When compared to other algorithms Naive Bayes is more accurate (95.9943) and faster. It takes as less as 0.02 seconds which is very less when compared to other types of Data Mining algorithms.

Divya Tomar et al [5] suggested the knowledge towards prerequisites for health domain also regarding suitableness decision for available system. There is no single information mining systems which provide for steady results for different types of data in healthcare. Those execution of data mining techniques varies concerning illustration for every the dataset that we need picked for those test. Classification, clustering and mining is used by the information mining systems in healthcare to increase their ability for decision making in patient health. On the basis of the seriousness of the disease a patient can be classified into “high risk” and “low risk”. For these purposes we can use the classification methods like KNN (K-nearest neighbour), DECISION TREE (DT), and SVM (support virtual machine), NN (Neural network). The result showed that that utilizing data mining learning doctor might effectively distinguish those viable cure, patients acquire expense compelling treatments, social insurance industry manages their client What's more social insurance insurers find at whatever cases of cheating for medicinal claim.

The existing framework need diverse sorts of data mining tools would accessible in the marketplace, every with their qualities also shortcomings. As per the current day, all the tools for Data mining, where we can do pre-processing, mining, and analysis data

with a various different data sets which supports various data mining functions such as clustering, visualization, pre-processing, feature selection and regression. There are many similar Tools, which have not yet, full-fledged like WEKA, Orange etc. All the data mining tools are grouped into three categories such as dashboards, text mining tools and traditional data mining tools.

The approach used by the existing system for the detection of diabetes from the data was complex and time consuming. Those system used combination of complex algorithm to find the result. The authors categorized the drawbacks of misclassification which reduce the performance of the algorithm. The public health agencies can use these sources of error to develop techniques to refine the algorithms and improve the efficiency of the data and also reduce the errors. Additionally using these algorithms in the public health systems helps to enhance and validated the work done.

III. PROBLEM DEFINITION

Data Mining is one of the most booming area of research that become increasingly popular in health organization, mostly in healthcare field may be a need for proficient explanatory procedure for identifying obscure also profitable data for health information. In health industry, Data Mining provides medical treatment for the patients at less expensive, detecting the sources of diseases and implementation of treatment methods. The information produced toward those wellbeing associations will be exceptionally limitless also mind boggling because of which it will be was troublesome with investigate the information in place will aggravate significant choice viewing patient's health.
The health organizations generate large amount of data which causes difficulty in analysing data. In such situation the concept of data mining comes into picture. Which is an analytical methodology can be used to classify the patients having same type of health issues. Therefore, there will be requirement to produce a capable device to analysing what's more extracting important information from the gathered data. Data mining with Classification Algorithms plays a vital role in the field of medical diagnosis to diagnose the disease.

IV. METHODOLOGY

Many Data Mining tools are available like WEKA, R tool etc. But specifically there is no tool available for diagnosing diabetes, so the proposed system proposes a prediction tool for diagnosing diabetes in pregnant and non-pregnant young obese ladies. The design is shown in figure 1.

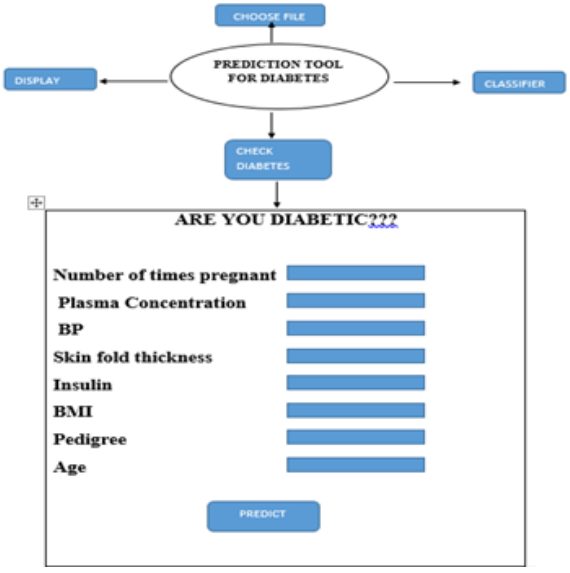

Figure 1. System design

Diabetes is often known as queen of all diseases as it is seen in people of all ages. Different doctors make use of different symptoms to predict the chances of a patient to be diabetic. However a standard set of parameters that directly contribute to diabetes have not yet been identified, so the proposed system focuses on identifying such a parameters set and selecting an efficient algorithm would greatly aide patients to identify their possibility of getting diabetes without the need for doctors or other expensive equipment.

A. Dataset Description

Sources:

Original owners: National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database:
Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231
Date received: 9 May 1990

Data Set Details

1. Name : Diabetes
2. Attributes : 8 plus class
3. Tuples : 768
4. Learning : Supervised
5. Missing Values : None
6. Normalized : No
7. Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

Past usage: Smith,~J.~W., Everhart,~J.~E., Dickson,~W.~C., Knowler,~W.~C., \& Johannes,~R.~S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In {\it Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261 265). IEEE Computer Society Press.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

For Each Attribute

1. Number of times pregnant: Throughout pregnancy, those placenta makes hormones that might prompt a build-up from claiming sugar in the blood. Usually, pancreas make sufficient insulin response on handle that. In not, your glucose levels will Ascent Furthermore might foundation gestational diabetes.

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test: Over fasting adults, blood plasma glucose ought further bolstering not surpass 7 mmol/l alternately 126 mg/dL. Maintained higher levels for glucose result in harm of the blood vessels what's more of the organs they supply, prompting those difficulties for diabetes.

3. Diastolic blood pressure (mm Hg): Hosting diabetes increments your danger of Creating high blood pressure Furthermore different cardiovascular problems, in view diabetes adversely influences those arteries, predisposing them on atherosclerosis.

4. Triceps skin fold thickness (mm): Measurement of skin fold thickness is mandatory to identify diabetic patients at risk early to prevent the development of cardiovascular disease and protect them against added complications

5. 2-Hour serum insulin (mu U/ml): A lady with insulin response resistance, typical glycaemic levels might a chance to be recognized after a glucose test in view the pancreas will must discharge overabundance insulin response so as with stay with the glucose in the ordinary extend. Therefore, a single "2-hour post-glucose insulin response level" gives the idea on be a dependable pointer from claiming insulin response imperviousness to PCOS patients.

6. Body mass index (weight in kg/(height in m)^2): Overweight particularly obesity, especially In more youthful ages, significantly expands lifetime danger for diagnosed diabetes, same time their effect once diabetes risk, an aggregation expectancy, Also diabetes span diminishes with ageists.

7. Diabetes pedigree function: Gives a few information for diabetes mellitus history for relatives and the hereditary relationship for the individual's relatives of the patient.

8. Age (years): One important measure for ensuring health in later years is controlling blood glucose levels, since high blood glucose tends to accelerate the effects of aging and increases the risk of developing diabetes complications.

9. Class variable (0 or 1): the result of the test "positive" or "negative".

B. ALGORITHM USED

Classification is a process of assigning entity to an already defined class by analysing the features. For example on the basis of the level of the disease a patient can be classified into "high risk" and "low risk" that is the task of assigning instances to pre-define classes. Three classification algorithms are used to find the accurate result.

- ✓ The Sequential Minimal Optimization (SMO) Algorithm: SMO solves the SVM QP problem by decomposing it into QP sub-problems and solving the smallest possible optimization problem, involving two Lagrange multipliers, at each step.
- ✓ Naive Bayes algorithm: The Naive Bayes classifier works on a simple, but comparatively intuitive concept. In addition, in some cases it is also seen that Naive Bayes outperforms many other comparatively complex algorithms.
- ✓ J48 Decision Tree: A decision tree is a predictive machine-learning model that decides the target value of a new sample based on various attribute values of the available data.

C. PREDICTION METHODOLOGY

Figure 2. prediction methodology

For predicting the result, we use three different algorithm. For selecting appropriate algorithm, the analytic tool Weka is used. The entire classification algorithm was implemented on the diabetes dataset. Among those SMO, J48 and Naïve Bayes gave the more accurate result. The Weka API is included in NetBeans and a model is build. The new data which is entered through the proposed system is passed to the model and the results are obtained. Then all the results are compared and the accurate result is given to the user.

D. UI INTERFACE

User Interface used in this research work is shown in figure3.

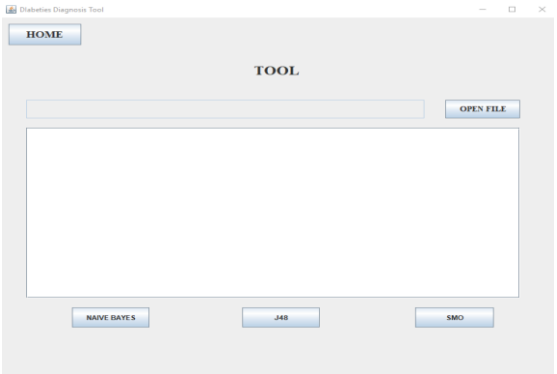

Figure 3. UI Design

E. RESULTS AND DISCUSSIONS

In healthcare industry, there is growing necessity of a well-organized analytical technique to capture essential and unidentified information from health data. The technique follows a process capturing and gathering the data, learning about the data and examining and evaluating the data to discover the origin which aids to cure diseases. Due to the vast data produced by the healthcare industry, computation and analysis of the data becomes a very difficult task.

Data Mining comes to aid in such crucial situations. In healthcare industry, Data Mining delivers health by diagnosing the reason of diseases and application of medical treatment at a very nominal rate. It is a tool for analysis which is used to organize the patients having similar kind of medical complexities. So, an urge to create a prevailing technique to examine and mine vital information from the vast and complex data of the healthcare industry. The model which has been developed in the research, plays a very vital part. The model requires the data from the user. The taken data is passed to the SMO, J48 and Naïve Bayes model, which was build prior to help in classifying the new data. Once the classification is completed, each model gives its own class value, corresponding to the data provided by the user to whichever class value it fits best. From the gathered class value, the result is obtained. Result is declared by taking into consideration the fact that which class occurs more number of times out of the class value provided by each of the models. The possible class values are positive and negative. This means you are diabetic or not. By testing with the collected data, the accuracy of the built model was quite high. The model also predicts the chance of the user to get diabetes in future, which help the user to take precautionary measures so that he/she does not get diabetes in future.

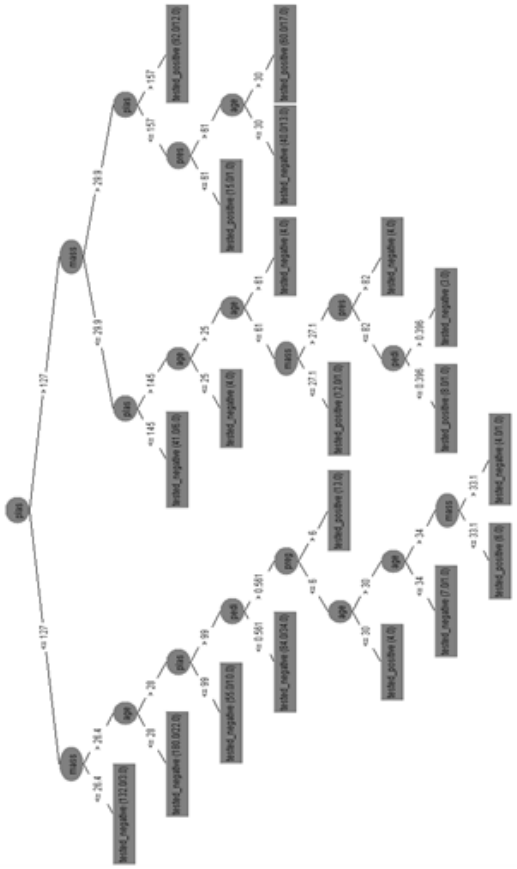

Figure 4. 5 J48 Flowchart

V. CONCLUSION

Diabetes mellitus is known as queen of all diseases because it leads to acute complications including Diabetic Ketoacidosis and Nonketotic hyperosmolar coma, Stroke, Cancer, Blindness, Kidney failure and so on. Diabetes mellitus is caused due to high blood sugar level. Diabetes is one of the most common diseases in the present generation that is why this topic is chosen for this project. The project helps to analyze and predict the possibility of getting Diabetes in pregnant and non-pregnant women based on the relevant attributes such as Age, BMI and Plasma concentration with the help of rules, which is extracted by Decision Tree. Three classification algorithms are taken into considerations that are J48, SMO and Naïve Bayes methods.

Only three classifications are used in the system to predict the result. There are other test options like use training set, percentage split other than cross validation. The system uses only cross validation. The

test option can be changed while using the system, it is fixed. In cross validation itself the no. of folds are fixed. The system automatically taken as ten. That also can't be changed.

A database can be added to the system so that the user details can be stored whenever they use the system. In future, the stored information can be used to generate reports. The limitation of not having more than one test option should be changed. The no. of folds for the cross validation should be changed from fixed to variable.

VI. REFERENCES

[1]. Dr. D. Ashok Kumar and R. Govindasamy," Performance and Evaluation of Classification Data Mining Techniques in Diabetes", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6, 1312-1319, 2015.

[2]. Mukesh kumari, Dr. Rajan Vohra ,Anshul arora," Prediction of Diabetes Using Bayesian Network", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178.

[3]. V.Karthikeyan, I.Parvin Begum,K.Tajudin,I.shahina Begam,"comparative of data minning classification algorithm(CDMCA)in diabetes disease prediction",International journal of computer applications(0975-8887) Volume 60-No.12,December 2012.

[4]. Mr. Chintan Shah, Dr. Anjali G. Jivani," Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE – 31661

[5]. Divya Tomar and Sonali Agarwal ,"A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5, pp. 241-266,2013.