International Conference of Sciences, Engineering & Environmental Technology (ICONSEET), 6(8): 57 – 62, 2021
ISSN 0794-9650

www.repcomseet.com

Automation of Diabetics Prediction System using Data Mining Technique

Ateko Busayo C. , Onyeka Ndidi C., Olatunji Abiodun F., Babafemi Olusola F. and AbdulGaniyu Toyeeb K.

Department of Computer Science, Federal Polytechnic Ede, Osun State, Nigeria.
Author's Email: ayclara16@gmail.com

Abstract- Diabetes is one of the most common diseases in the world. Complications of this disease include nephropathy, cardiac arrest, blindness, and even mutilation of the body. The accurate diagnosis of this condition is very important. This study was to identify and provide a model for prediction of Diabetics using data mining approach. In this study, diabetes is predicted using significant attributes such as Pregnancies, glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. Various tools are used to determine significant attribute selection, prediction and association rule mining for diabetes. Our finding shows a strong association between diabetes and the body mass index (BMI) of an individual, also their glucose level, which was extracted via the Naïve Bayes Classification. Artificial neural network (ANN) was implemented for the prediction of diabetes. The ANN technique provided the best accuracy for the prediction and may be useful to assist medical professionals with the treatment decisions.

Keywords: Diabetes, Data Mining, Artificial Neural Network, BMI

1.0 INTRODUCTION

The term "diabetes" is a disease that happens once the blood sugar within the body, conjointly referred to as blood glucose, is simply too high. Blood sugar is the main source of energy and comes from the food we have tendency to eat. In step with doctors, diabetes occurs when a gland known as pancreas does not release a hormone called insulin in sufficient quantity. Insulin is a hormone that carries sugar from the bloodstream to various cells to be used as energy. Lack of insulin disrupts the body's natural ability to make use of insulin produce accurately. As a result of this, high levels of glucose are discharged in urine. In the long -term, diabetes when not properly managed can lead to organ failure, cardiovascular diseases and disrupts alternative functions of the body. WHO (World Health Organization) has listed diabetes as one of the four major NCDs (Non Communicable Diseases) within the world nowadays (World Health Day, 2016).

Diabetes is divided into two distinct types; type 1 diabetes enforces the necessity for artificially infusing insulin through medicines or by injections and type 2 diabetes, pancreas create insulin, however it is not effectively employed by the body. The majority of individual with diabetes are affected by type 2 diabetes. Diabetes was a common problem among adult's specifically old folks however due to dynamic lifestyles diabetes affects children too. Type 1 diabetes is unpreventable because of the various external environmental stimulants which result in the destruction of body's insulin producing cells. However, dynamic lifestyle to achieve the desired body weight and procure the physical activities can help to prevent type 2 diabetes to enlarge. Diabetes impacts men and women proportionately; there are over 12 million men with diabetes and 11.5 women with diabetes. Therefore, predicting diabetes manually generally looks not to be objective and it consumes a lot of time and cost. Diabetes treatment focuses on controlling blood sugar levels to stop varied symptoms and complications through diet and exercise.

Data mining is a new concept used for retrieving information from a large set of data. Mining means using available data and processing it in such a way that can be used for decision-making. Data mining thus has evolved based on human needs which can help humans in recognizing relationship patterns and forecasts based on pre-set rules and stipulations built into the program (Alexandra Twin, 2020). Data mining helps in pattern identification and categorizing data records by conducting cluster analysis, identification of odd records conjointly known as detecting anomalies and association rule mining or dependencies.

As we can realize from these facts, problems related to diabetes are many and quite costly. It is a very serious disease because, if not treated properly and on time, it could lead to very serious complications, may be the death of the patient. This makes diabetes one of the main priorities in medical science research.

As per the existing system, patients have to visit a diagnostic center, consult their doctor and wait for a day or more to get their result. Moreover, each time they want to get their diagnosis report, they need to waste their money in vain. But with the rise of Machine Learning approaches, the need for dietetics prediction system which will find solution to the problem using data mining will be develop.

2.0 RELATED CONCEPTS

Diabetes a non-communicable disease is resulting to long-term complications and serious health issues. A report from the World Health Organization (WHO) addresses diabetes and its complications that impact on individual physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred because to the risk factors of diabetes such as cardiovascular and other diseases. Diabetes prediction can be achieved using several classifiers which are discussed below:

2.1 NEURAL NETWORKS

Inspired by the means biological nervous systems process information, Artificial Neural Networks (ANNs) are composed of interconnected parts named neurons, processing and cooperating to determine solutions to specific issues. Similar to humans, the learning process of ANNs is based on examples. Instead of a set of instructions for the accomplishment of a specific task, they are given examples to analyse and find a way to solve the problem (Maind, S.B. & Wankar, 2014).

2.2 KNN

It is a classification technique that classifies the new sample supported similarity measure or distance measure. The measure includes three distance measures Euclide an distance, Manhattan, Minkowski. The steps for KNN is given below.

1. Training phase of the algorithm consists of solely storing the feature sample and class label of training sample.
2. Classification phase: the user must define a “k” value for the classification of the undefined sample for the k number of the class labels, therefore the unlabelled sample can be classified into the defined class based on the feature similarity.
3. Majority of choice classification happens for unlabelled class. The value of the k can be selected by various techniques like heuristic technique.

2.3 RANDOM FOREST

It is supervised learning, used for classification and Regression. The logic behind the random forest is bagging technique to create random sample features. The distinct between the decision tree and the random forest is the process of finding the root node and splitting the feature node will run randomly.

2.4 SVM

Originated from statistical learning theory, SVM is a supervised machine learning model that uses classification algorithms for both classification or regression challenges (Sunil Ray, 2017).

2.5 DECISION TREE

It is a supervised learning methodology that is used for solving classification problems. Decision tree is a technique which iteratively breaks the given dataset into two or more sample data. The goal of the method is to predict the category value of the target variable.

2.6 NAIVE BAYESIAN

The Naive Bayes relies on Bayes theorem with the independence assumption between the predictors. Naïve Bayesian methodology takes the dataset as input, performs analysis and predicts the category label using Bayes’ Theorem. It calculates a probability of class in input data and helps to predict the category of the unknown data sample. It is a powerful classification technique suitable for large datasets.

2.7 REVIEW OF RELATED WORK

Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study BY Luís Chaves and Gonçalo Marques at Polytechnic of Coimbra, ESTGOH, Rua General Santos Costa, 3400-124 Oliveira do Hospital, Portugal. The authors emphasize on diabetics condition that is well-known in the 21st century. In late 2019, a new public health concern was rising (COVID-19), with a selected hazard regarding individuals living with diabetes. Medical institutes have been grouping data for years. The authors designed the system to achieve predictions for pathological complications, which will prevent the loss of lives and improve the quality of life using data mining processes. The authors compare the classifier that can be used for diabetes system. The results

suggest that Neural Networks ought to be used for diabetes prediction. The proposed model presents associate group of AUC of 98.3% and 98.1% accuracy, an F1-Score, Precision and Sensitivity of 98.4% and a Specificity of 97.5%.

- A Prediction Technique in Data Mining for Diabetes Mellitus by Harleen et al. proposed a system which supported a technique in data mining for diabetes disease prediction. The proposed system has three main steps which are: preprocessing, feature extraction and parameter evaluation. In preprocessing step, the empty and anomalies sets are off from the used dataset. Besides that, the helpful hidden patterns and relationships of the dataset are explored in the feature extraction step in order to improve the decision making result. Naive Bayes and the achieved rates are 73.8%, 76.3% respectively.

3.0 RESULT

The framework use medical parameters of diabetes to classify diabetes in a patient. The steps involved are general data collection, data pre-processing, classification and prediction.

- Data Collection, public dataset of patient having diabetes was used.
- Data pre-processing was done to get rid of noise and null fields.
- Classification was done using Naïve Bayes classifier to classify the dataset into categories.

The following images represent the various screen design displayed to the users when interacting with the system:

3.1 PREDICTION PAGE

This is the page that accepts inputs data from the User after some physical test and processes it to give output. The screenshot of the prediction page is shown in figure below

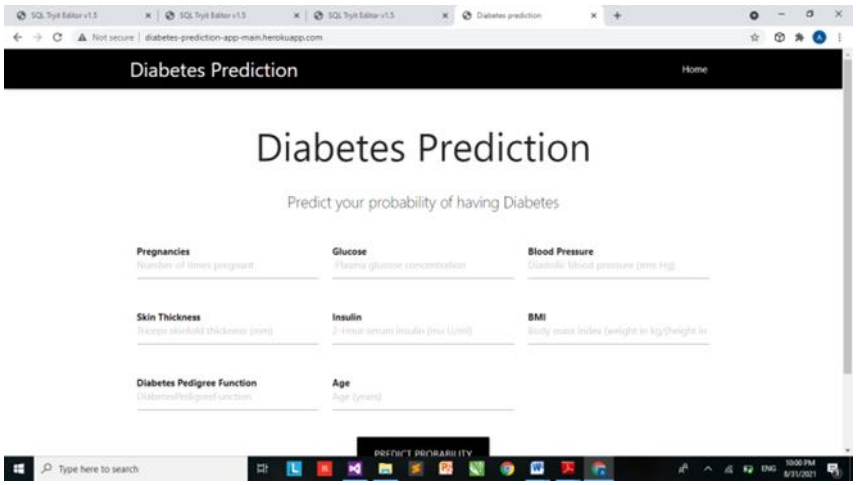

Fig. 1: Prediction Page

3.2 RESULT PAGE

This is the page that displays the probability of patience having diabetes after the processing of the inputs data. The screenshot of the prediction page is shown in figure below:

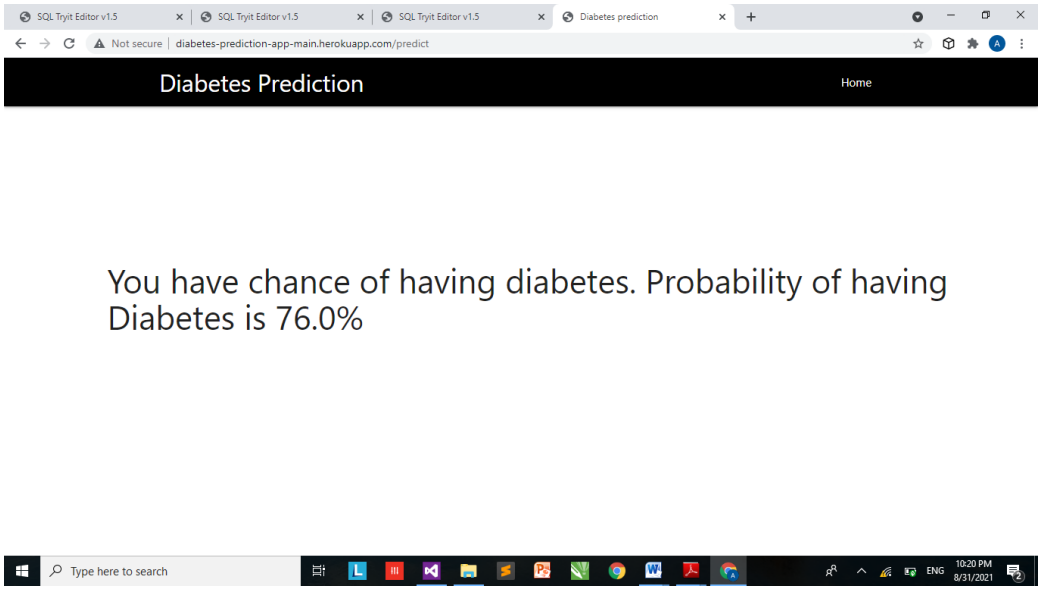

Fig. 2: Result Page

3.3 DESCRIPTION OF DATA SET

Diabetes is predicated using the following significant attributes:

Table 1: Description of Dataset

S/N	ATTRIBUTE	DESCRIPTION
1	Pregnancies	Pregnancy count of women
2	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	Skin Thickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin (mu U/ml)
6	BMI(Body Mass Index)	Body mass index (weight in kg/(height in m)2)
7	DPF	Diabetes pedigree function
8	Age	Age of a person

3.4 TEST DATA AND RESULT

Table 2: Test Data and Result

Pregnancies	Glucose	BloodPressure	Skin Thickness	Insulin	BMI	Diabetes PedigreeFunction	Age	Outcome (%)
6	**148**	**72**	**35**	**0**	**33.6**	**0.627**	**50**	**71**
1	**85**	**66**	**29**	**0**	**26.6**	**0.351**	**31**	**8**
8	**183**	**64**	**0**	**0**	**23.3**	**0.672**	**32**	**76**
1	**89**	**66**	**23**	**94**	**28.1**	**0.167**	**21**	**6**
0	**137**	**40**	**35**	**168**	**43.1**	**2.288**	**33**	**88**
5	**116**	**74**	**0**	**0**	**25.6**	**0.201**	**30**	**17**
3	**78**	**50**	**32**	**88**	**31**	**0.248**	**26**	**9**
10	**115**	**0**	**0**	**0**	**35.3**	**0.134**	**29**	**62**
2	**197**	**70**	**45**	**543**	**30.5**	**0.158**	**53**	**70**

4.0 CONCLUSIONS

Data Mining has the ability to support clinical decision support systems. A massive amount of data is being collected from public dataset. These data can be used to support healthcare facilities and public health. Detecting diseases might dramatically influence how a person will live to the rest of his days. In this study, Naive Bayes classification methodology was used to a publicly accessible diabetes data set. The method achieved above 86% efficiency compare to other data mining classification methodology. A public data set containing information of 550 patients between 20 and 90 years old was used. The results have shown that Neural Network is an effective method to predict diabetes disease.

5.0 RECOMMENDATIONS

With the rate at which diabetes is increasing among the people, there is a need of prediction of diabetes. The present study shows how importance machine learning is in the healthcare industry for decision-making and also in reducing the cost of diagnosis. The main contribution of the work is used a model that best for predicting of diabetes and has also emphasized the importance of feature selection and proper handling of data. It would be interesting in the future work to know whether body size and height could be used in the dataset and find the role these parameters play in the prediction of diabetes. Also future work should be done on improving the accuracy of the prediction by increasing the numbers of training data.

6.0 DATA AVAILABILITY

Previously reported data were used to support this study and are available at https://www.kaggle.com/c/diabetes-prediction/data

REFERENCES

Alexandra Twin (2020): Data Mining published by Dotdash publishing family

Chaurasia V. and Saurabh. P (2017): Early Prediction of Heart Diseases Using Data Mining Techniques: Caribbean Journal of Science and Technology, revised version. 10pages.

Dangare, Chaitrali and Apte, Sulabha (2012): A Data Mining Approach for Prediction of Heart Disease Using Neural Networks. International Journal of Computer Engineering and Technology (IJCET), Volume 3, Issue 3.

Maind, S.B. & Wankar (2014): Research Paper on Basic of Artificial Neural Network.

International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 1 96 – 100

Mary K. Obenshain (2015): Application of Data Mining Techniques to Healthcare Data

Published online by Cambridge University Press.

Sunil Ray(2017: Understanding Support Vector Machine(SVM) algorithm from examples (along with code)