Inernship Report

A Internship Report submitted in partial fulfillment of the requirements for the degree of

Bachelors of Technology (B.Tech)
In
Computer Science and Engineering

Submitted by

Aasurjya Bikash Handique,
(CSB18017)

Under the supervision of
Y Vishnuvardhan
Chief Director
Exposys Data Labs

SCHOOL OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
TEZPUR UNIVERSITY
NAPAAM - 784028, ASSAM, INDIA
December 2021

Acknowledgment

Firstly, I would like to express my deep and sincere gratitude to my Project guide **Y Vishnuvardhan**, Chief Director, Exposys Data Labs for giving us the opportunity to work under him and providing us encouragement and valuable guidance in bringing shape to this project work. It was a great privilege and honor to work and study under his guidance. I would also like to thank him for his friendship and empathy. Secondly, I am very much thankful to **Dr. Bhogeswar Borah**, Head of Department of Computer Science & Engineering for giving us the opportunity to present our work and ideas.

I am thankful to all the Professors and Faculty Members in the department for their teachings and academic support and also to our friends without whom this project could not have been successful.

Contents

Abstract

Diabetes could be a chronic disease with the potential to cause a worldwide health care crisis. As per International Diabetes Federation 382 million people live with diabetes across the full world. By 2035, this can be doubled to 592 million. Diabetes mellitus or just diabetes may be a disease caused thanks to an increased level of glucose.

Various traditional methods, supported physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is sort of a challenging task for medical practitioners because of the complex interdependence of assorted factors as diabetes affects human organs like kidneys, eyes, heart, nerves, feet, etc.

Data science methods have the potential to learn other scientific fields by shedding new light on common questions. One such task is to assist make predictions on medical data. Machine learning is an emerging scientific field in data science managing the ways during which machines learn from experience. The aim of this project is to develop a system which will perform early prediction of diabetes for a patient with higher accuracy by combining the results of various machine learning techniques.

This project aims to predict diabetes via three different supervised machine learning methods including SVM, Logistic regression, ANN. This project also aims to propose an efficient technique for the sooner detection of diabetes disease.

1 Introduction

1.1 Diabetes Mellitus

Diabetes is one amongst deadliest diseases within the world. It's not only a disease but also a creator of different varieties of diseases like coronary failure, blindness, kidney diseases, etc. The normal identifying process is that patients must visit a diagnostic centre, consult their doctor, and remain for each day or more to urge their reports. Moreover, every time they need to urge their diagnosis report, they have to waste their money vainly. Diabetes Mellitus (DM) is defined as a gaggle of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency leads to elevated glucose levels (hyperglycaemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one in every of the foremost common endocrine disorders, affecting quite 200 million people worldwide. The onset of diabetes is estimated to rise dramatically within the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), in keeping with the etiopathology of the disorder. T2D appears to be the foremost common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is believed to flow from to autoimmunological destruction of the Langerhans islets hosting pancreatic- cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other sorts of DM, classified on the premise of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood sugar levels (fasting plasma glucose = 7.0 mmol/L).

1.2 Machine Learning

Machine learning is the scientific field dealing with the ways during which machines learn from experience. For several scientists, the term "machine learning" is a dead ringer for the term "artificial intelligence", providing the chance of learning is the main characteristic of an entity called intelligently in the broadest sense of the word. The aim of machine learning is that the construction of computer systems that may adapt and learn from their experience. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is alleged to be told from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. With the increase of Machine Learning approaches, we've got the flexibility to seek out an answer to this issue, we've got developed a system using data mining which has the power to predict whether the patient has diabetes or not. Furthermore, predicting the disease early results in treating the patients before it becomes critical. data processing has the power to extract hidden knowledge from a large amount of diabetes-related data. thanks to that, it has a significant role in diabetes research, now over ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a better accuracy. This research has focused on developing a system supported three classification methods namely, Support Vector Machine, Logistic regression and Artificial Neural Network algorithms

2 Existing Method

The normal identifying process is that patients must visit a diagnostic centre, consult their doctor, and remain for each day or more to urge their reports. Moreover, every time they need to urge their diagnosis report, they have to waste their money vainly. Diabetes Mellitus (DM) is defined as a gaggle of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency leads to elevated glucose levels (hyperglycaemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one in every of the foremost common endocrine disorders, affecting quite 200 million people worldwide. The onset of diabetes is estimated to rise dramatically within the upcoming years.

3 Proposed System

Classification is one among the foremost important decision-making techniques in many world problems. During this work, the most objective is to classify the information as diabetic or non-diabetic and improve the classification accuracy. For many classifications problem, the upper number of samples chosen but it doesn't result in higher classification accuracy. In many cases, the performance of algorithm is high within the context of speed but the accuracy of knowledge classification is low. The main objective of our model is to attain high accuracy. Classification accuracy may be increase if we use much of the information set for training and few data sets for testing. This survey has analysed various classification techniques for classification of diabetic and non-diabetic data. Thus, it's observed that techniques like Support Vector Machine, Logistic Regression, and Artificial Neural Network are most suitable for implementing the Diabetes prediction system.

3.1 Artificial Neural Network

The artificial neural network is far similar as natural neural network of a brain. Artificial Neural networks (ANN) typically contains multiple layers or a cube design, and also the signal path traverses from front to back. Back propagation is that the use of forward stimulation to reset weights on the "front" neural units and this can be sometimes tired combination with training where the proper result's known. More modern networks are a small amount freer flowing in terms of stimulation and inhibition with connections interacting during a way more chaotic and complicated fashion. Dynamic neural networks are the foremost advanced, therein they dynamically can, supported rules, for brand spanking new connections and even new neural units while disabling others. Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer. The input neurons define all the input attribute values for the info mining model. In our work, the quantity of neurons is 7, since each item in our data set has 7 attributes, including: Glucose, force per unit area, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and age. For the hidden layer, hidden neurons receive inputs from input neurons and supply outputs to output neurons. The hidden layer is where the assorted probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a specific input to the hidden neuron. Mathematically, a neuron's network function f(x) is defined as composition of other functions gi (x), which may further be defied as a composition of other functions. The important characteristic of the activation function is that it provides a smooth transition as input values change, sort of a chickenfeed in input produces a little change in output. the synthetic neural networks are applied to tend to fall within the broad categories. Application areas include the system identification and control (vehicle control, trajectory prediction, process control, natural resources management), quantum chemistry, game playing and deciding (backgammon, chess, poker), pattern recognition (radar systems, face identification, seeing and more), sequence recognition (gesture, speech, handwritten text recognition), diagnosis, financial applications (e.g. automated trading systems), data processing (or knowledge discovery in databases, "KDD"), visualization and e-mail spam filtering. Artificial neural networks have also been accustomed diagnose several cancers. An ANN based hybrid carcinoma detection system named HLND improves the accuracy of diagnosis and also the speed of carcinoma radiology. These networks have also been used to diagnose glandular carcinoma. The diagnoses are often accustomed make specific models taken from an oversized group of patients compared to information of 1 given patient. The models don't rely upon assumptions about correlations of various variables. Colorectal cancer has also been predicted using the neural networks. Neural networks could predict the result for a patient with colorectal cancer with more accuracy than the present clinical methods. After training, the networks could predict multiple patient outcomes from unrelated institutions.

3.2 Support Vector Machine

The Support Vector Machine (SVM) was first proposed by Vapnik, and SVM is a set of related supervised learning method always used in medical diagnosis for classification and regression. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, so called structural risk minimization principle. SVMs can efficiently perform nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 7.

3.3 Logistic Regression

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labour force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. In this paper, Logistic regression was used to predict whether a patient suffered from diabetes, based on seven observed characteristics of the patient.

4 Methodology

In this modern era, human beings encounter different health issues. Most of the health issues are due to the food habits of the individuals. In this project work, a predictive approach is proposed to pre-treat Diabetic Mellitus. The proposed approach has three phases namely data collection, data storage and analytics. This approach plays an important role in predicting diabetes and pre-treating diabetic patients. The phases in the proposed approach for diabetic prediction are presented in Fig. 1. In the first phase, data collection is done through IoT devices and other sources. The collected data are cleansed using pre-processing techniques. Phase two deals with data storage. The pre-processed data are stored in warehouses. To store massive amount of data, cloud storage is used. The data stored in the cloud are analyzed to establish association between the various parameters such as BP, BMI, Air Pollution level etc., with Diabetic Mellitus. The third phase of the proposed approach deals with Predictive Analytics where the decisions are taken based on association rules with respect to diet pattern, physical fitness, current medicine intake etc.

5 Live Demo

Click Here for Demo Video. Click Here for Jupyter Notebook.

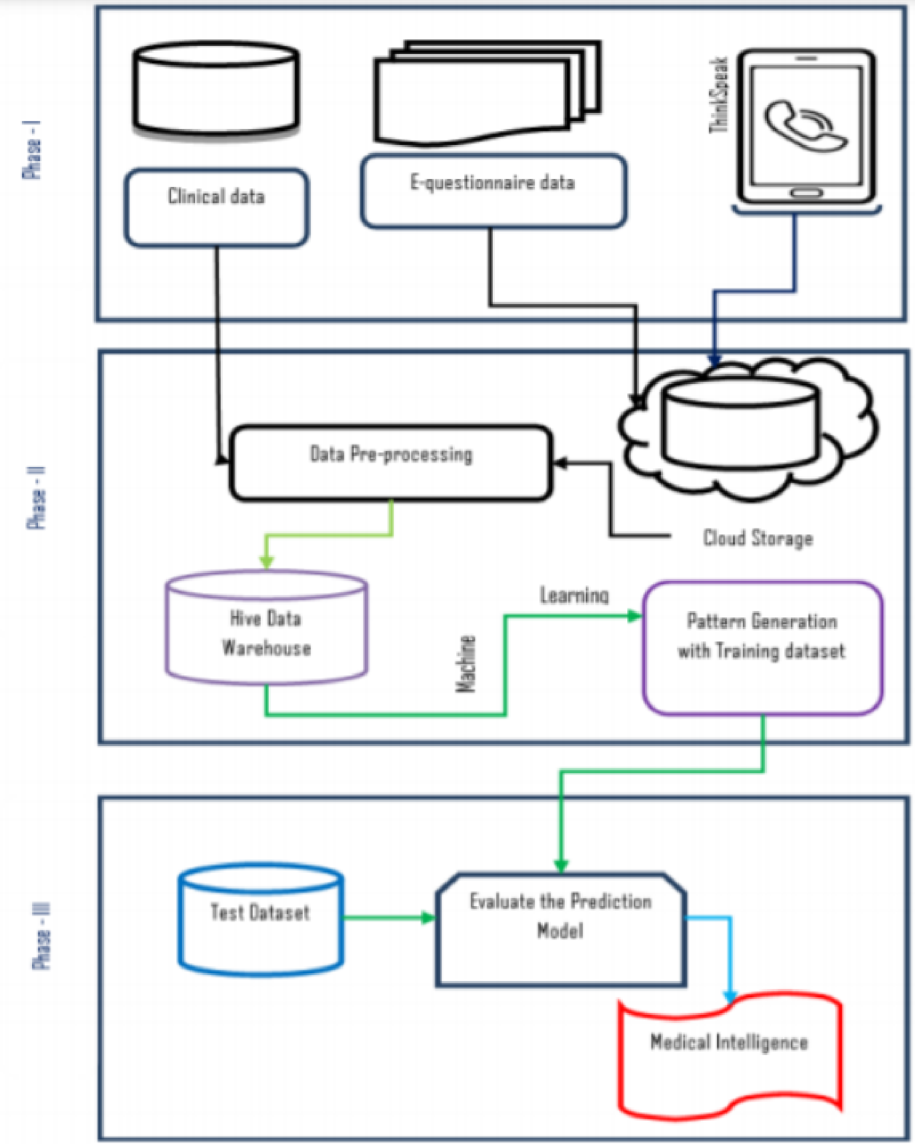

Figure 1: Methodology Diagram for Predicting Diabetes Mellitus

6 Conclusion

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of a large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decisions about the disease status.