Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5443 - 5447
Received 15 May 2021; Accepted 20 May 2021.

Early-Stage Detection of Diabetes Using Exploratory Machine Learning Algorithms

Atishay Jain[1], Yashovardhan Malhotra[2], M. Karthikeyan[3]
[1]Student, Department of Computer Science and Engineering SRMIST, kattankulathur, Chennai, India
[2]Student, Department of Computer Science and Engineering SRMIST, kattankulathur, Chennai, India
[3]Assistant professor, Department of Computer Science and Engineering SRMIST, kattankulathur, Chennai, India
aj8829@srmist.edu.in
ys9540@srmist.edu.in
karthikm1@srmist.edu.in

ABSTRACT

Diabetes is an ongoing illness and can be severe enough to cause a standard medical emergency. According to the International Diabetes Federation more than 300 million people are affected by this disease across the globe . By the late 2030s, this disease will affect more than 600 million people worldwide . Diabetes is caused due to long term high blood sugar levels or blood glucose . According to various tests, there are several common techniques available for detecting diabetes. Be that as it may, the early detection of diabetes is a very tedious task due to its complex dependence on various variables as it can affect organs such as the eye,kidney,heart and many others. AI is an upcoming field and it deals with the way a machine reaches a conclusion or a result based on the information fed to it. The aim of this project is to develop a framework that can more accurately identify the early chances of diabetes in a patient by linking the results of various AI methods. This task plans to foresee diabetes by means of three distinctive directed AI techniques including: SVM, Logistic relapse, ANN. This venture ––likewise expects to propose a powerful procedure for prior identification of the diabetes sickness.

Keywords:
SVM, ANN, k-NN, DT.

Introduction

Diabetes Mellitus

Diabetes is considered to be a very deadly sickness and affects the majority of the people. It isn't just an illness yet additionally a maker of various types of sicknesses like cardiovascular failure, blindness, kidney infections, etc. The typical hallmark is that patients have to go to an analysis center, consult their doctor and wait at least a day for their reports. They have to get their analytical report, and have to spend a lot of money. Diabetes is characterized on the basis of some metabolic troubles, especially abnormal emission of insulin and additionally through the activity. A person whose body cannot produce enough insulin has elevated blood sugar levels (hyperglycemia) and impaired digestion of starches, fat and proteins. It is considered as the most common endocrine disease that affects more than two hundred million humans worldwide. The occurrence of diabetes is expected to rise exponentially in the years to come .DM has primarily two types , Type 1 Diabetes and Type 2 Diabetes, depending on the severity of the problem. Type 2 Diabetes (T2D) is the most widely recognized type of diabetes (90% of every single diabetic patient), fundamentally described by insulin obstruction. The fundamental driver of T2D incorporates way of life, active work, dietary propensities and heredity, while Type 1 diabetes is thought to cause autoimmune obliteration of the islets of Langerhans, thereby relieving pancreatic β-cells. Only about 10% of the people worldwide are affected by T1D.

Machine Learning and Artificial Intelligence

Machine Learning or ML is the way to make a machine learn from the given data and results known as the training set and then make accurate predictions. For some researchers, the term "AI" is indistinguishable from the term "man-made intellectual capacity" because the ability to learn is the fundamental characteristic of a

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5443 - 5447
Received 15 May 2021; Accepted 20 May 2021.

substance. The main task for AI is to develop PC frameworks that can be customized and benefit from our expertise. To explain the principle of AI we can give a formal definition as :A pc program is actually supposed to gain E in relation to a series of tasks T and a quantified achievement P if its representation of tasks in T, as estimated by P, improves with E. For this topic we have created a framework that takes in different parameters from a person such as BMI, blood pressure etc and predicts whether a person has diabetes. In addition, anticipating the disease early indicates treatment for patients before it becomes basic. Information mining can separate hidden information from huge amounts of information related to diabetes. So you have a critical job in diabetes research like never before. The aim of this review is to create a framework that can be used to more accurately predict a patient's diabetes risk level.. This exploration has zeroed in on building up a framework dependent on three arrangement techniques such as, Support Vector Machine, Logistic relapse and Artificial Neural Network calculations.

Supervised Learning

Supervised Learning is a machine learning task where the input data along with the correct output data,known as the training set, is provided to a machine learning model. A supervised learning algorithm analyzes the training set and develops a function based on the results and uses this function to map out new examples i.e, the algorithm learns from the inputs and the correct outputs and hence is able to give accurate results. Supervised Learning is widely used in Bioinformatics, Database marketing, Pattern Recognition, Speech Recognition etc.

Unsupervised Learning

Unsupervised Learning is a machine learning technique where the model itself works and learns to detect patterns and information without any user supervision ,i.e. this model does not need a training set to learn from. It generally works on unlabelled data and is helpful in finding insights from the data. Unsupervised learning is considered to be closer to AI than supervised learning as it can be compared to a Human learning new things on its own. This type of model is widely used in Clustering and Association.

Literature Review

Authors N. Sneha and Tarun Gangil in the paper titled "Analysis of Diabetes Mellitus for Early Prediction by Selection of Optimal Characteristics" developed an algorithm using Naive Bayes, Decision Tree and Random Forest that best fits the data with respect to diabetic and non-diabetic patients. The percentage prevalence of the disease measured at its highest using the SVM is 45.7% which is very low.

In the 2018 article titled "Predicting Diabetes Mellitus Using Machine Learning Techniques" published by Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang, the concept of Principal Component Analysis (PCA) and the maximum relevance of the minimum explains redundancy. The main advantage of the article is that Principal Component Analysis (PCA) and Minimal Relevance of Minimal Redundancy (mRMR) decrease the dimensionality. The results showed that the prediction with random forest could achieve the highest precision

Another paper published by Meherwar Fatima, Maruf Pasha titled ‘Survey of Machine Learning Algorithms for Disease Diagnostic’ focuses on Supervised, unsupervised, Semi-Supervised, Reinforcement and evolutionary learning. Although the paper had several useful facts, the disease prevalence percentage measured is very low.

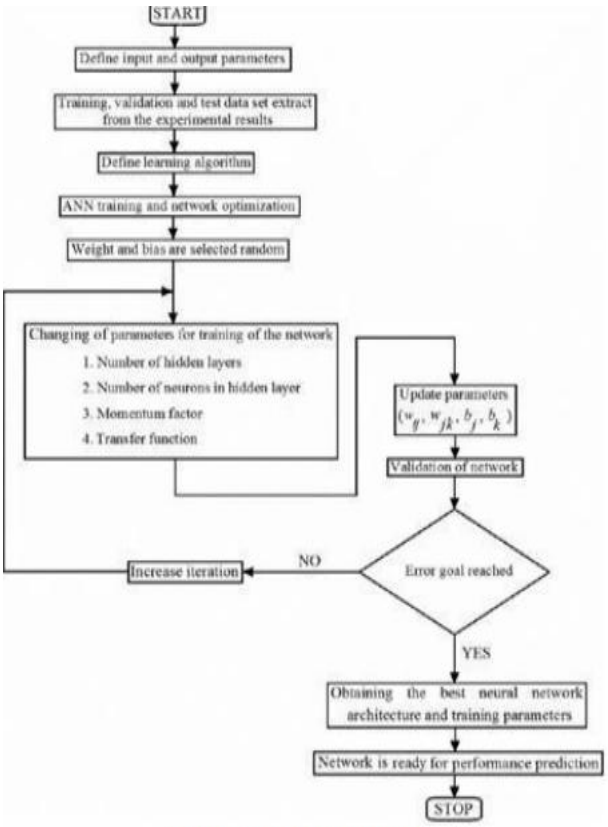
The authors Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh and Gregor Stiglic in the article ‘Early detection of type 2 diabetes mellitus using machine learning-based prediction models’ used Unsupervised, Semi Supervised machine learning models to show that the prediction with SVM could reach the highest accuracy. The drawback of the paper is that it cannot predict the type of diabetes

Proposed System

Grouping is quite possibly the main dynamic strategy in numerous genuine issues. In this, the main focus is to group the information as diabetic or non diabetic and improve the order exactness. For some order issues, higher number of tests were picked however it doesn't prompt higher arrangement exactness. As a rule, the presentation of calculations is high with regards to speed yet the exactness of information grouping is low. The principle

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5443 - 5447
Received 15 May 2021; Accepted 20 May 2021.

objective of our model is to accomplish high precision. Order exactness can be expanded on the off chance that we utilize a large part of the informational collection for preparing and few informational indexes for testing. This study has investigated different arrangement procedures for grouping of diabetic and non-diabetic information. Along these lines, it is seen that methods like Support Vector Machine, Logistic Regression, and Artificial Neural Network are generally appropriate for executing the Diabetes forecast framework.



(3.1)System Design Flowchart

Artificial Neural Network

The artificial neural networks are quite similar to the neural arrangement and working of a brain or cerebrum. Artificial Neural Networks (ANNs) are usually made up of different layers or a block plan, and the shape of the Signal that moves from the front to the back. The back generation is the way in which the forward incitement is used to restore charges on the "front" neural units. For the most part, the Artificial Neural Network organization consists of organizational levels and work, organizational levels include: input layer, cover layer, and performance/yield layer. Information neurons characterize all information quality qualities for the information mining model. In this project the number of neurons taken are 7 as we have 7 attributes in our information set which are:blood glucose levels, blood pressure of a person , thickness of the skin, blood insulin levels, Body Mass Index, diabetes pedigree function, and age. For the hidden layer, the covered up neurons receive contributions from the input neurons and give outputs for generating neurons. A work on the organization of neurons f (x) is numerically characterized as a synthesis of different capacities gi (x), which can also be resisted as an arrangement of different capacities. The enactment job is to make smooth progress when the estimated information changes, much like a small change in input that results in minimal changes in performance. The areas of application include the recognizable test and control framework, quantum science, game and dynamics, design recognition and more.

Support Vector Machines

The Support Vector Machine (SVM), is a group of supervised learning methods that are majorly used in the medical field for classification purposes.It is based on statistical learning and is one of the most robust prediction methods.Apart from linear classification SVMs can successfully perform non-linear classifications as

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5443 - 5447
Received 15 May 2021; Accepted 20 May 2021.

well. It is based on the theory that given a set of points each belonging to one of the two classe, SVM will predict that in which class a new point will lie in, for our case it would be diabetic or non-diabetic. People have started gaining interest in SVM as a classifier model in research related to machine learning. It is found that SVM offers high accuracy results in terms of classification . SVM is a suitable model for binary classification so we chose SVM to predict diabetes where the feature dimension is 7.

Methodology

The main aim of this work is to design a prediction machine learning algorithm which uses the significant features and gives accurate results for the classification which should be as close to the outcomes of the clinical tests. The proposed system focuses on selecting the correct attributes that will help in the prediction of the occurrence of diabetes mellitus in an early stage.

Data Analysis

Once prepared, Neural organizations could anticipate different patient outcomes based on inconsistent factors.The following glycemic control indices were recorded for each patient: plasma glucose levels, mean and SD of recorded capillary blood glucose levels; VG during the first 7 days and during the 28 days of the study; The glycemic variability was measured as the mean SD of the daily capillaries. Glucose level, the coefficient of variation (CV) according to the equation $Glu_{CV} = Glu_{SD} \times 100 / Glu_{M}$ and the glycemic lability index (GLI). The GLI was calculated using the Ryan equation modified for critically ill patients and corrected for the number of blood extractions as follows:

$$GLI \{((mmol/l)^2/h) \times day^{-1}\} = \sum^{N}_{n=1} (Gluc_n - Gluc_{n+1})^2/\{ (N-1) \times (h_{n+1}-h_n)\}$$

Where $Gluc_n$ is the nth reading for the patient at the time n measured in mmol/L, Nis the total number of readings performed that day, and h_n is the time in hours of the nth reading at time n.

Results

Accuracy Table

CLASSIFICATION MODEL	ACCURACY
k-NN	**72.55**
J48	**76.12**
SVM	**81.31**
Random Forest	**70.56**

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5443 - 5447
Received 15 May 2021; Accepted 20 May 2021.

Error Table

CLASSIFICATION MODEL	ERROR RATE
k-NN	**33.56**
J48	**29.72**
SVM	**24.15**
Random Forest	**35.63**

Conclusion

This document provides us with a system for the early prediction of diabetes for people who may have it in the future. This can bring the cure to millions of people as they can prevent it early and is a boon to the medical world. It could be clearly seen that the Support Vector Machine proved to be the most effective and accurate in predicting correctly whether a person has diabetes mellitus or not.

References

1. Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction
2. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering
3. Alan Siper, Roger Farley and Craig Lombardo, “Machine Learning and Data Mining Methods in Diabetes Research”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.
4. Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.
5. Berry, Michael, and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997
6. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016
7. Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform.