International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Patient Readmission Prediction Due to Diabetes using Machine Learning Classification

P. Adlene Ebenzer, Rishikesh bhattalwar, Harshit Patel, Rupesh Kumar

Abstract-Diabetes mellitus is a general illness of body caused due to a group of metabolic disorders and conditions where the sugar level readings over a prolonged period are very high. It affects different organs of the human body which thus harm a large number of the body's system and tissues to micro level, particularly blood veins, nerves and also skin. It usually occurs when body has malfunctioned pancreas or there is insulin resistance. As there have been huge advancements in machine learning field which is widely used in solving different real life community level problems including health care and also due to presence of vast data from different medical care centres and hospitals which can play an important role in building a machine learning model which can predict weather a person is suffering from diabetes by using data sets from different hospitals and medical care centres. We have collected data from different hospitals over past 10 years where we consider different factors that determine whether admission of a person into hospital is required or not. Depending upon the previous medical history of the person, it can be determined that weather or not the person is readmitted into the hospital and within what time period. Hence in this paper we try to construct a model where we predict weather a person is readmitted to hospital or not .Main focus is to help health care professionals, medicine practitioners and people who suffer from its symptoms improve their treatment process by predicting the chance of the person having diabetes, here by decreasing cost of treatment and enabling the concerned person to be self-aware of their medical condition.

Keywords: machine learning, logistic regression, data classification, diabetes.

I. INTRODUCTION

Diabetes is an illness that occurs when the body doesn't make or use the hormone called insulin properly or to its optimal use. It causes too much blood glucose (sugar) to build up in the blood. There are 2 main types of diabetes that are medically visible. Type 1 diabetes occurs when one's body doesn't produce any insulin whatsoever at all. It is usually referred to as worse diabetes because it is usually found not only in children but also teenagers in range from 15-19, but it may appear in adults, too. Type 2 diabetes is indicated at the time when your body doesn't produce enough insulin or doesn't use the insulin as it should.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Ms. P Adlene Ebenzer*, Assistant Professor(O.G), SRM Institue of science and technology.
Rupesh Kumar, Pursuing B.tech, SRM Institue of science and technology.
Harshit Patel, Pursuing B.tech at SRM Institue of science and technology.
Rishikesh Bhattalwar, Pursuing B.tech, SRM Institue of science and technology.

In the past, physicists determined that only adults and elderly people were at risk of developing type 2 diabetes. However, an increasing number of younger lot in the India are now being diagnosed with the deadly disease.
Doctors think this increase is mostly because more children are overweight or obese and are less physically active or it can also be due to the fact that they are malnutrition. Recently , in the past few years there have been new tests results indicating the presence of pre - diabetes.Pre - diabetes occurs when blood sugar levels are more than they should be, but not more or high enough to reportedly be diagnosed as diabetes. Pre-diabetes greatly increases the risk of developing type 2 diabetes usually in adults over 40. The good news is that, if a person is suffering with prediabetes, he or she is reportedly able to prevent or delay the chances of full-blown 2nd type diabetes by making lifestyle changes on daily basis. These include eating a healthy diet, reaching and maintaining a healthy weight, and exercising regularly with keeping the track of the health by visiting the doctor for regular check-ups if in any doubt. symptoms vary from person to person. The early stages of diabetes have very few symptoms as reported. You may not know you have the disease. But by now, the affect may already be affecting to your eyes, your kidneys, and your cardiovascular system. Common symptoms include too much hunger, too much thirst, frequent urination on daily basis, unexpected weight loss or gain, Fatigue or tiredness, unclear visuals, Slow-healing bruises, sore muscles, or bruises. Dry, itchy skin, tingling or numbness in the hands or feet.Detection of diabetes might also depend on its causes which include but not limited to can be: Weight: obesity and malnutrition can be single most dangerous factor which might show signs of diabetes. More overweight one is, Chances of diabetes increase as insulin resistance increases. Age: age can be factor as data shows people older than 50 years of age are more prone to diabetes than compared to younger age group. Family history: If diabetes runs in family then risks of diabetes at a certain age increase.Smoking and alcohol. Alcohol and tobacco use may increase your risk of type 2 diabetes. There are many different tests which can prove the presence of diabetes.

II. RELATED WORKS

1. The prediction of long-term diabetes complication risk is important in our process of medical decisions making. Predefined guidelines for the prevention of Type 2 Diabetes Mellitus (T2DM) helps to calculating the severity. Cardiovascular Disease (CVD) risk to establish appropriate treatment. The objectives of the study is to implements the use of logistic regression which is a part of machine learning techniques towards the development of sophisticated models able to predicts the risk of treatable or untreatable CVD incidence in T2DM patients.

The challenges of handling the unstructured nature of the available datasets is to predict weather a person is suffering to diabetes or not.

The demerits of this journal is poor diabetes prediction accuracy, but in this project we provide a certain criteria and boundaries which help the patients either he should go to hospital or not. There are many different methods for grouping the decisions of the primary model that are applied and similarly assessed. 2. In this they proposed a methods for screening for the presence of type 2 diabetes on the bases of signals obtained from a pulse oximeter. The system that is the screening system contains two parts: the first is that they analyze the various signals obtained from the pulse oximeter, and the second subsist of a machine-learning module. In this project we use logistic regression. The system subsist of a front end the use of that is it withdraw a set of feature form the pulse oximeter signal. The set of features are given as the input of a machine-learning algorithm that is logistic regression that determine the class of the input sample, i.e., whether the person had diabetes or not. The demerits of this journal is that here they uses only physiological measurements. In this project We show a screening method for identify diabetes that has a performance proportionate to the glycated hemoglobin test, does not require blood extraction, and returns results in less than 5 min.

3. Large amount of the health related data is being produced in different phases of health system. Because of the size of the data it will be difficult to process the data and then analysis it. But we can use the different machine learning based approaches such as logistic regression that can be process the data. Machine learning based approaches will provide the effective ways to data for curing of the patients. Even helps in forecast the future of the diabetes disease, by these set of datasets we predict in this project weather there is any need to hospitalize or not to the person .So by using the machine learning technique we reduced the cost of testing the different test for a particular disease. Patients past history for various parameters can be contribute to his probability of different health related problems. By using Association clustering and Time Series based data mining in Continuous data early warning system can be developed, but there is disadvantage of this is that there is less measurements and also less accuracy, so we use logistic regression. This prediction based system can define the disease (diabetes) while analyzing his existing parameters. Some level of care we can protect the patient from disease.

4. Linear modeling is the mostly used common used statistical technique to get on unknown relationship among different essential random variable of interest because of its uniformity and clarify random variable relationship. In our paper, we use or utilize linear models under which we use logistic regression to study glycosylated hemoglobin, which can be a benchmark of the diseases of diabetes from which patient is suffering. We want to search other things from which other predictors or signal indicators have the most suitable predominance power on glycosylated hemoglobin. The dataset which we collected is collected from the Indian Diabetes Association. Meanwhile, the dataset is deficient due to missed data problems. We use utilize different EM algorithm to implement and to study about the linear model such as logistic regression from the inimical missing data. Converging rate and ruggedness against starting or initial values are inspect .Likewise, we had prove the convergence of the EM algorithm in a more efficient general settings. In expansion, we rate the accomplishment of EM at different missing rate and connect the results with other different two other methods that are basically used to deal with missing data. Experimental results depict that the EM algorithm have a improved performance than other different methods used in various missing rates in their applications.

5. Data science has wide range which contains different methods that have the capability to benefit other different scientific fields by throw out a new light on many common questions raised. One such task assigned is help to make predictions on the medical data. Diabetes mellitus or simply diabetes is a threatening disease induce due to the increase level of blood glucose. Various different long established methods, based on physical and chemical inspection, are prepared for diagnosing diabetes. These methods forcefully based on the data mining techniques can be efficiently usable for high blood pressure risk prediction. In our paper, we analyze the initial prediction of diabetes via five diverse data mining methods in addition to it, Logistic regression, GMM, SVM, ANN, ELM, ANN. The research outcomes proves that ANN (Artificial Neural Network) provides the highest accuracy than other techniques, in our paper we use logistic regression instead of ANN for prediction of diabetes.

III. SYSTEM ARCHITECTURE

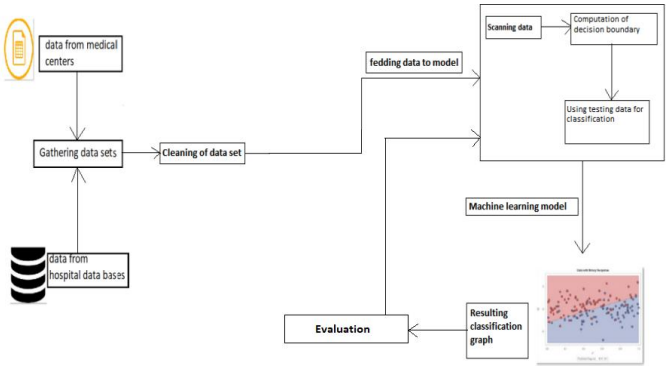


Fig. 1 System architecture

IV. MODULE DESCRIPTION

The architecture diagram has been split up in three parts where each part describes the behaviour of the module. These modules namely are:

Data gathering

Data cleaning

Feeding data to model and evaluation

A. Gathering Of Data Sets

The data gathered from the internet and hospital databases combined gives us a raw data regarding the hospital readmission. The data gathered consists of over 100000 entries and these entries are categorized by 50 columns. These 50 columns are the different parameters on which the data sets will be cleaned, tested and used through the module. The data sets are can be read through panda a function into the module where it will be trained and tested and will produce output after evaluation. The data sets used is in .csv format i.e. Comma Separated Values which are easier to read and visually more comfortable while browsing through data.

Data gathered is consisting of many entries (rows) so that training would be more efficient for the machine learning model. When we train the model on less entries then prediction accuracy is very low as the data sets would be split into training and testing sets where training consists of 75 % of the data and testing consists of 25% of data.

B. Cleaning Of Data Sets

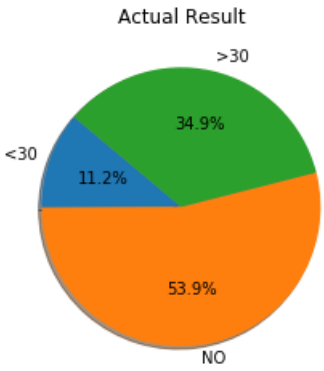
This is an important part of the process where data gathered is cleaned and used for prediction. Why this part is important is that the data must be uniform and must not contain irregularities such as duplicate column names, column names in different cases, inclusion of null values, string values for a binary state etc. These irregularities cause the inaccuracies while predicting. Hence these irregularities are removed by different functions and methods employed by the python libraries. Null values are removed by replacing them with either value 0 or with the mean value of the particular column. Usually mean value is preferred. Column names are converted into same case to increase readability. Duplicate column names are avoided usually by renaming the column according to the data it holds. Data set may hold duplicate string values such as 'YES','NO','NAN', etc which are converted into binary values for reference.

C. Feeding Data To Model And Evaluation

After the data has been cleaned and is ready to be used, it is passed to the logistic regression model. Logistic regression is a classification machine learning algorithm used to classify the given data into two or more groups depending upon the deterministic factors we use on the data set. It uses a sigmoid function given as 1/(1+e^-xt)Where x is the given parameter i.e. the input and t is the evaluation matrix .On using the function , we get an output y which if above the decided criteria is classified into one group , if below it is classified into another. This separation is represented by the classification boundary which separates the output classification groups. After production of an output, is verified for accuracy .if the output we classified into the group is correct then it will be the final output to produced, else sit is checked if the classification is done correctly or not with the evaluation matrix and training data then it will be processed again with the model until there is a definite output classification.

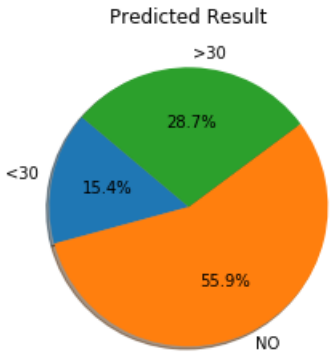
V. CONCLUSION

Hereby this project is concluded by getting the results that were expected from the machine learning model. The model that we used was logistic regression and KNN (k-nearest neighbour) where both are classification algorithms and used on categorical data. After the data cleaning process the data was used by the logistic regression model in order to produce a prediction .The accuracy achieved by the model was approximately 74 %.This means that 74 % chance that the predictions that were produced were accurate .In order to compare this prediction accuracy, we tried to use another classification algorithm KNN which gave approximately just over 60% of accuracy. Hence we concluded that for given data set the, logistic regression produced more accurate results in comparison to KNN algorithm. This statistics were produced after applying the logistic regression classifier. The prediction was close to the expected results as it produced 75% accuracy.

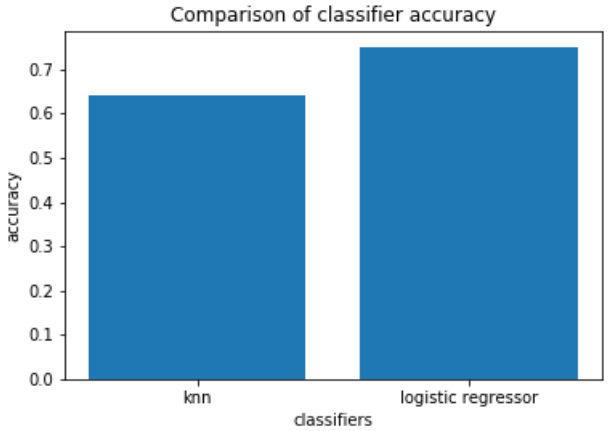


Pie Chart-:1

This is the actual statistics that were originally provided with the data set.



Pie Chart-:2



Histogram :-1

This gives an intuition about the comparison between both the classifiers.

REFERENCES

1. Comparison of machine learning approaches towards assessing the risk of developing cardiovascular disease as a long-term diabetes. (https://www.ieee.org)
2. Type 2 Diabetes Screening Test by Means of a Pulse Oximeter. (https://www.ieee.org)
3. Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction. (https://www.ieee.org)
4. The EM algorithm for a linear regression model with application to a diabetes data. (https://www.ieee.org)
5. Application of data mining methods in diabetes prediction. (https://www.ieee.org)
6. Diabetestestshttps://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451

AUTHERS PROFILE

Ms. P Adlene Ebenzer Assistant Professor(O.G) at SRM Institue of science and technology. Work experience in area of Digital Imaging

Rupesh Kumar 3rd year cse undergraduate Pursuing B.tech at SRM Institue of science and technology.

Harshit Patel 3rd year cse undergraduate Pursuing B.tech at SRM Institue of science and technology.

Rishikesh Bhattalwar 3rd year cse undergraduate Pursuing B.tech at SRM Institue of science and technology