DIABETES PREDICTION USING DATA MINING

By

IIT2018138 CHINMAY TAYADE

Under the guidance of Prof . Shirshu Verma

INTRODUCTION

A group of diseases that result in too much sugar in the blood (high blood glucose).Diabetes mellitus is the fourth most high mortality rate disease in the world and it is also a cause of kidney disease, blindness, and heart diseases. Data mining techniques support a medical decision for a correct diagnosis, treatment of disease in such a way it minimizes the workload of specialists. This project aims to predict diabetes using data mining and machine learning techniques.In this project the objective is to predict whether the person has Diabetes or not based on various features like Number of Pregnancies , Insulin Level , Age ,BMI etc. The dataset used in this project has been taken from kaggle .This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases .

INTRODUCTION

Diabetes is a diseases that result in too much sugar in the blood (high blood glucose).Diabetes mellitus is the fourth most high mortality rate disease in the world and it is also a cause of kidney disease, blindness, and heart diseases. Data mining techniques support a medical decision for a correct diagnosis, treatment of disease in such a way it minimizes the workload of specialists. This project aims to predict diabetes using data mining and machine learning techniques.In this project the objective is to predict whether the person has Diabetes or not based on various features like Number of Pregnancies , Insulin Level , Age ,BMI etc. The dataset used in this project has been taken from kaggle .This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases .

MOTIVATION

Diabetes is an increasingly growing health issue due to our inactive lifestyle. If it is detected in time then through proper medical treatment, adverse effects can be prevented. To help in early detection,technology can be used very reliably and efficiently. Using data mining and machine learning we will build a predictive model that can predict whether the patient is diabetes positive or not.

APPLICATION

The recent report of WHO shows a remarkable hike in the number of diabetic patients and this will be in the same pattern in the coming decades also.Diabetes can also act as a means for other diseases like heart attack ,kidney damage and somewhat blindness.Early identification of diabetes is an important challenge.

Data Mining plays an important role in diabetes research. An early prediction of diabetes can help in reducing the risk for more deadly disease and can help to take precautions from early stage.The implementation of this predictor will help to reduce the stressful process, doctor's face during prediction of diabetes.

PROBLEM STATEMENT

This project was visualised keeping in view the increase in the number of diabetic patients. Many people does not know that they are suffering from diabetes.When diabetes is not treated it could become life threatening disease. Technology can help in early identification of people who are prone to diabetes. With the help of data mining, machine learning techniques and technology the risks for Type-II diabetes can be identified early and with proper treatment Type-II diabetes can be controlled thereby reducing negative impacts of diabetes. To help in early detection, technology can be used very reliably and efficiently.

LITERATURE SURVEY

S.No	Paper Title / Author	Basic Approach	Achievement	Limitation
1.	Classification of Diabetes using Random Forest with feature Selection Algorithm K.Koteswara Chari, M.Chinna babu, Sarangarm Kodati	Predict the level of occurence of diabetes using Random Forest , a Machine Learning Algorithm.This was done using the patients Electronic Health Record(EHR).	Decision tree accuracy of 75.2.Bagging with Decision Tree but rather accuracy 81.3, Random Forest accuracy 85.6, Random Forest with Feature Selection accuracy 92.02.	It can only be used for low-resource treatment process . The precision can be improved using more technologies and with true parameters.

S.No	Paper Title/Author	Basic Approach	Achievements	Limitations
2.	Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach Rahul Joshi,Preeti Mulay,Minyechii Alehegn	The proposed method provides better accuracy of 93.62% in case of PIDD using stacking meta classifier. In case of large dataset 130-us hospital an ensemble method provides better accuracy than single prediction algorithm.	In the proposed system two datasets one is large (130_US) and other is small (PIDD) used for analysis. The proposed method provides better accuracy of 93.62% in case of PIDD using stacking meta classifier. In case of large dataset 130-us hospital an ensemble method provides better accuracy than single prediction algorithm.	Implementation of hybrid algorithm is difficult and ambiguity in the output could be found.The precision can be improved using more technologies and with true parameters.

S.No	Paper Title/Author	Basic Approach	Achievements	Limitations
3.	Prediction of Diabetes using Ensemble Techniques Prema N S, Varshith V, Yogeswar J	In this paper, various machine learning algorithms are applied to predict diabetes, based on specific attributes. The performances of the algorithms are compared in terms of accuracy, voting based ensemble techniques is applied for the normalized pima diabetes data for which a highest accuracy is achieved.	Prediction of diabetes is done using ensemble voting classifiers for pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for data set by using 10-fold cross validation and by spitting data into 30% testing and 70% training.	For different classification techniques more resources means on the dataset different Algos are applied which could be time and cost consuming even 10-fold cross validation is even tremendous.

S.No	Paper Title	Basic Approach	Achievements	Limitations
4.	Diabetes prediction using Machine Learning Techniques Pramila M .Chawan, Tejas N.Joshi	The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN.	Machine Learning has a great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and generic diabetes risk dataset.	Not much accuracy is achieved in the process.

ASSOCIATED CHALLENGES

1. The main problem is that we are trying to solve to improve the accuracy of the prediction model, and have to make model adaptive to more than one dataset.
2. Extraction of pattern in the dataset or its analysis and identifying relationships in the dataset .

3. Efficient training of the dataset so that we can predict the possibility of the disease with greater accuracy.

4. The values provided by the user must be accurate for accurate prediction,if the values are provided incorrectly the model will give wrong result.

WORK PLAN

- Technologies To Be Used:

1. Python
2. Flask
3. Scikit Learn
4. Pandas
5. Numpy

- Dataset to be used:

The dataset to be used in this project has been taken from kaggle .This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases .

- Work plan:

We will be building and hosting a Flask web app.

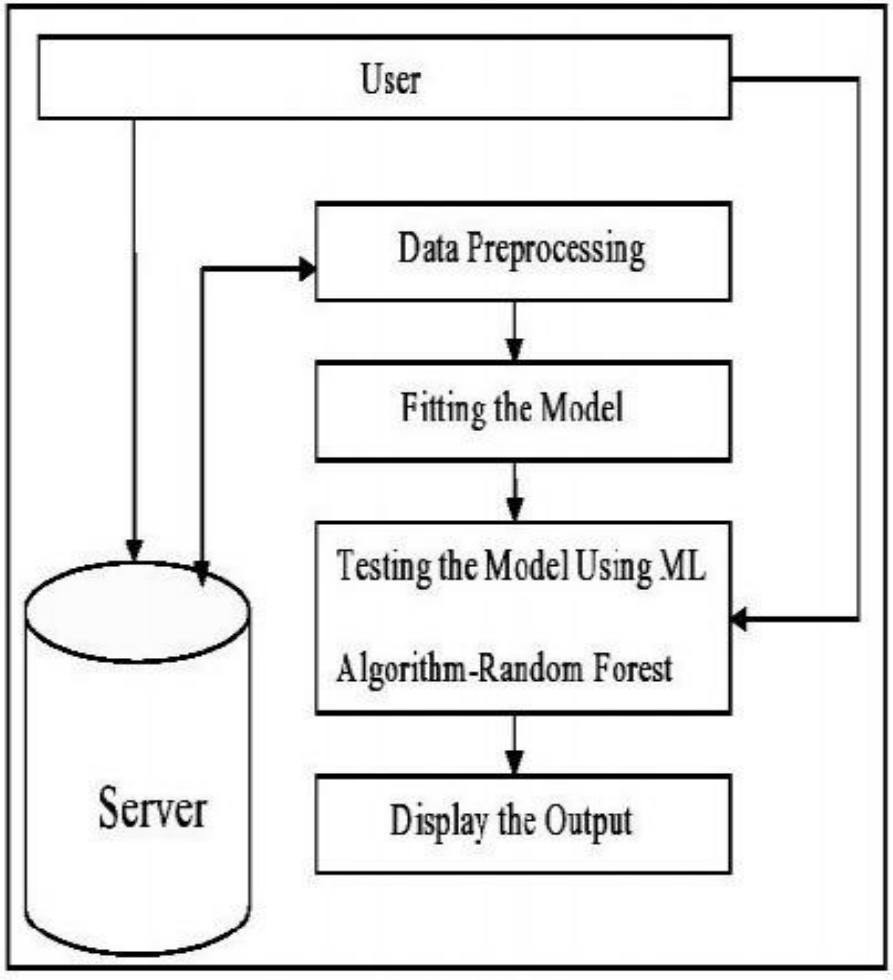
1. Use Data Mining to extract the usable dataset from the dataset of National Institute of Diabetes and Digestive and Kidney Diseases.

2. Used Random Forest Classifier[1] to predict whether diabetes positive or not. Random Forest algorithms are often used for classification and regression tasks and also it is a type of ensemble learning method[2].The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

3.Train a Machine learning Model using scikit - learn.

4.A user has to put details like Pregnancies - Number of times Pregnant ,Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test,Blood Pressure - Diastolic blood pressure (mm Hg),Skin Thickness - Triceps skin fold thickness (mm),Insulin Level - 2 hour serum insulin (mu U/ml),Body Mass Index(BMI) - Body mass index (weight in kg/(height in m)^2),Age - Age of the person.

5.Once it gets all the field information , the prediction is displayed on the page whether the person is diabetes positive or not.


User
Data Preprocessing
Fitting the Model
Testing the Model Using ML
Algorithm-Random Forest
Server
Display the Output

CONCLUSION

Our Project aim to built a web app using Data Mining to extract hidden knowledge using large amount of diabetes related data to predict the diabetic risk level of patient with higher accuracy . It can prove to be a very good web app in early prediction of diabetes thus reducing the risk of more deadly disease in near future so that the person can take necessary medication and precautions from early stages to avoid such risks.

References

[1]https://www.ijitee.org/wp-content/uploads/papers/v9i1/L35951081219.pdf

[2]https://www.researchgate.net/publication/337275136_Diabetes_Prediction_Using_Machine_Learning_Techniques

[3]https://www.researchgate.net/publication/338819080_Diabetes_Analysis_And_Prediction_Using_Random_Forest_KNN_Naive_Bayes_And_J48_An_Ensemble_Approach