DIABETES MELLITUS PREDICTION USING

MACHINE LEARNING TECHNIQUES

Submitted by,

Karthick S

Karthik V

Prabagaran G

**ABSTRACT**

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in2040, the world’s diabetic patients will reach 642 million,which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health for accurate predictions. The aim of the project is to build a machine learning model that can efficiently discover the rules to predict the risk level of patients using Machine Learning Algorithms based on the given parameter about their health (In this project we are using decision tree to predict diabetes mellitus). Then we evaluate the performance of the model in terms of different parameters like classification accuracy which comes under Supervised. An Web application is built from where the patient health features are entered and depending on the entered parameters, the machine learning model integrated to application will predict the type of diabetes and according to the type of diabetes diet plan for the person will be displayed on the UI.

**INTRODUCTION**

**OVER VIEW**

Diabetes is one of the major health problems of all over the world. Diabetes mellitus is classified into four broad categories: type 1, type 2, gestational diabetes and other specific types. All forms of diabetes increase the risk of long-term complications. These typically develop after many years (10–20), but may be the first symptom in those who have otherwise not received a diagnosis before that time. The criteria for diagnosing diabetes in pregnancy

have been given the World Health Organization.

criteria are as follows,

1. x fasting plasma glucose ≥ 7.0 mmol/l (126 mg/ dl)

2. 2-hour plasma glucose ≥ 11.1 mmol/l (200 mg/dl) following a 75g oral x glucose load.

3. random plasma glucose ≥ 11.1 mmol/l (200 mg/ dl) in the presence of diabetes symptoms.

Diagnostic criteria for diabetes in non-pregnant individuals are based on the relationship between plasma glucose values and the risk of diabetes-specific Micro vascular complications .

Data mining predicts the future by modelling. Predictive modelling is the process by which a model is created to predict an outcome. The data mining process for diagnosis of diabetes can be divided into five steps, though the underlying principles and techniques used for data mining

diabetic data bases may differ for different projects in different countries. Data mining is one of the "Knowledge Discovery in Databases" processes. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. This process has become an increasingly pervasive activity in all areas of medical science research. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and statistics, including hypothesis testing, clustering, classification, and regression techniques. In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics,genetics, medicine, and education.

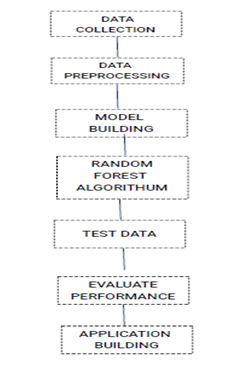
**PURPOSE:**

The aim of the project is to build a machine learning model that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health. Then we evaluate the performance of the model in terms of different parameter like classification accuracy AUC-ROC Curves.A Web application is built from where the patient health features are entered and depending on the entered parameters, the machine learning model integrated to application will predict the type of diabetes and according to the type of diabetes diet plan for the person will be displayed on the UI.

**THEORITICAL ANALYSIS**

**BLOCK DIAGRAM:**

Block diagrams here represents the data collection,data preprocessing and further steps we also it includes the decision tree algorithm then at last we done web application.



**EXPERIMENTAL INVESTIGATIONS**

**DATA COLLECTION**

For the purpose of this study, Pima Indian Diabetes Dataset (PIDD) is considered as it is the best dataset for the present study. It contains 768 records. Every record has 9 attributes out of those one attribute is class variable. All 9 attributes contains only numeric data. Each record contains information about single patient.

**DATASET**

PREG: This column indicates how many times a person is pregnant.

PLAS: This indicates plasma glucose concentration at 2 h in an oral glucose tolerance test.

PRES: This shows diastolic blood pressure.

SKIN: This indicates thickness of skin at triceps.

TEST: This demonstrates insulin level.

MASS: It demonstrates body mass index which is ratio of weight and height.

PEDI: It demonstrates how much probability a person can inherit diabetes from ancestors.

AGE: It provides or shows age of the person.

Class: It is a variable which contains only 0 or 1 1 indicates person having diabetes and 0 indicates person not having diabetes.

**DATA PREPROCESSING**

**ImportingThe Libraries:**

Pandas,Numpy,Matplotlib.pyplot,Seaborn and Sklearn are some llibraries we used here.

**Importing The Dataset :**

We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program) and read it using a method called read\_csv which can be found in the library called pandas.

**DataVisualization :**

To get clear understanding about each column we visualize the data

**Taking care of Missing Data:**

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common ideas to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.Initatialy there was no missing data but most of the columns are filled with 0, so we tried to replace the 0 with mean initally. When replacing the 0 with mean value tha accuracy was not that good which we expected.Then we tried with median in which it showed better accuracy.

**Splitting x and y and Label Encoding and OneHotEncoding:**

we need to split the dataset where input column which is independent variable and output column(class column) which is dependent variable.In This dataset first 8 columns are stored in x and last column is stored in y. No encoding required becuase

the values are not strings.

**Feature Scaling:**

The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data. A lot of machine learning models are based on Euclidean distance. If, for example, the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square difference dominates over the other square difference. In the machine learning equations, the square difference with the lower value in comparison to the far greater value will almost be treated as if it does not exist. We do not want that to happen. That is why it is necessary to transform all our variables into the same scale.

**Splitting dataset into independent variable and dependent variable :**

Now we need to split our dataset into two sets — a Training set and a Test set. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import test\_train\_split from model\_selection library of scikit. Now to build our training and test sets, we will create 4 sets— X\_train (training part of the matrix of features), X\_test (test part of the matrix of features), Y\_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices) , Y\_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices). We will assign to them the test\_train\_split, which takes the parameters

— arrays (X and Y), test\_size.

**MODEL BUILDING**

**Training And Testing The Model:**

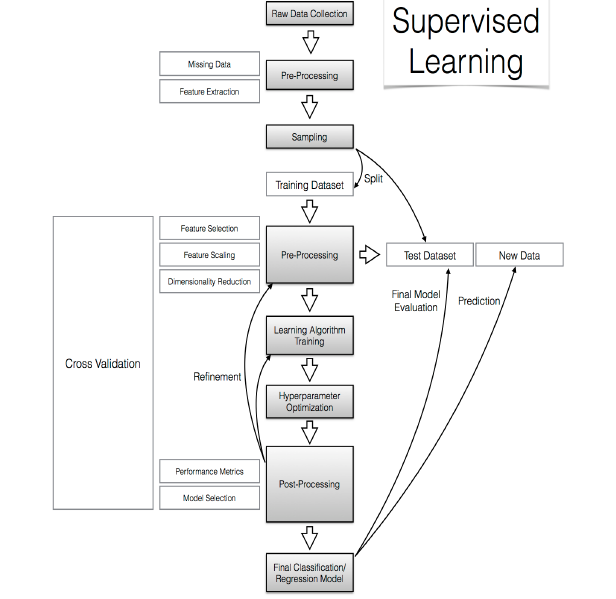
A machine learning model can be a mathematical representation of a real-world process. To generate a machine learning model you will need to provide training data to a machine learning algorithm to learn from.Here we implemented decision tree Classificatiion Algorithm to build our model.The model uses any one of the models that we had chosen. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data.

**Evaluation:**

Once this is done we can calculate the performance of the decision tree model by calculating accuracy.If the model shows accuracy more than 80 it menas model is trained well.AUC curve is also drawn

here.

**FLOWCHART**



**RESULT**

We have got the desired results of more than 80% accuracy for prediction of diabetes by using different classifiers. And the AUC curve also shows about 0.77 .

**ADVANTAGES & DISADVANTAGES**

With this we can know our risk factor. Early diagnosis of diabetes and pre-diabetes is important so that patientscan begin to manage the disease early and potentially prevent or delay the serious disease complications that can decrease quality of life. And there is also a disadvantage in this, in some extent there may be wrong prediction.

**APPLICATION**

By making the algorithm predict correctly we implemented that in a web applicatioṇ. We have made a user friendly web applicationin which user can know the riskfactor and also there are some messages and tips in our application which will be more helpful to the user.Web application has about diabetes mellitus, Risk factors of diabetes mellitus

and also some tips to keep our body health.This is applied in hospitalites in which tech can predict the risk factor of the patient easily.

**CONCLUSION:**

The machine learning methods can support the doctors to identify and cure diabetic diseases. We shall conclude that the improvement in classify ation accuracy helps to make the machine learning models get better results. The performance analysis is in terms of accuracy rate among all the classification techniques such as decision tree, logistic regression, k-nearest neighbors, naive bayes, and SVM , decision tree. We have also seen that the accuracy of the existing system is less than 70% hence we proposed to use a combination of classifiers . We have found

that our system provides us with 84.44 % of accuracy for decision tree Classifier.

