**Prediction of diabetes**

**ATUL**

**13-05-2021**

# ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. 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 different Classification techniques including: DecisionTree, Knn etc. This project also aims to propose an effective technique for earlier detection of the diabetes disease.

# Table Of Contents

1.INTRODUCTION

2.EXISTING METHOD

3.PROPOSED METHOD WITH ARCHITECTURE

4.METHODOLOGY

5.IMPLEMENTATION

6.CONCLUSION

# INTRODUCTION

**1.Diabetes**

Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in 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), according to the etiopathology of the disorder. T2D appears to be the most 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 thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis 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 glucose levels (fasting plasma glucose = 7.0 mmol/L.

2**.Problem**

**The problem tell us to train on different models and try to find out best model which is suitable to predict diabetes**

# EXISTING AND PROPOSED METHOD WITH ARCHITECTURE

The existing methods are various classification technique in machine learning.

The various classifiers are:

Naive Bayes Classifier: This classifier can also be known as a Generative Learning Model. The classification here is based on Bayes Theorem, it assumes independent predictors. In simple words, this classifier will assume that the existence of specific features in a class is not related to the existence of any other feature.

Decision Trees: This classification algorithm builds the regression models. These models are builded in form of structure which is similar to tree – a tree like structure is created by this classifier. It keeps on dividing the data set into subsets and smaller subsets which develops an associated tree, incrementally

Random Forest: This classification algorithm are similar to ensemble learning method of classification. The regression and other tasks, work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees

Nearest Neighbor:As the name suggests the nearest neighbour algorithm is based on the nearest neighbour and this classification algorithm is supervised. It is also called as k- nearest neighbour classification algorithm. A cluster of labeled points are used to understand how the other points should be labelled.

* Support vector machine (SVM): This is also one of the classification algorithm which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the n represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.
* **Gradient** **boosting** is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient **boosted** trees, which usually outperforms random forest.

# PROPOSED METHOD

The proposed system predicts the disease of diabetes in patients with maximum accuracy. We shall talk about various machine learning, the algorithm which can help in decision making and prediction. The xgboost has highest accuracy.So I will proposed xgboost

0 KNN 0.729282

1 SVC 0.740331

2 LR 0.762431

3 DT 0.745856

4 GNB 0.734807

5 RF 0.767956

6 GB 0.773481

7 xgbc 0.779006

8 adab 0.779006

**START**

Define features and output variable

Import machine learning algorithms and append them in one list

Now split the data into training and test set using train\_test\_split

We also find score using k-cross validation

After training every model ,we find their accuracy score. Here we can clearly see that XGBoost has highest accuracy score.

End

Finally we find confusion matrix and draw ROC Curve

# Methodology And Implementation

1.Business Understanding

We are going to predict whether a patient have diabetes or not.

And we are going to solve through classification techniques

2.Analytical Approach

We can use patient data to solve this problem. Patient data must include glucose ,insulin level. By examining all the related features we can reach to the solution.

3.Data collection

Data is already provided here but we can also download it from Kaggle provided by PIMA india.

4.Data Prepration

From the histograms we can observe that many patients have zero BMI,bloodpressure,glucose,insulin,skin thickness but we cannot cannot remove features like insulin and skin thickness because we will loose a valuable informations and we also cannot replace by mean value because suppose we replace blood pressure with mean value then it will send wrong message to the model.For insulin and skin thickness we also cannot remove zero values because they consist of major part of the data.So we will only remove that rows for which all together values pf bmi,bloodpressure and glucose are zero.

5.**Modeling**

We define the features and dependent variables and split the data set into train and test data.

After that we fit the training data in all classification models.

6.**Evaluation**

We evaluate using accuracy score and we get the following result:

KNN 0.729282

1 SVC 0.740331

2 LR 0.762431

3 DT 0.745856

4 GNB 0.734807

5 RF 0.767956

6 GB 0.773481

7 xgbc 0.779006

8 adab 0.779006

# Implementation

In this section we shall discuss about the actual steps which were implemented while doing the experiment. We shall explain the stepwise procedure used to analyse the data and to predict the data accuracy for prediction of diabetes. The system consists of the following main steps:

We have selected a diabetic dataset named PIMA Indian Diabetes Dataset which consists of 768 instances classified into two classes : diabetic and non-diabetic with eight different risk factors: number of times pregnant , plasma glucose

concentration of two hours in an oral glucose tolerance test , diastolic blood pressure, triceps skin fold thickness, two hour serum insulin , body mass index , diabetes pedigree function ang age.

Feature Selection is the procedure where we automatically or manually select those features which contribute most to your prediction variable or output you are interested in. If there is irrelevant features in our data then it can decrease the accuracy of the models.

1. We are taking a diabetic dataset which is PIMA Indian dataset.
2. For pre -processing of data we have use histogram and check distribution of every features.
3. We found Adaboost and XGBoost to be the best out of all the fclassifiers in the aspects of accuracy, since they give better accuracy.
4. Below I have provided confusion matrix of all classifiers:

[[ 32 30]

[ 19 100]]

KNN

[[ 29 33]

[ 14 105]]

SVC

[[ 37 25]

[ 18 101]]

LR

[[39 23]

[23 96]]

DT

[[38 24]

[24 95]]

GNB

[[ 34 28]

[ 14 105]]

RF

[[ 38 24]

[ 17 102]]

GB

[[ 37 25]

[ 15 104]]

xgbc

[[ 39 23]

[ 17 102]]

adab

# Conclusion

1. **The machine learning methods can support the doctors to identify and cure diabetic diseases. We shall conclude that the improvement in classification 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, k-nearest neighbors, naive bayes, SVM , random forest , adaboost , xgboost. We have also seen that the accuracy of the existing system is less than 80% hence we proposed to use a combination of classifiers known as Hybrid Approach. Hybrid approach takes advantage by aggregating the merits of two or more techniques. We have found that our system provides us with 74.5 % of accuracy for Decision Tree Classifier, 77.90% accuracy for XGBoost Classifier. We have therefore found that the best among all the above classifiers is XGBoost and AdaBoost.**