International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1365–1373

Type 2 Diabetes Prediction Using Lightgbm

Abha Parihar [1], Suraj Bhoste [2], Shubham Patil [3], Shreya Kapgate [4], *Sanjeev Wagh* [5]

[1,2,3,4] *Students, Department of Information Technology,*

Government College of Engineering Karad, India

[5]*Head, Department of Information*

Technology, Government College of

Engineering Karad, India.

Abstract

Today every second person in the world is suffering from diabetes. Diabetes occurs when the level of glucose abnormally increases in the blood. Among the various types of diabetes Type 2 diabetes is the most common type. The symptoms in type 2 diabetes develop very slow or are absent in the patient. So it becomes almost impossible for the patient to recognize type 2 diabetes in the early stages. Conventional methods take too much time and effort. The proposed system predicts type 2 diabetes in patients using LightGBM algorithm on Pima Indian Dataset. The Pima Indian Dataset has eight features that were chosen to form the basis for forecasting the presence of diabetes within five years in Pima Indian women. Advantages of LightGBM over other algorithms include its faster training speed and efficiency, lower memory usage, better accuracy. To compare the performance of LGBM we used SVM and Random forest algorithms on the same dataset. LightGBM (cv) gives accuracy 0.9062 which is higher in comparison to SVM and Random forest.

Keywords: *LightGBM, Support Vector Machine, Random Forest, Pima Indian Dataset, Gradient Boosting Decision Tree, Type 2 Diabetes.*

1. Introduction

Diabetes is a very common disease causing a stress for large portion of population of world. Diabetes occurs when the level of glucose abnormally increases in the blood as body cannot metabolize carbohydrates properly. Type 1, Type 2 and Gestational are the type of diabetes. Among these Type 2 diabetes is the most common type of diabetes, accounting for around 90% of all diabetes cases [1]. Type 2 diabetes occurs when blood sugar level is too high. Blood glucose is the main source of energy, which comes from the food we eat. Insulin helps glucose get into our cells from bloodstreams to get energy. But in type 2 diabetes, instead of moving into cells, glucose remains in bloodstream. As a result blood sugar level increases, so more insulin is produced by pancreas and eventually the cells producing insulin become impaired and can't meet the need of body. Weight, fat distribution, inactivity, family history, age are some factors that may increase the risk of type 2 diabetes. Possible complications faced by the patient could be nerve damage, kidney damage, heart attack, vision loss, slow healing of wounds.

There are a lot of ways out there to prevent this disease. These ways mainly focus on improving lifestyle such as keeping your weight less, eating a balanced diet and getting enough exercise. Development can be done in this field by inventing some wireless sensors for early prediction of disease. Gene structure concept could be used to design wireless sensors [2]. There are many problems related to this disease so, detection is very important in diagnosis of this disease. Conventional methods take too much time and efforts. In this paper we have used LightGBM algorithm for classification of Pima Indians Dataset. LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithm. To compare performance of LightGBM we used SVM and Random forest algorithms on the same dataset. The Pima Indian Dataset has eight features that were chosen to form the basis for

ISSN: 2233-7857 IJFGCN

forecasting the presence of diabetes within five years in Pima Indian women.

2. Literature Survey

In medical industry, various studies are being done on Type 2 Diabetes Detection based on Pima Indian Diabetes Dataset.

In [3] a system is proposed using CNN and CNN- LSTM. The author used combination of both algorithms to detect the disease. CNN gave an accuracy of 93.6% and CNN-LSTM combination gave 95.1% accuracy. LSTM is efficient only for large dataset, so we need to take large dataset for using LSTM. The main advantage of using this method is that Deep learning techniques were introduced for the first time for the diagnosis of diabetes using HRV data as input.

In [4] Type 2 Diabetes is predicted by machine learning techniques like SVM, Random Tree (RT) and ANN. Various results were obtained in which the accuracies were obtained as follows: SVM gave 90.1%, ANN gave 88.02% and RT gave 83.59%. The training accuracy for Diabetes dataset is 65.8% and the testing accuracy is 78.2% for the SVM classifier. The predictions are slower, which may lead to major challenges for applications.

Stacked Autoencoders used in [5] proposes a Deep Neural Network framework for diabetes data classification. This classification model achieved accuracy of 86.26%. The author has used many evaluation metrics such as precision, recall, specificity and F1 –score for the evaluation of the model. The important thing about the model is that this model gives better performance than other models with precision value of 90.66% and recall of 87.92%. This model can be used as powerful tool for the disease diagnosis process.

In [6], the author has used Naive Bayes Network to predict patients with type 2 diabetes. Classifier was applied to construct Naïve Bayes model. The accuracy of the resulting model was 72.3%.The advantage of this model that it observes the uncertainty and can observe the system change for the evaluation of diagnosis procedure.

In [7] the author has used two common boosting algorithms, Adaboost.M1 and LogitBoost. LogitBoost is slightly more accurate than Adaboost. M1 classification model. The accuracy of LogitBoost classification was 95.30% using 10- fold cross validation. In this model, both of the algorithms gave excellent performance when used for the classification modelling of the respective disease based on the large amount of data. In [9] author gives the solution of prediction of disease by using XGBOOST. In this paper the accuracy is being increased by xgboost. The main advantage of this method is that it is used to predict the disease with maximum accuracy and fast execution time. After many iterations the obtained accuracy was increased from 77% to 90%. The speciality of using xgboost is that is works on generic loss which gives maximum execution time of almost three times faster than that of Adaboost algorithm.

For the detection of type 2 diabetes we are using LightGBM algorithm. LightGBM is essentially an improvement on the Gradient Boosting Decision Trees. Although conventional GBDT is being widely used for its accuracy and efficiency some improvements can be made to it to obtain more efficient and a faster algorithm without doing trade- off between accuracy and efficiency.

The larger the data instances the more training time an algorithm requires. This gives birth to the data sampling in which only a subset of data is used to draw satisfactory conclusions from the dataset. Some papers have used data sampling to reduce the training time of the GBDT [11] [12].

In LightGBM [8] two novel techniques have been proposed to solve this challenge.

Gradient-based One-Side Sampling (GOSS)

In GBDT those data instances with higher gradient generally have higher information gain. This

ISSN: 2233-7857 IJFGCN

suggests that data instances with gradients greater than some threshold will have more information gain than those with gradients which are lesser than the threshold. So the data instances with small gradients do not or contribute very less to the information gain of the model. This can be used to randomly drop the data instances with small gradients in order to make the number of total data instances smaller.

This is relatively a better technique than randomly sampling all the data instances.

Exclusive Feature Bundling (EFB)

In many real datasets there are large number of features which decides the outcome of the model. Hence choosing right set of feature yields very good accuracy, efficiency and can avoid overfitting model [13] [14] [15]. In EFB bundling of the features which are exclusive which means that they rarely take any nonzero values simultaneously. Then optimal bundling problem is reduced to a graph coloring problem and can be solved by greedy method.
Since GBDT is an ensemble approach to decision tree the main cost in LightGBM lies in learning the decision tree. Finding the best split points is a very time consuming task. LightGBM uses histogram- based algorithm. This algorithm does not find the best split by brute forcing, instead it creates discrete bins in which features are stored and uses these bins to construct histograms during training of model.

3. Proposed Architecture

A. Objectives

Following are the objectives of proposed system:

1. To implement LightGBM on Pima Indian Dataset for the classification of dataset for Type 2 diabetes.
2. To achieve maximum accuracy on the dataset using LightGBM.
3. To compare other conventional algorithms such as Random Forest and SVM with LightGBM in terms of accuracy, time to train data and memory for that particular algorithm.

B. Dataset

In this system, PIMA Indian Diabetes Dataset has been used which is available at Kaggle Platform. The main objective is to diagnostically predict the patient that the patient is suffering from diabetes or not. All the data present in the dataset is about females of Pima Indian heritage and are at least 21 years. This dataset contains medical record of females infected with Type 2 Diabetes with the minimum age of 21. This medical records contains total 9 attributes through the disease will be predicted. The 9 main features have been shown below in the table [10].

Feature no.	Feature Description
1	Number of times pregnant
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure(mm Hg)
4	Triceps skin fold thickness(mm)
5	2 Hour serum insulin (mu U/ml)

ISSN: 2233-7857 IJFGCN

6	Body mass index(weight in kg/(height in m) ^2)
7	Diabetes pedigree function
8	Age (Years)
9	Class variable (1:tested positive for diabetes,0:tested negative for diabetes)

Table 1: Dataset Feature Description

C. System Architecture

The proposed system represents the architecture for type 2 diabetes prediction. This system consists of standard dataset which is obtained from Kaggle .This dataset is further preprocessed so that the dataset can be used efficiently to predict the outcomes of the disease. This preprocessing is done manually by changing values in the columns based on the outcome of that particular row. Python's sklearn library has used for splitting the dataset in training and testing parts.

After pre-processing, the data is segregated. This segregated data is trained with the LightGBM Classifier. After the prediction is done we can obtain the results of the process. To evaluate the model many metrics can be used. We have given more emphasis on the accuracy, precision and f1score as these metrics greatly influence the betterment of the model.

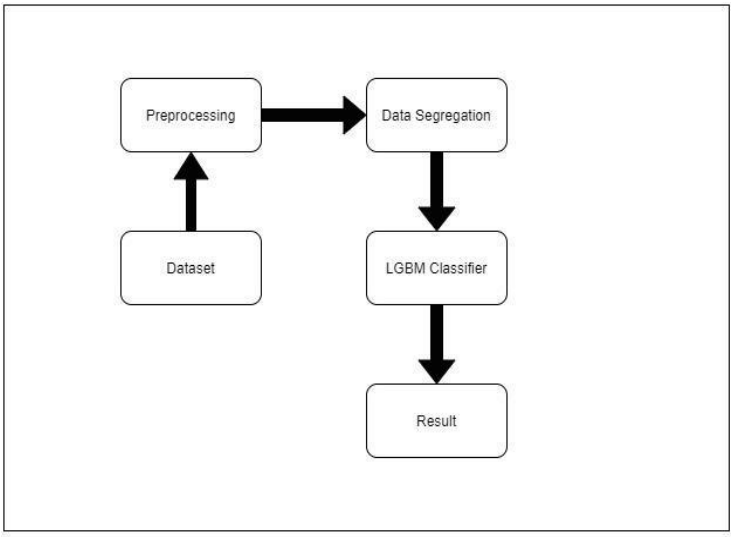

Fig 1: System Architecture

4. Implementation Details

A. LightGBM

LightGBM is a type of gradient boosting framework which grows the tree horizontally. It handles large size of data and requires less memory.

ISSN: 2233-7857 IJFGCN

LightGBM follows following steps:

1. Input – dataset, no of iterations
2. Build a decision tree (learner)
3. Loop
    3.1 Gradient descent to minimize the loss
    3.2 Choose a weak learner to make predictions
    3.3 Optimize the loss function by adding a decision tree (weak learner)
    3.4 If loop has run no. of iterations then
        3.4.1 Stop
4. End loop
5. Return the trained model

B. SVM

SVM algorithm is a powerful supervised learning algorithm that can be used for building both classification and regression models. It is based on hyperplanes that are used to classify a set of given objects. It is used for binary classification like presence or absence or diabetic or non-diabetic etc. SVM follows following steps:

1. Define features and target for your data.
2. SVM Classifier tries to draw a plane based on the neighbouring data instances.
3. This plane is then used to predict values.
4. Evaluate the SVM model.

C. Random Forest

Random forest is a type of ensemble learning method which is used for classification, regression and many other tasks. The main purpose of random forest is creation of decision trees on data samples and obtaining the prediction from each of them and then selecting the best solution by means of voting. It avoids the over-fitting by averaging the results.

Random forest algorithm follows following steps:

We can understand the working of Random forest through following algorithm:

1. Select some sample from the dataset.
2. Now, a decision tree will be constructed for every sample.
3. Prediction will be done on the basis of matrices.
4. Prediction result for every decision tree will be obtained.
5. The best prediction result will be used as a final prediction result.

D. DFD

The following Fig 2 shows the step wise representation of type 2 diabetes prediction system. Initially we got the dataset from the repository. This data is further preprocessed using Python's sklearn library and then segregation is performed by splitting the data into training and testing data. After segregating, LGBM classifier is applied and the parameters are defined.

ISSN: 2233-7857 IJFGCN

The model is now trained for the classification purpose. After the completion of training, the prediction of the output is done in the testing phase which is done separately. For the evaluation purpose accuracy is checked whether it is more than the threshold value.

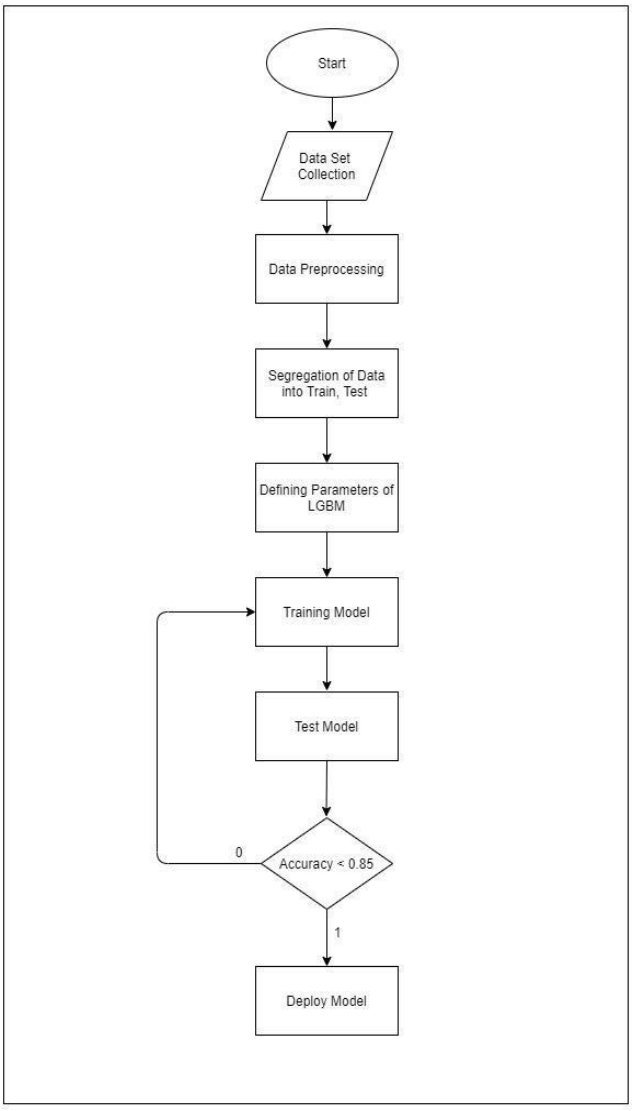

Fig 2: DFD Level

5. Results

The most common type of diabetes among these three is type 2 diabetes. It has severe impact on human health. One way to minimize effects of the disease is by early detection. So detection is very important in diagnosis of this disease. Keeping this as a goal, type II diabetes prediction system was developed. The algorithms were tested on Pima Indians Dataset. The parameters for diabetes prediction were analyzed and selected based on their correlation with type II diabetes.
The table 2 shows the comparisons based on accuracy, precision and F1score.

	Accuracy	**Precision**	**F1score**
LightGBM CV	**0.9062**	**-**	**-**

ISSN: 2233-7857 IJFGCN

LightGBM	0.8802	0.8550	**0.8368**
Random Forest	0.8697	**0.8873**	0.8344
SVM	0.7395	0.75	0.6428

Table 2: Result comparison of algorithms

We have implemented two different variations of LightGBM. The first model of LightGBM is python's API named lightgbm.cv. We have used this model to define the 5 fold cross validation on the train and test dataset. This model got the highest accuracy among all other algorithms that we have implemented. GBDT boosting is used to boost the model.

The second model is the one without the cross validation on the dataset. It gave fairly good accuracy and other metrics. For this model we have used around 50 boosting rounds of GBDT boosting with 15 early stopping rounds.

For Random Forest we have used sklearn library's 'RandomForestClassifier' which has max depth 2 and gini criterion to measure the quality of the split. Sklearn SVC classifier has been used for implementation of the SVM on the PIMA Indian Diabetes dataset. Linear kernel of the classifier was the best kernel since the outcome was binary.

Binary Error – Mean: X axis represents the number rounds that has been done on the dataset and Y axis shows the mean binary error.

1. **Train Data**

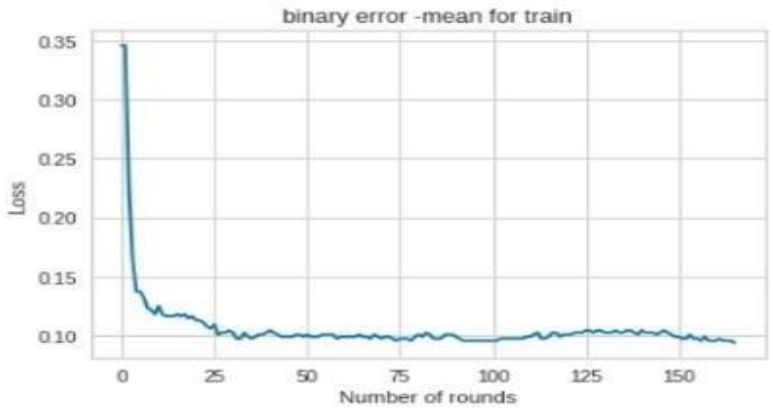

Fig. 3: Graph of mean Binary error for train

From the above Fig 3 it can be observed that while training the model it has made some harsh changes in the decision tree which is what the oscillations in the graph represents. Furthermore it is very clear from the graph that loss of the model is greatly reduced within the first 10 rounds. From this we can infer that some parameter changes made from random initialization turn out to be very significant improvement.

2. **Test Data**

ISSN: 2233-7857 IJFGCN

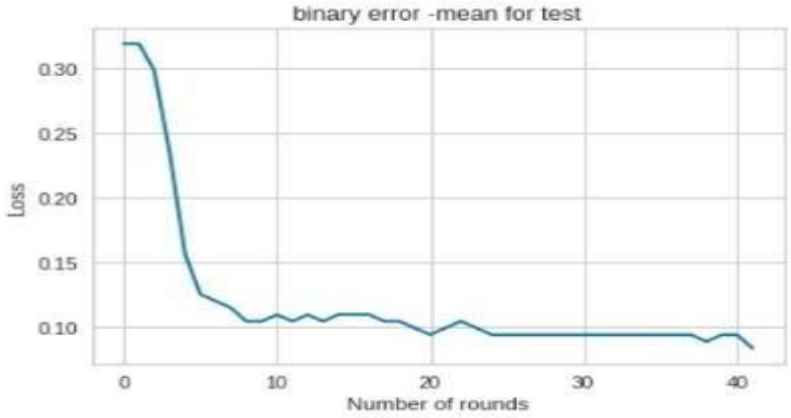

Fig. 4: Graph of mean Binary error for test

Graphs of Both the train and test binary error- mean are somewhat same in pattern since the same model with exact parameters configurations is used to train and test. The sudden drop in the loss can be seen in the above fig.4 before 10 rounds as we have seen while training the model. After about 10 rounds the loss seems to be steadier. This means that loss is no more affected by the parameter change. Hence this is the point where learning on the prediction should be stopped.

6. Conclusion

Currently, there rarely is such a system which will give earlier prediction of this disease. Early disease prediction and identification is critical task in the prevention and control of such chronic diabetic disease leading to saving lives.
Type 2 diabetes prediction is using LightGBM implemented successfully. PIMA Indian dataset was trained for LightGBM, Random Forest and SVM individually. LightGBM CV algorithm gave higher accuracy as compared to Random Forest and SVM. It leads to produce an intelligent, autonomous data collection system to predict Type
2 Diabetes from some parameters and reduce human efforts by automating the Type 2 Diabetes prediction. Hence by using the above approach successfully, prediction of type 2 diabetes was performed and the result was obtained which predicted the diabetes based on the parameters provide by medical organization. Further the accuracy can be improved through newer technologies. If in future study on diabetes reveals some other important feature which could affect performance of model greatly then this feature can be implemented in this model.

References:

[1] International Diabetes Federation, Diabetes Atlas, (9th ed) (2019) Available at www.idf.org

[2] Sanjeev Wagh, Ramjee Prasad, “Energy Optimization in Wireless Sensor Network through Natural Science Computing: A Survey”, 2013.

[3] Swapna G, Soman KP, Vinayakumar R , “Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals”, 2018.

[4] Minyechil Alehegn, Rahul Raghvendra Joshi, “Type II Diabetes Prediction using combo of SVM, ANN and Random Tree”, August 5, 2019.

[5] Kannadasan K, Damodar Reddy Edla, Venkatanareshbabu Kuppili , “Type 2 diabetes data classification using stacked autoencoders in deep neural networks”, 8 December, 2018.

[6] Yang Guo, Guohua Bai,Yan Hu, “ Using Naïve Bayes Network for Prediction of Type-2

ISSN: 2233-7857 IJFGCN

Diabetes",2012.

[7] Peihua Chen and Chuandi Pan, "Diabetes classification model based on boosting algorithms", 2018.

[8] Guolin Ke , Qi Meng , Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye1 , Tie-Yan Liu1, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree" 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[9] Priyadarshini. P, "Prediction of Diabetes Mellitus Using XGBoost Gradient Boosting", Vol- 5, Iss-4, Spl. Issue-2 Dec.-2017.

[10] Standard Dataset of Pima Indian from kaggle https://www.kaggle.com/kumargh/pimaindiansdiabetescsv

[11] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al., "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". The annals of statistics, 28(2):337–407, 2000.

[12] Charles Dubout and François Fleuret, "Boosting with maximum adaptive sampling", in Advances in Neural Information Processing Systems, pages 1332–1340, 2011.

[13] Wei Wang; Yan Huang; Yizhou Wang; Liang Wang, "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction" 25 September 2014.

[14] Ian T. Jolliffe and Jorge Cadima, "Principal component analysis: a review and recent developments"

[15] S. Balakrishnama, A. Ganapathiraju, "LINEAR DISCRIMINANT ANALYSIS - A BRIEF TUTORIAL".

ISSN: 2233-7857 IJFGCN