Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 165 (2019) 292–299

Procedia Computer Science

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019

Diabetes Prediction using Machine Learning Algorithms

Aishwarya Mujumdar[a], Dr. Vaidehi V[b]

[a] *Vellore Institute of Technology, Chennai, India*
[b] *Mother Teresa Women's University, Kodaikanal, India*

Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

Keywords: Diabetes Mellitus; Big Data Analytics; Healthcare; Machine Learning

I.INTRODUCTION

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data.

Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are

*Corresponding author : +91- 8698018735
Email Id-aismu05@gmail.com

10.1016/j.procs.2020.01.047

suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to **629 million**.[1]
Diabetes Mellitus (DM) is classified as-
Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. **Type-2** also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.**Type-3** Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person.

A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction. The paper is organized as follows-
Section II-gives literature review of the work done on diabetes prediction earlier and taxonomy of machine learning algorithms. Section III-presents motivation behind working on this topic. Section IV gives diabetes prediction proposed model is discussed. Section V gives results of experiment followed by Conclusion and References.

II.LITERATURE REVIEW

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization.[4] Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result.[5] K. Rajesh and V. Sangeetha (2012) used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently.[8] Humar Kahramanli and Novruz Allahverdi (2008) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes.[9] B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used.[10] Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour.[7] Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes.[11]
Fig 1, represents taxonomy for Machine Learning Algorithms that can be used for diabetes prediction.
The task of choosing a machine learning algorithm includes feature matching of the data to be learned based on existing approaches. Taxonomy of machine learning algorithms is discussed below-
Machine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised learning, Semi-supervised learning.

Fig1. Taxonomy of Machine Learning Algorithms for Diabetes Prediction

A.The Supervised Learning/Predictive Models

Supervised learning algorithms are used to construct predictive models. A predictive model predicts missing value using other values present in the dataset. Supervised learning algorithm has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to new dataset. Supervised learning includes Decision Tree, Bayesian Method, Artificial Neural Network, Instance based learning, Ensemble Method. These are booming techniques in Machine learning.[3]

B. Unsupervised Learning / Descriptive Models

Descriptive models are developed using unsupervised learning method. In this model we have known set of inputs but output is unknown. Unsupervised learning is mostly used on transactional data. This method includes clustering algorithms like k-Means clustering and k-Medians clustering.[3]

C. Semi-supervised Learning

Semi Supervised learning method uses both labeled and unlabeled data on training dataset. Classification, Regression techniques come under Semi Supervised Learning. Logistic Regression, Linear Regression are examples of regression techniques.[3]

III. MOTIVATION

There has been drastic increase in rate of people suffering from diabetes since a decade. Current human lifestyle is the main reason behind growth in diabetes. In current medical diagnosis method, there can be three different types of errors-

1. The false-negative type in which a patient in reality is already a diabetic patient but test results tell that the person is not having diabetes.
2. The false-positive type. In this type, patient in reality is not a diabetic patient but test reports say that he/she is a diabetic patient.
3. The third type is unclassifiable type in which a system cannot diagnose a given case. This happens due to insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type.

However, in reality, the patient must predict either to be in diabetic category or non-diabetic category. Such errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithm and data mining techniques which will provide accurate results and reduce human efforts.

IV.PROPOSED METHOD

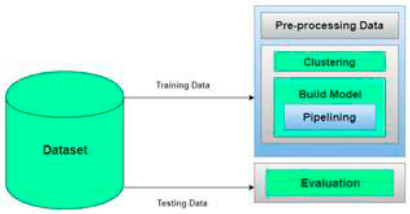


Fig 2. Diabetes Prediction Model

Fig 2., shows architecture diagram for diabetes prediction model. This model has five different modules. These modules include-

i. Dataset Collection
ii. Data Pre-processing
iii. Clustering
iv. Build Model
v. Evaluation

Let's have a look at each model briefly.

i. Dataset Collection

This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. Dataset description is given below-
This Diabetes dataset contains 800 records and 10 attributes.

Table 1. Dataset Information

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Job Type(Office-work/Field-work/Machine-work)	No
Outcome	C

ii.**Data Pre-processing**

This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

iii.**Clustering**

In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our record.

Algorithm:

- Choose the number of clusters(K) and obtain the data points
- Place the centroids c_1, c_2, ... c_k randomly
- Steps 4 and 5 should be repeated until the end of a fixed number of iterations
- For each data point x_i:
  -find the nearest centroid(c_1, c_2 ,..c_k)
  -assign the point to that cluster
- for each cluster j = 1..k
  new centroid = mean of all points assigned to that cluster
- End

iv.Model Building

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier.

Algorithm 1: Diabetes Prediction using various machine learning algorithms

Generate training set and test set randomly.
Specify algorithms that are used in model
mn=[KNN( ), DTC( ), GaussianNB( ),
LDA(),SVC(),LinearSVC(),AdaBoost(), RandomForestClassifier(), Perceptron(),
ExtraTreeClassifier(), Bagging(),
LogisticRegression(),
GradientBoostClassifier()]

```
for(i=0; i<13; i++) do
Model= mn[i];
Model.fit();
model.predict();
print(Accuracy(i),confusion_matrix,  classification_report);
```

End

Algorithm 2: Diabetes Prediction using pipeline

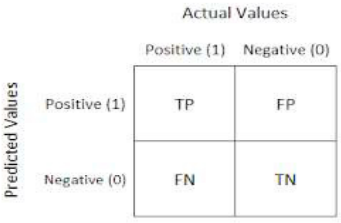
Step1: Import required libraries.
Step2: Import diabetes dataset.
Step3: Create pipeline for algorithms giving highest accuracy.
Step4: Add theses pipeline to a dictionary where all pipelines will be stored.
Step5: Fit the pipelines in training dataset.
Step6: Compare accuracies of all pipelines added.
Step7: Prediction and identification of the most accurate model will be done on test data.
Pipelines work by allowing for a linear sequence of data transforms to be chained together culminating in a modelling process that can be evaluated. The goal is to ensure that all of the steps in the pipeline are constrained to the data available for the evaluation, such as the training dataset or each fold of the cross-validation procedure.

v.Evaluation

This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score.
Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as-

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model.



Where, TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as-

$Accuracy = \frac{TP+FN}{N}$ Where, N:Total number of samples

F1 score-It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

$$F1 = 2 * \frac{1}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right)}$$

F1 Score tries to find the balance between precision and recall.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as-

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: It is the number of correct positive results divided by the number of *all* relevant samples. In mathematical form it is given as-

$$Precision = \frac{TP}{(TP + FN)}$$

V. RESULTS

After applying various Machine Learning Algorithms on dataset we got accuracies as mentioned below. Logistic Regression gives highest accuracy of 96%.

Table 2. Accuracy Table

Algorithms	Accuracy
Decision Tree	86%
Gaussian NB	93%
LDA	94%
SVC	60%
Random Forest	91%
Extra Trees	91%
AdaBoost	93%
Perceptron	76%
Logistic Regression	96%
Gradient Boost Classifier	93%
Bagging	90%
KNN	90%

Confusion Matrix for Logistic Regression is given below-

Table 3. Confusion Matrix for Logistic Regression

	Diabetic	Non-Diabetic
Diabetic	93	5
Non-Diabetic	4	138

The different performance measures that are being compared are Accuracy, F1-Score, Precision and Recall. The Confusion matrix for the algorithm with highest accuracy is mentioned in Table 3.
Visualization of these accuracies helps us to understand variations among them clearly.

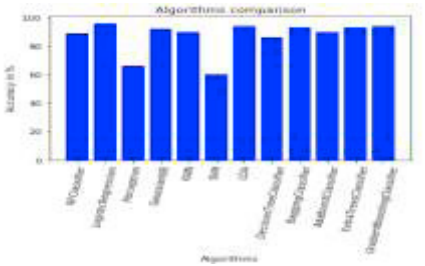


Fig 3.Comparison of various machine learning algorithms based on accuracies

Table 4. Comparison between accuracies of PIMA Diabetes Dataset and Diabetes Dataset used in this paper

Algorithms	Accuracy with PIMA Dataset	Accuracy with Diabetes Dataset used in this paper
Logistic Regression	76%	96%
Gradient Boost Classifier	77%	93%
LDA	77%	94%
AdaBoost Classifier	77%	93%
Extra Trees Classifier	76%	91%
Gaussian NB	67%	93%
Bagging	75%	90%
Random Forest	72%	91%

Decision Tree	74%	86%
Perceptron	67%	76%
SVC	68%	60%
KNN	72%	90%

Result of Algorithm 2:
Using Pipelining, we got highest accuracy of 97.2% for Logistic Regression

Table 5. Pipelining Results

Algorithms	**Accuracy**
AdaBoost Classifier	98.8%
Gradient Boost Classifier	98.1%
Random Forest Classifier	98.1%
Logistic Regression	97.5%
Extra Trees Classifier	96.3%
Linear Discriminant Analysis	95%

CONCLUSION

In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Logistic Regression gives highest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. We have seen comparison of machine learning algorithm accuracies with two different datasets. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely non-diabetic people can have diabetes in next few years.

REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.
[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[9]Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
[10] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
[11] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.