Diabetes Data Analysis and Prediction Model

Nisha N, [2]Amruta Renake, [3]S Y Pattar

[1]Student, [2]Student, [3]Professor
Department of Medical Electronics,
BMS College of Engineering, Bangalore,India

Abstract— **Diabetes Mellitus is one among critical diseases and lots of people are suffering from this disease. Diabetes Mellitus is caused due to age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. 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 andappropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcareindustries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patternsto 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 factorslike Glucose, BMI, Age, Insulin, etc. Classification accuracy is boostedwith new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.**

Index Terms—Diabetes Mellitus, Data Analytics, Prediction model.

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 suffering from it. Around 425 million people suffer from diabetes according to 2017statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to629 million.[1] Diabetes Mellitus (DM) is classified asType-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 injectinsulin 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 complicationsassociated 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 currentand 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 isconsidered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimumflaws. Existing method for diabetes detection is uses lab tests suchas 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 datamining techniques for diabetes prediction.

.

II. LITERATURE REVIEW

The analysis of related work gives results on varioushealthcare 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. Thissystem predicts type of diabetes and alsorisks 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. Patill 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] ManiButwall and Shraddha Kumar (2015) proposed a model using

Random Forest Classifier to forecast diabetes behaviour.[7] Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes.[11].

III. Methodology

A. Data Collection and Analysis

The dataset is collected from Kaggle platform. The proposed work is restricted only for the female dataset. This module includes data collection and understanding the data to study thepatterns and trends which helps in prediction and evaluatingthe results. This Diabetes dataset contains 768 records and 8attributes: Number of Pregnancies, Glucose level, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

B. Data Standardization:

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 selectedattributes 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.

C. Train Test Split:

20% of the data were used to test the model. Other 80% of thedata were used to train the model.

D. Build a Predictive Model:

The proposed method uses support vector machine to design the prediction model. A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After givingan SVM model sets of labeled training data for each category,they're able to categorize new text.

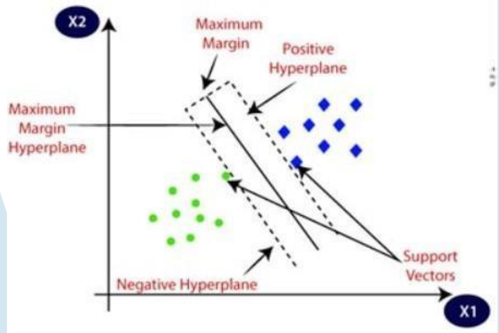


Fig: Support vector machine

The proposed work uses SVM classifier which predictswhether the data shows the subject is diabetic or not.

. Evaluation:

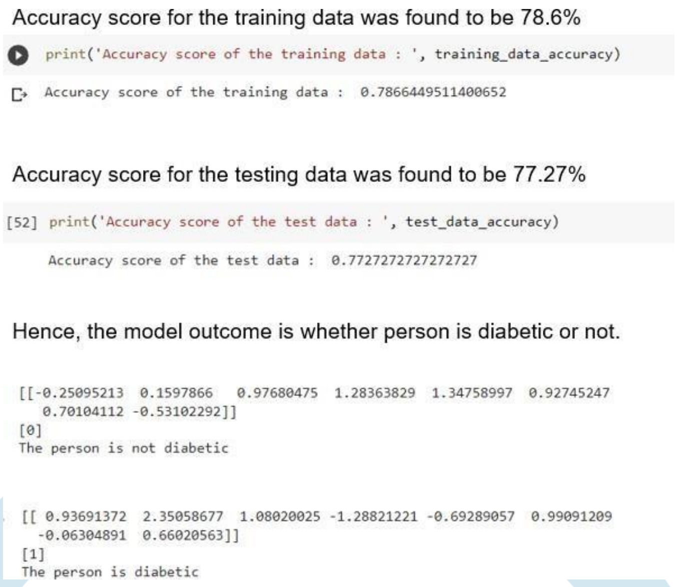
This is the final step of prediction model. Here, we evaluate theprediction results using various evaluation metrics like classification accuracy. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as

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

Table: Referred papers Algorithm and Accuracy

Algorithms	**Accuracy**
Decision Tree	86%
Gaussian NB	93%
LDA	94%
SVC	60%

According to the referred paper, the accuracy they have achieved using SVM classifier is 60%. But, in the proposedwork we have achieved 77.27%.



Accuracy score for the training data was found to be 78.6%

```
print('Accuracy score of the training data : ', training_data_accuracy)

Accuracy score of the training data :  0.7866449511400652
```

Accuracy score for the testing data was found to be 77.27%

```
[52] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.7727272727272727
```

Hence, the model outcome is whether person is diabetic or not.

```
[[-0.25095213  0.1597866   0.97680475  1.28363829  1.34758997  0.92745247
   0.70104112 -0.53102292]]
[0]
The person is not diabetic
```

```
[[ 0.93691372  2.35058677  1.08020025 -1.28821221 -0.69289057  0.99091209
  -0.06304891  0.66020563]]
[1]
The person is diabetic
```

IV. CONCLUSION

In this study, machine learning algorithm is applied on the datasetand the classification has been done using SVM algorithms which gives highest accuracy of 77.27%. It is clear that the modelimproves 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

V. ACKNOWLEDGMENT

Taking immense pleasure in thanking Prof. S.Y. Pattar and also Department of Medical Electronics, for inspiring and giving an opportunity to work on this project. We would like to thank ourinstitute for their support to work on the topic of Image Processing.

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 learning & 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 forthe 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 20.