Diabetes Data Analysis and Prediction Model

Nisha N, 2Amruta Renake, 3S Y Pattar

1Student, 2Student, 3Professor
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 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 factor 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.

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 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 the best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] 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 users 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 datamining techniques for diabetes prediction.

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 risk 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. Naive 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

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 the patterns and trends which helps in prediction and evaluating the results. This Diabetes dataset contains 768 records and 8 attributes: 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 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.

C. Train Test Split:

20% of the data were used to test the model. Other 80% of the data 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 giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.

![Support vector machine](image.png)

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

E. Evaluation:

This is the final step of prediction model. Here, we evaluate the prediction 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

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

Table: Referred papers Algorithm and Accuracy

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>86%</td>
</tr>
<tr>
<td>Gaussian NB</td>
<td>93%</td>
</tr>
<tr>
<td>LDA</td>
<td>94%</td>
</tr>
<tr>
<td>SVC</td>
<td>60%</td>
</tr>
</tbody>
</table>

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

In this study, machine learning algorithm is applied on the dataset and the classification has been done using SVM algorithms which gives highest accuracy of 77.27%. 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.

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 our institute for their support to work on the topic of Image Processing.

REFERENCES