DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

M. Ramakrishna Murthy*1, P. Ruchita*2, Ch. Bharat Teja*3,
M. Manoj Kumar*4, T V S Lingeswararao*5

*1Professor, Department Of Computer Science & Engineering, Anil Neerukonda Institute Of
Technology & Sciences, Visakhapatnam, Andhra Pradesh, India.
*2,3,4,5Students, Department Of Computer Science & Engineering, Anil Neerukonda Institute Of
Technology & Sciences, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

Diabetes is an ineradicable disease that can be found in most of the people nowadays. Due to hectic schedules people are unable to focus on their health. The food we are consuming is fragmented into glucose; these fragments will be delivered into blood. The pancreas releases a hormone named as insulin when the glucose levels are high. This insulin plays a vital role in transporting the glucose to cells that can be used as energy. To maintain a sustainable life detection of diabetes in early stage will be beneficial. Machine learning algorithms will be a productive approach as it will be trained & test with vast data and it enhances itself with upcoming future predictions. In this article, various algorithms like Random Forest, Naive Bayes are used, Decision Tree and trained with our collected dataset. Among these three algorithms it was observed that Random Forest produced accurate results.

Keywords: Machine Learning, Random Forest (RF), Naive Bayes, Decision Tree.

I. INTRODUCTION

Today diabetes is a very common disease. Earlier diabetes was observed in adults and old age. But these days diabetes was reported in teen age also. These are some aspects for developing diabetes such as family history, age, food habits, and high blood pressure, obese.

In general, there are two types of diabetes. In type 1 diabetes production of insulin is less due to this insulin producing cells in pancreas will be affected by immune system. Types 1 diabetes develops due to family history. In the second type of diabetes the body is found to act as a resistant to insulin therefore this results in need of less insulin than required this type of irregularity is found due to procrastination of exercising, unwholesome food and obesity. Type 2 diabetes develops due to obese, family history and inactive lifestyle. It was found that there are multiple risks involved if the diabetes is not controlled / detected in an early stage. During a research result from reputed article, it was found that younger aged people are suffering from type-1 diabetes, women are giving birth to a child that is weighted over 9 pounds, due to unhealthy diet, people are experiencing overweight and obesity, people are also suffering from polycystic overy syndrome etc., all these side effects and diseases are found due to unhealthy diet and unwholesome food.

During development stage of our model various literature review papers were researched and found that traditional approaches were already implemented. So, to implement a model that is unique and futuristic modernised dataset is considered i.e., is a digital medical dataset that consists readings of people eating junk food, average people taking for exercising and the day-to-day lifestyle. This is totally different from traditional approach where a fixed set of values are taken using prediction this results in inaccurate results. Therefore, efficient results were not found in their approach. To overcome this problem including digital dataset, machine learning algorithms like Random Forest, Decision Trees & Naive Bayes are applied. And out of these algorithms Random Forest performed well with our digital dataset and the results were found out be better and efficient.

II. LITERATURE SURVEY

KM Jyothi Rani Proposed a system for predicting diabetes based on Machine learning algorithms. In this paper they have used the dataset which contains 9 features and 2000 entries out of which outcome describes 0 means no diabetes, 1 means diabetes. They have used 5 machine learning algorithms in this paper out of these 5 algorithms Decision Tree algorithm provides training accuracy as 98% and testing accuracy as 99%.
Raja Krishnamoorthi proposed a diabetes healthcare disease prediction framework using machine learning techniques. The dataset contains 768 rows and 9 columns and 90% of the data is used for training and 10% used for the testing purpose and they performed hyper-parameter tuning to evaluate the Machine Learning models and used to increase the accuracy. Out of 5 algorithms best one is identified and hyper parameter tuning has been applied to provide better accuracy as a result of 86%

Desmond Bala Bisandu proposed a system for diabetes prediction using data mining techniques. In this paper there are 5 parameters based on which diabetes is predicted and data is pre-processed to remove noise and to remove null values and classification and prediction was done using Naive Bayes Classifier and efficiency was around 95%

B. Suvarnamukhi proposed a big data processing system which uses machine learning techniques for predicting diabetes. Due to rapid increase in technology the data is stored in the form of electronic records (EHR) and this data is processed using big data and for prediction of diabetes ELM is used and compared with other algorithms and diabetes which is predicted of 3 types

Mitush Soni proposed machine learning algorithms for providing better accuracy in diabetes prediction. In this paper the dataset contains 500 negative outcomes means no diabetes and 268 positive outcomes means diabetes and For Predicting accurately they have used 6 machine learning algorithms and among these 6 algorithms random forest algorithm predicts with 77% accuracy

N. Sneha1 and Tarun Gangil has designed a model for Analysis of diabetes mellitus for early prediction using optimal features selection The dataset consists of 2500 entries and 15 attributes and 768 items used for testing and they have used 5 algorithms out of which support vector machine provides 77% accuracy.

Abdullah A. Aljumah and M.G Ahmad proposed a data mining application to predict diabetes in young and old patients using regression-based mining technique. The dataset is used is a NCD risk factor report from Ministry of health report, Saudi Arabia and using data mining analysis on data set they have predicted the effectiveness in young and old group for different treatments.

Salliah Shafia and Prof. Gufran Ahmad Ansari designed a model for Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach. this research uses the WEKA tools to predict diabetes in patients from Pima India Diabetes Data Set consists of 7 attributes and 767 entries and in this paper, they have used 3 classification algorithms out of which Naïve bayes provides 74% accuracy.

R M Anjana prepared a report on Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India. In this report they conducted a survey on urban and rural parts of india to estimate prevalence of diabetes and prediabetes and in the report, Chandigarh was found to be have highest diabetes percentage.
III. PROPOSED SYSTEM

![Diagram showing the process]

This model helps to predict diabetes with better accuracy. We experimented with different classification algorithms.

1. Dataset Description - The data is gathered from Kaggle website which is named as Diabetes Health Indicators Dataset. It contains 253,679 entries of data and each record consists of 22 columns.

<table>
<thead>
<tr>
<th>S No.</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes_012</td>
</tr>
<tr>
<td>2</td>
<td>HighBp</td>
</tr>
<tr>
<td>3</td>
<td>HighChol</td>
</tr>
<tr>
<td>4</td>
<td>Cholcheck</td>
</tr>
<tr>
<td>5</td>
<td>BMI (Body mass index)</td>
</tr>
<tr>
<td>6</td>
<td>Smoker</td>
</tr>
<tr>
<td>7</td>
<td>Stroke</td>
</tr>
<tr>
<td>8</td>
<td>HeartDiseaseorAttack</td>
</tr>
<tr>
<td>9</td>
<td>PhysActivity</td>
</tr>
<tr>
<td>10</td>
<td>fruits</td>
</tr>
</tbody>
</table>
Correlation Matrix: It is used to demonstrate the relation between the attributes and displays matrix as a output

2. Data Pre-processing -This phase model handles inconsistent data, missing values and other impurities that could cause effectiveness of data.

   Data Pre-processing is done to improve the quality and to obtain accurate results.

   A. Missing values removal - Instances with zero as worth are removed. Through eliminating irrelevant instances, we make feature subset and this process is called features subset selection, which help to work faster.

   B. Splitting of data - After removal of irrelevant instances, data is normalized in training and testing the model. When data is splitted then we train the efficient algorithm on the training data set and keep test data set aside.

   C. Apply Machine Learning – After pre-processing of the data we will split the data into training and testing parts, 80% of the data for training part and 20% of the data for the testing part and now we will train the data using machine learning classification algorithms. These algorithms include Random Forest, Decision Trees, Naïve Bayes. We will train the data using these algorithms and after training the data we will measure the accuracy using test data.
3. **Evaluation** - In this step, we evaluate the prediction results using different performance metrics such as confusion matrix, accuracy, precision, recall, and F1 score.

**Confusion Matrix** – Confusion matrix is used to describe the performance of the Algorithms and gives matrix as an output.

![Confusion Matrix Diagram]

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

**Accuracy** - It is the ratio of number of corrected predictions to by total number of observations. The model is best if it has high accuracy

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

**Precision** – It is the ratio of correctly predicted positive results by the total predictive positive results

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

**Recall** – It is the ratio of correctly predicted positive results by the total results in actual class

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

**F1 Score** – It is weighted average of Precision and Recall.

\[
F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

4. **Save the model** - In this stage, we will compare the accuracies of each model and we will save the model with highest accuracy and we will use that model for predicting the disease.

**IV. EXPERIMENTAL ANALYSIS**

Confusion matrix used to describe the performance of the algorithms and here we will see the confusion matrix for 3 algorithms.

![Confusion Matrix for Naive Bayes Classification Diagram]

*Fig 2: Confusion matrix for Naive Bayes classification*
In this project we have used 3 algorithms and the above table describes the performance metrics of these algorithms and out of those 3 algorithms Random Forest gives better results in terms of accuracy, F1 Score, Recall Score and precision Score.
V. CONCLUSION

We have successfully built a model where it will predict whether a patient has diabetes or not using 3 machine learning algorithms which are Decision Tree classifier, Naïve Bayes and Logistic Regression. Out of these 3 algorithms Logistic Regression gives 84.86% accuracy.

VI. FUTURE WORK

The above model is used to predict whether a person has diabetes or not using their health records and in future we can build a perfect model using deep learning techniques and providing best accuracy and further we can also build a Web application using flask so that users can give the parameters and based on those attributes the model will predict.

VII. REFERENCES


[9] R M Anjana, R pradeep, “Prevalance of Diabetes and prediabetes (impaired fasting and/or impaired glucose tolerance) in urban and rural india, Indian Council of Medical Research, volume 54, pp 3022-3027, 2011.