Time Prediction of Diabetes Mellitus using Machine Learning Classification Algorithm

Veda G. Gaonkar¹, Veena Roshini Pinto², Jeneena Maria Monteiro³, Manasa. M⁴, Dr. V. K. Annapurna⁵
Department of Computer Science and Engineering
The National Institute of Engineering, Mysuru, India

Abstract:
Diabetes Mellitus is one of the most common chronic diseases which not only increases the blood sugar level but also makes an individual vulnerable to other health ailments. In severe cases patients also have a likelihood of developing other health issues like heart ailments, kidney disease, nerve damage etc. Timely prediction of diabetes can help individuals to incorporate a healthy lifestyle to minimize the chances of falling prey to the disease. Machine learning algorithms have proven beneficial in predicting several health conditions more precisely and quickly. Machine learning allows the user to provision a computer algorithm an immense amount of data, have the algorithm analyze and make data-directed recommendations with resolutions based only on the input data. The paper features developing a web application for time prediction of diabetes whose prime users are the doctors and authorized health care personnel. The proposed model uses K-nearest neighbor (KNN) machine learning algorithm for identification and time prediction of diabetes mellitus.

Keywords: KNN, Diabetes Mellitus, Accuracy.

I. INTRODUCTION
Diabetes mellitus (DM) is a chronic condition, with high blood glucose. Primary prevention of diabetes mellitus averts increased risks at later stages. Individuals suffering from diabetes have the risks of acquiring chronic kidney diseases, where the symptoms are imperceptible. Some of the symptoms include, frequent thirst, abnormal loss of weight, loss of appetite and blurry vision. The International Diabetes Federation (IDF) lays out the latest data on DM in the Diabetes Atlas. In 2019, the number of patients suffering with diabetes worldwide was close to 463 million. From the perspective of the population growth, it speculates that the number will reach 700 million, by 2045. With a view to research the high-risk group of DM, it is necessary to harness advanced information technology. Therefore, Machine Learning is an appropriate technology for this problem. Machine learning (ML) is the computer science area that focuses on analyzing and interpreting data patterns and structures to facilitate learning, reasoning, and decision-making outside human interaction. The modeled system is a browser based application, specifically designed for the doctors. It uses machine learning approaches for identification and time prediction of diabetes in a new patient by utilizing previously collected data. The main objective is diabetes identification and time prediction using KNN algorithm, which belongs to the supervised learning domain. The introduced system is implemented using Visual Studio, C# and .NET framework. The work considered for prediction is divided into the following sections; Section II contains the related analysis on the prediction of diabetes mellitus in recent years; Section III offers an overview of the dataset and its relevant definition, followed by Section IV, where the proposed method is evaluated; Section V covers the algorithm model and its summarized description and finally followed by future enhancements and the conclusions drawn from this work.

II. RELATED WORKS
This section provides some of the related research works carried on over the past several years.

Deepti Sisodia [1], proposed the comparative analysis of diabetes based on accuracy. Experimental Analysis was performed on Pima Indians Diabetes Database. Algorithms used were Naïve Bayes, SVM (Support Vector Machine) and Decision Tree. Naive Bayes algorithm was considered as the finest supervised machine learning method of this research because it gave higher accuracy with regards to other classification algorithms with an accuracy of 76.30%. Shi et al. considered that preventing T2DM (Type 2 Diabetes Mellitus) should be oriented toward individuals. Therefore, they emphasized on setting up a diabetes risk assessment model and drafted a diabetes risk score system based on mobile devices.Han Wu [2], proposed the prediction model for diabetes using classification algorithms. Pre-processing, associating, clustering, classifying, and the visual interface were carried out using Weka toolkit. The logistic regression algorithm was used to categorize the remaining data. The accuracy of this prediction model was upto 95.42%. D. Vigneswari [3], proposed the prediction model for diabetes mellitus using 5 classifiers of Machine Learning such as Random tree, RepTree, random forest, Logistic Model Tree (LMT),C4.5. In this study, the LMT classifier achieved an accuracy of 79.31% with an average time of 0.49s to build the model. M. Durgadevi [4], used the following classification algorithms namely RBF (Radial Basis Function) network, ant-miner, Bagging and Adaboost for the prediction of diabetes mellitus. These Machine Learning classifiers have been tested with 3 sets of type II diabetes datasets obtained from the UCI (University of California, Irvine) machine learning repository in terms of specificity, F-score, accuracy, kappa,
sensitivity and chance agreement. Ant-Miner has achieved a Kappa value of 0.982 for the Pima dataset.

III. DATASET DESCRIPTION

A data set consists of data corresponding to the executive summary of a single database table or a single statistical data matrix in which each table column represents a specific variable and each row corresponds to a given data set member.

Training dataset—The perceptions in the training set structures the experience that the algorithm uses to learn and whenever required, to construct a model. In supervised learning problems, every perception comprises a watched yield variable and at least one watched input factors.

Testing dataset—The test set involves perceptions used to assess the presentation of the model utilizing some exhibition metric.

A. Dataset parameters

The dataset constitutes records collected from sources such as UCI machine learning repository, Kaggle and Datahub. 21 parameters are taken into consideration which include age, gender(M/F), duration of diabetes (DOD), diabetes diagnosed (DD), sugar tested value, last eye examination, family history of diabetes, diabetes treatment, diabetes treated from, past smoked, smoke per day, started smoking, drinking, weight, height, abdominal circumference, BP (Blood Pressure) systolic, BP diastolic.

IV. PROPOSED SYSTEM

This section elaborates on the modeled system, which is composed of frontend modules and databases. The modeled system is a medical application used by Endocrinologists. Machine learning classification algorithm is applied on old patients’ dataset and a model is created for recognizable proof and early prediction of diabetes in another patient. The objective is to predict diabetes dependent on the properties, such as age, weight, blood test report, BP and so on. The modeled system is a real-time application which uses Visual Studio and SQL (Structured Query Language) Server as front-end and back-end technologies respectively.

A. Architecture

The system architecture is illustrated in the Fig.1. The work aims at designing a real-time web application which can be accessed by the medical practitioners for predicting the disease. The confirmation of diabetes disease and time prediction is done using KNN algorithm. The doctors or authorized health care workers need to register on the application and upload patient details for the prediction. As depicted in Fig 1, the outcome is inferred based on the presence of the disease. Further, if the outcome is negative, the duration of occurrence of the disease is provided.

B. Modules

a) Index Module

The Index page is the starting page which is viewed by users when he/she first opens the web application. It is designed to give an overview of the purpose of the web application.

b) Admin Module

The admin can login through the admin login module. The job of the admin is to monitor the new users registering on the site, maintenance and also has the option of deleting existing users if necessary.

c) Register Module

New users (doctors) can register/signup into the application using this module. It asks the basic details like the user ID, phone number, mail ID etc. Any new user first needs to register in order to login.

d) User Module

Registered members (doctors) can login with the valid id and password. The members/users can view the training dataset used for prediction. The members can then perform the prediction for patients by entering the various parameters that include patient details like name, age, gender etc. as indicated in the page. If the individual is classified as ‘YES’ i.e. he/she is suffering from diabetes, then no further analysis is done. On the other hand, if the patient is classified as a ‘NO’ i.e. he/she is not suffering from diabetes, then time prediction is carried out. The members can view the bulk prediction of multiple patients using the result analysis tab. The prediction results are obtained using KNN algorithm for the testing dataset. The Result analysis tab gives the information of accuracy and time taken for the calculation by KNN. Discussion forum enables members to post various queries and topics related to the disease or for that matter any medical queries. The posted comments, topics are visible to all the registered members.

C. Database

The database that is created in the backend using SQL Server is the Medical database which includes the following tables:

- Admin: AdminTable {adminId, password} is used to store the admin id and password.
- Comments: Comments Table {CommentId, TopicId, Member Id, Comment, PostDate} is used to store the comments data in the discussion forum.
- Members: Members Table {memberId, password, Name, Mobile, EmailId, RegisteredDate} is used to store the data of doctors.

Figure.1. System Architecture
V. ALGORITHM MODEL

The K-nearest neighbors (KNN) algorithm is a basic, simple to-execute supervised machine learning algorithm that can be utilized to take care of both regression and classification problems. It is broadly expendable, in actuality, since it is non-parametric and consequently doesn't make any basic suppositions about the dissemination of information.

KNN algorithm stores all the accessible information and characterizes another information point dependent on the likeness. Further, when new information shows up, it very well may be handily ordered into a well suit category by utilizing KNN algorithm. The algorithm at the training stage just stores the dataset and when it gets new information, it groups that information into a classification that is a lot like the new information. In this work, a dataset with 21 parameters is considered. A total of 400 records are taken and divided into testing and training dataset. The training and testing datasets consist of 350 and 50 records respectively. Each record of testing dataset is evaluated against every record of training dataset by applying the Euclidean distance formula for all parameters. The outcome is either positive or negative for the expectation of diabetes illness. If the outcome is negative, the time of getting diabetes is predicted as after 6 months or 1 year or 2 years or 3 years.

The working of KNN is depicted in Fig. 2. One of the challenging parts of KNN is deciding the value of K.

• An extremely low incentive for K, for example, K=1 or K=2, can be noisy and lead with the impacts of anomalies in the model.

• Enormous values for K are acceptable, however it might lead to few troubles.

In this work, initially K value was equal to 10 and after applying the algorithm till K=35, it was concluded that K=25 was the most optimal one and gave the least number of incorrectly classified data.

The metric used to find the nearest neighbor is Euclidean’s Distance formula

\[ h \approx \sqrt{(i_1 - i_2)^2 + (g_1 - g_2)^2} \]

Here, 
\( h \) is the distance between current testing data and training data, \( i_1, i_2, g_1, g_2 \) being parameters.

Calculating the accuracy:

\[ accuracy \approx \left( \frac{outcome\text{count}}{actual\text{count}} \right) \times 100 \]

Here, 
Outcome count is number of records correctly classified. Actual count is the number of records in the testing dataset. Time taken by KNN algorithm is found by making use of session [output] variable present in c#. The first argument of the session [output] gives the time.

VI. EXPERIMENTAL ANALYSIS AND OUTCOMES

The Table 1 below depicts the cases and inferred results considered while developing the application

<table>
<thead>
<tr>
<th>ID</th>
<th>CASES</th>
<th>EXPECTED OUTCOME</th>
<th>INFERRED RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>PREDICTION</td>
<td>The members are able to enter the patient details and prediction results are shown on the screen</td>
<td>Prediction results are displayed. If the patient is classified as a ‘NO’, then the time is predicted</td>
</tr>
<tr>
<td>2.</td>
<td>RESULT ANALYSIS</td>
<td>The members should be able to view the bulk prediction results of multiple patients</td>
<td>Bulk prediction is displayed in the form of a table</td>
</tr>
<tr>
<td>3.</td>
<td>ACCURACY</td>
<td>Members should be able to view the accuracy obtained and time taken by the algorithm to calculate the result</td>
<td>Accuracy and time taken are displayed in the form of a table</td>
</tr>
</tbody>
</table>
VII. CONCLUSION AND FUTURE WORKS

Machine Learning classification algorithm was analysed for the identification and early prediction of diabetes mellitus. In this work, K- Nearest Neighbours algorithm gave an efficiency of 98%. Machine learning has played a great role in revolutionizing the prediction of diabetes with the help of advanced computational methods and availability of diabetes risk dataset. The scope of the current system can be enhanced by using various data visualization techniques for better visual analysis of the prediction. In future, machine learning algorithms can be enhanced and made to run on real time patient data in hospitals, by rigorously and continuously training it, in order to infer better results.

VIII. REFERENCES

[1]. Deepti Sisodiaa, Dilip Singh Sisodia; “Prediction of Diabetes using Classification Algorithms”, International Conference on Computational Intelligence and Data Science (ICCIDS 2018)


[4]. M. Durgadevi,” Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes”, International Conference on Advanced Computing (ICoAC) Pages 84-86


