Abstract: Diabetes is one of the deadliest diseases which causes an increase in blood sugar. Many problems can occur if diabetes remains undetected and unidentified. The lengthy identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this difficult problem. The true objective of this study is to design a model which can determine the likelihood of diabetes in patients with maximum accuracy. Therefore, we are using machine learning classification algorithm, namely Decision Tree, in this experiment to detect diabetes at an early stage. However, early determination of diabetes is quite challenging process for medical practicing people due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot, etc. Data science methods have the ability to benefit other scientific fields by finding new light on common questions. One such benefit is to help make predictions on medical data. Machine learning is an emerging scientific field in data science to deal with the ways in which machines learn from their experience.

Index Terms: Smart Health Prediction System, Clinical Predictions, Machine Learning, Decision Tree, Predictive Analysis, Health Care.

I. INTRODUCTION

Diabetes poses a great threat to human health and body. The characteristic of diabetes is that the blood glucose is either higher than the normal level, which is caused by defects or problems in insulin secretion or its improper biological effects, or both [3]. Diabetes can lead to a disastrous damage and non-functioning of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Diabetes is divided into two categories, which includes type 1 diabetes, T1D and type 2 diabetes, T2D [4]. Patients with T1D diabetes are younger, mostly less than 30 years old. The typical symptoms include increased thirst, frequent urination and high blood glucose levels. This kind of diabetes cannot be cured efficiently with oral medications alone and the patients require insulin therapy, T2D, type 2 diabetes occurs more commonly in middle-aged and elderly people [8]. This kind of diabetes cannot be cured efficiently with oral medications alone and the patients require insulin therapy. T2D, type 2 diabetes occurs more commonly in middle-aged and elderly people, that is more often associated with the occurrence of excess weight gaining, extreme tension, thickening of the walls of arteries, and other diseases. Rise over the years is due to type 2 diabetes, major factors are obesity and overweight among adults [5].

In medicine, the diagnosis of diabetes is done by fasting the blood glucose, tolerance of glucose, and randomly taken blood glucose levels. We can control it as much easily as it is diagnosed earlier. Machine learning also helps people to make a judgment about diabetes according to their data calculated by the examination, which may function a reference for doctors. For machine learning method, the most important problems are how to select the valid features and the correct classifier. Recently, numerous algorithms are used to predict diabetes, such as support vector machine (SVM), decision tree (DT), logistic regression and so on. In paper [6] author Asma A. AlJarullah studied the data set clearly and followed two step pre-processing, first step was focused on attribute selection, some irrelevant attributes with missing values were removed and in second step numerical discretization is applied on data obtained from step 1, now this processed data is given to decision tree classifier and achieved 78.17% accuracy. Machine learning methods are widely used in diabetes prediction, and they get good results. Decision tree is one of the popular machine learning methods in medical field, which has great classification power. So, in this study, we used decision tree algorithm to predict the diabetes.

II. METHODS AND MATERIAL

1. Existing System

A major challenge facing the health care organizations like hospitals, medical centers is the provision of quality services at an affordable cost. Quality service consist of diagnosing the patients correctly and administering the treatments that are effective. Poor clinical decisions can lead to the harmful consequences [6], which, therefore cannot be accepted. Hospitals should also reduce the cost of clinical tests [7]. Most hospitals today employ some sort of the information systems to manage their health care or patient’s data. Unfortunately, this data is rarely used to support the clinical decisions. There is a wealth of hidden information in this data that is largely unused. This raises an important question that how to use this data to convert it into the useful information that can enable the health care practitioners to make intelligent clinical decisions. This is the main motive for this project. Everyone can be a patient at some time or another, and we all need good medical care. We assume that doctors are medical experts and that they do a good research for all their decisions. However, that is not the case always. Nevertheless, they cannot possibly commit to memory all the knowledge they need for every situation, and they probably do not have it readily available. Disadvantage of an existing system would be that the patients have to visit the doctor in person and still does not get proper treatment, as the doctors are unable to predict the exact disease. Human error is possible to be avoided using the computer assisted quality decisions. It is poor when there is large amount of data to be classified. In addition, accuracy of decisions decreases when the humans are put into stress and the immense work.

1Shubham Salunke, 2Shubham Rajiwaide, 3Deepak Yadav, 4S.K. Sabnis

123 UG Student, 4Mentor, Information Technology,
Rajiv Gandhi Institute of Technology, Mumbai, India
2. Proposed System

Patient ID (PID) data set of National Institute of Diabetes and Digestive and Kidney Diseases. PID is comprised of 768 instances. Eight numerical attributes IS representing for each patient in data set[9].

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of times pregnant</td>
</tr>
<tr>
<td>2</td>
<td>Plasma glucose concentration in an oral glucose tolerance test</td>
</tr>
<tr>
<td>3</td>
<td>Diastolic blood pressure (mm/Hg)</td>
</tr>
<tr>
<td>4</td>
<td>Triceps skin fold thickness (mm)</td>
</tr>
<tr>
<td>5</td>
<td>2-hour serum insulin (µU/ml)</td>
</tr>
<tr>
<td>6</td>
<td>Body mass index (kg/m²)</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes Pedigree function</td>
</tr>
<tr>
<td>8</td>
<td>Age (years)</td>
</tr>
</tbody>
</table>

Table 1: PID Data set.

The proposed diabetes prediction system has two important stages that work together to achieve the desired results. The 1st stage of the proposed system is the data preparation, and the 2nd stage is the classification. However, the input into the system is the PID data set and the output will represent the healthy or diabetic individual[10].

3. Working of The System

- According to the figure, it is a two-tier architecture. We provide a form that would contain input field which are to be filled as per the data specified in the patient report. The input field only takes numerical input.
- On the basis of input given by the patient the diabetes would be predicted. If the patient then a list of food items is shown that would help patient to maintain their diabetic level.
- On the basis of information given by the patient the query is generated and the database response to the query.

![Figure 1: Main Stages of Proposed System](image)

![Figure 2: System Working](image)
4. Algorithm Used

ID3 Machine Learning Algorithm

Iterative Dichotomizer 3 is also called as ID3 algorithm, It is a classification algorithm that follows a ravening attitude of constructing a decision tree by selecting a best property that yields maximum IG (Information Gain) or minimum H (Entropy).

Entropy is a measure of the amount of unreliability in the data set S. Mathematical Representation of Entropy is shown below

\[ H(S) = -\sum_{c \in C} p(c) \log_2 p(c) \]

In ID3, entropy is calculated for each residual attribute. The attribute with the smallest entropy is used to split the set S on that particular iteration.

Entropy = 0 implies it is of pure class, that means all are of same category.

Information Gain IG(A) tells us how much unreliability in S was lessen after splitting set S on attribute A. Mathematical representation of Information gain is shown below

\[ IG(A,S) = H(S) - \sum_{t \in T} p(t) H(t) \]

Figure 3: Flowchart of ID3 Algorithm
5. Technologies Used
   a) .NET Framework:
      The .NET Framework is Managed by Microsoft’s Code programming model for building the applications on servers, windows clients and mobiles or IOT devices. It is a software technology that is available with several Windows OS.
   b) C# framework:
      C# is a professional programming language. C# is a programming language designed for building a wide range of organizational applications that run on the .NET Framework.
      C# focuses on providing a safe, modern, simple, high-performance oriented system .NET development.
   c) Microsoft SQL Server 2008:
      It provides an enterprise database platform for management that your organization needs to adapt quickly in a fast-changing environment.

III. RESULTS AND DISCUSSIONS

**Figure 4: Patient Registration**

**Patient Registration:** if patient is a new user, he will enter his personal details, he will put a user id, and password through which he can login to the system.

**Figure 5: Patient login**
**Patient Login:** - patient login to the system using his own login id and password.

**Homepage:** patient gets here numerous options by which he can add his parameters to check the diabetes probability.

---

**Figure 6: Disease Prediction**

**Diabetes Prediction:** - patient will enter the parameters for the analysis as per the calculated report which then used by applying the algorithm to check the probability of diabetes.

---

**Figure 7: Process Outcome**

**Process Outcome:** - based on the added parameters by the patient, the system will apply decision tree algorithm to check the severity of the result if it is yes or no.
IV. FUTURE SCOPE

Hidden knowledge will be extracted from the historical data in the proposed system, by preparing data sets by applying the decision tree algorithm. Predicting the diabetes probability can be done only if the system responds that way. These data sets will be compared with the incoming queries and the final report will be generated using the decision tree algorithm. More work can be done in the future by using more data set related to diabetes disease by using different data reduction methods and machine learning algorithms to improve the classification and get the accurate results. For better accuracy and prediction of diabetes disease, the data sets that will be used must be properly oriented with good quality and free from any faults, errors, inconsistencies and the missing values. This web based application can be further enhanced into an Android app. This will be made available to the users on mobile phones and then its use can be further increased accordingly. Also, features like getting the doctor online on chat so that patients can directly talk to the concerned doctors, we can also provide the options for patients to contact the nearest best doctor in case of positive results. This will make this web application predictable in true sense.

V. CONCLUSION

Machine learning has the great ability to improve the prediction diabetes risk with the help of advanced methods of computation and availability of the large amount of the diabetes risk data set. Detection of diabetes in its early stages is the key element for treatment. This work describes a use of machine learning approach to predict probability of diabetes. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status. A fast and accurate diabetes prediction system is being used in this paper. The designed system uses a numerous instance within about 8 attributes for each one of the PID data set. The used data is preprocessed in order to remove the unwanted data, and leads to the faster processing time. Moreover, the process of dividing of the data set into the subsets using the ID3 algorithm, made an optimal classification result. The proposed system focuses on the features, analysis of the parameters and the classification parts. The propositions of these parts leads to an optimal solution. The results of experiments illustrated the effects of using the algorithms of the proposed system through achieving a higher classification rate that the other systems.

VI. REFERENCES