International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

# Implementation of Data Mining Algorithms for Diabetes Prediction

**Shradhda S. Chindage, Rohini M. Rajmane, Shravani S. Shinde, Shweta S. Gundale, Uday B. Mane(Asst. Prof.)**
Computer Science & Engineering, Shivaji University/Sanjay Ghodawat Institute, Atigre/ Kolhapur, Maharashtra, India

## ABSTRACT

The process of analyzing different aspects of data and aggregating it into useful information is called data mining. The goal is to provide meaningful and useful information for the users about the diabetes. With the rise of information technology and its continued advent into the medical and healthcare sector, the cases of diabetes as well as their symptoms are well documented. This research project aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. The monitoring module analyzes the laboratory test reports of the blood sugar levels of the patient and provides proper awareness messages to the patient through mail and bar chart.

**Keywords:** Classification, Data Mining, Decision Tree, Diabetes and Naïve Bayes.

## I. INTRODUCTION

Effects of diabetes have been reported to have a more fatal and worsening impact on women than on men because of their lower survival rate and poorer quality of life. WHO reports state that almost one – third of the women who suffer from diabetes have no knowledge about it [4]. The effect of diabetes is unique in case of mothers because the disease is transmitted to their unborn children. Strokes, miscarriages, blindness, kidney failure and amputations are just some of the complications that arise from this disease. The analyses of diabetes cases have been restricted to pregnant women.

In this paper, Decision Tree and Naïve Bayes algorithm has been implementing on a pre-existential dataset to predict whether diabetes is recorded or not in a patient. Results from both the algorithms have been compared and presented. Several other models have been formulated over the years that are used for diabetes prediction.

Nowadays, large amount of information is collected in the form of patient records by the hospitals. Knowledge discovery for predictive purposes is done through data mining, which is analysis technique that helps in proposing inferences. This method helps in decision-making through algorithms from large amounts of data generated by these medical centres. Considering the importance of early medical diagnosis of this disease, data mining techniques can be applied to help the women in detection of diabetes at an early stage, which may help in avoiding complications.

### 1.1 Overview of Diabetes

#### 1.1.1 Diabetes

Diabetes is a disease that occurs when the insulin production in the body is inadequate or the body is unable to use the produced insulin in a proper manner, as a result, this leads to high blood glucose. The body cells break down the food into glucose and this glucose needs to be transported to all the cells of the body. The insulin is the hormone that directs the

glucose that is produced by breaking down the food into the body cells. Any change in the production of insulin leads to an increase in the blood sugar levels and this can lead to damage to the tissues and failure of the organs.

Generally, a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol/L). There are three main types [13] of diabetes, viz. Type 1, Type 2 and Gestational.

### 1.1.2. Types of Diabetes

The two main types of diabetes are described below:
Type 1 – Though there are only about 10% of diabetes patients have this form of diabetes, recently, there has been a rise in the number of cases of this type in the United States. The disease manifest as an autoimmune disease occurring at a very young age of below 20 years hence also called juvenile-onset diabetes. In this type of diabetes, the pancreatic cells that produce insulin have been destroyed by the defence system of the body. Injections of insulin along with frequent blood tests and dietary restrictions have to be followed by patients suffering from Type 1 diabetes.

Type 2 – This type accounts for almost 90% of the diabetes cases and commonly called the adult-onset diabetes or the non-insulin dependent diabetes. In this case, the various organs of the body become insulin resistant, and this increases the demand for insulin. At this point, pancreas does not make the required amount of insulin. To keep this type of diabetes at bay, the patients have to follow a strict diet, exercise routine and keep track of the blood glucose. Obesity, being overweight, being physically inactive can lead to type 2 diabetes. Also with ageing, the risk of developing diabetes is considered more. Majority of the Type 2 diabetes patients have borderline diabetes or the Pre-Diabetes, a condition where the blood glucose levels are higher than normal but not as high as a diabetic patient.

### 1.1.3. Symptoms, Diagnosis and Treatment

The common symptoms of a person suffering from diabetes are:

- ✓ Polyuria (frequent urination)
- ✓ Polyphagia (excessive hunger)
- ✓ Polydipsia (excessive thirst)
- ✓ Weight gain or strange weight loss
- ✓ Healing of wounds is not quick, blurred vision, fatigue, itchy skin, etc.

Urine test and blood tests are conducted to detect diabetes by checking for excess body glucose. The commonly conducted tests for determining whether a person has diabetes or not are

- ✓ A1C Test
- ✓ Fasting Plasma Glucose (FPG) Test
- ✓ Oral Glucose Tolerance Test (OGTT).

## II. RELATED WORK

Design of prediction models for diabetes diagnosis has been an active research area for the past decade. Most of the models found in literature are based on clustering algorithms and artificial neural networks (ANNs).

A study conducted in [6] intended to discover the hidden knowledge from a particular dataset to improve the quality of health care for diabetic patients. In [7] Fuzzy Ant Colony Optimization (ACO) was used on the Pima Indian Diabetes dataset to find Set of rules for the diabetes diagnosis.

The paper [8] approached the aim of diagnoses by using ANNs and demonstrated the need for pre-processing and replacing missing values in the dataset being considered. Through the modified training set, a for training the set. Finally, in [9] a neural network model for prediction of diabetes based on 13 early symptoms of the disease was created with implementation using MATLAB.

However, nobody established a classification model based on probability and feature selection. Install the related work analytical techniques have been employed to produce reliable results but generally, the methods are time consuming since most employed a weighted approach.

Hence, there is a requirement of a model that can be developed easily providing reliable, faster and cost effective methods to provide information of the probability of a patient to have diabetes. In the present work, an attempt is made to analyse the diabetes parameters and to establish a probabilistic relation between them using Naïve Bayes and Decision Tree approach. For the purpose of analysis, the models are tested depending on the percentage of correctly classified instances in the dataset.

The paper [14] study includes the characteristics of diabetes and to find the number of people suffering from diabetes. This process is performed by considering the diabetic population of 249 instance and 7 unique attributes. The dataset of 249 instances are applied to WEKA tool and performed on algorithms such as Bayes network classifier, J48 Pruned tree, REP tree and Random forest. This survey was done to create awareness about the increasing population of diabetes among people all over the world and helps in knowing the status of the disease.

## III. PROPOSED WORK

As Existing, system works only on 15 attributes, which are the laboratory dataset of 150 peoples only, and one attribute at a time can processed is the main Drawback of Existing system, which we overcome through our new proposed system in "Diabetes Prediction and Monitoring Tool".

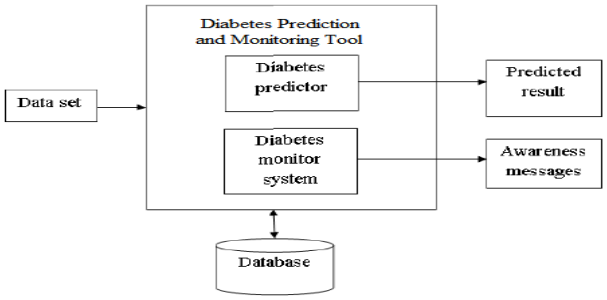


**Figure 1.** System Architecture

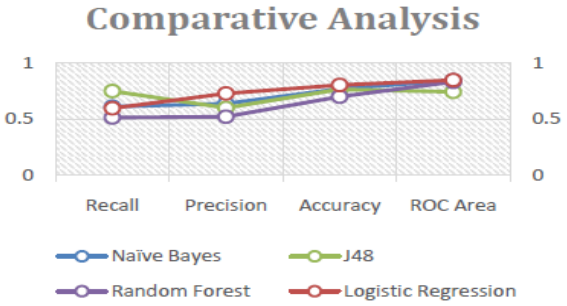


**Figure 2.** Comparative analysis of algorithm in terms of accuracy

In our proposed system, we are going to implement the J48 and Naïve Bayes algorithm. Comparing results of both the algorithm and give the proper prediction result to the end user. Prediction is on the medical report data and these reports are listed below.

- ✓ Blood Sugar Test: - Finds Your Plasma Level Concentration, can be done while fasting and within 2 hour after having meal.
- ✓ Blood Pressure Test: - Finds your Blood Pressure i.e. systolic and diastolic pressure.
- ✓ Skin Thickness: - Finds your Body composition and thickness of skin.
- ✓ Insulin Test: - Finds Insulin in your blood to find level of insulin resistance.
- ✓ Pima Test: - Finds how likely you are to have diabetes considering the here dietary traits of diabetes.

### 3.1 Objectives

- ✓ The present work is intended to meet the following objectives:
- ✓ Present a Decision Tree and Naïve Bayes model for diabetes prediction in pregnant women.
- ✓ Summarize Diabetes – types, risk factors, symptoms and diagnosis.

- ✓ To evaluate various parameters for performance evaluation of purposed system.

### 3.2 Overview of Methodologies

The present work intends to create a mining model based on two classification algorithms in order to provide a simpler solution to the problem of diagnosis of diabetes in women. The results have been analysed using statistical methods and are presented in the Section 3.2.1 and Section 3.2.2

#### 3.2.1 Decision Trees

Decision tree [3] is a tree structure, which is in the form of a flowchart. It is used as a method for classification and prediction with representation using nodes and internodes. The root and internal nodes are the test cases that are used to separate the instances with different features. Internal nodes themselves are the result of attribute test cases. Leaf nodes denote the class variable.

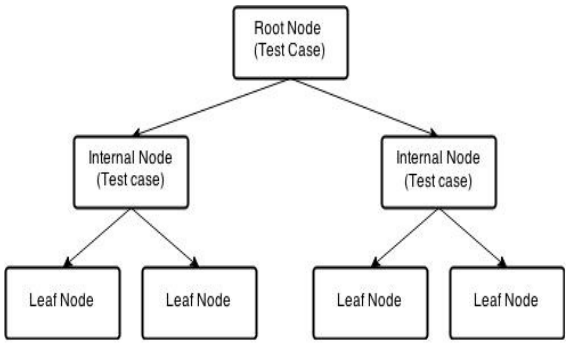


**Figure 3.** Sample Decision Tree Structure

Decision tree provides a powerful technique for classification and prediction in Diabetes diagnosis problem. Various decision tree algorithms are available to classify the data, includingID3, C4.5, C5, J48, CART and CHAID. In this paper, J48 decision tree algorithm [10] has been chosen to establish the model. Each node for the decision tree is found by calculating the highest information gain for all attributes and if a specific, attribute gives an unambiguous end product (explicit classification of class attribute), the branch of this attribute is terminated and target value is assigned to it.

#### 3.2.2 Naïve Bayes

The Naïve Bayes Algorithm is a probabilistic algorithm that is sequential in nature, following steps of execution, classification, estimation and prediction. For finding relations between the diseases, symptoms and medications, there are various data mining existing solution, but these algorithms have their own limitations; numerous iterations, binning of the continuous arguments, high computational time, etc. Naïve Bayes overcomes various limitations including omission of complex iterative estimations of the parameter and can be applied on a large dataset in real time. The algorithm works on the simple Naïve Bayes formula shown in Figure 3.

$$\text{Posterior Probability } P(c|x) = \frac{\text{Likelihood } P(x|c) \times \text{Class Prior Probability } P(c)}{\text{Predictor Prior Probability } P(x)}$$

**Figure 4.** Naïve Bayes Formula

## IV. METHODOLOGIES

### 4.1 Dataset Description and Pre-Processing

The paper explores the aspect of Decision Tree and Naïve Bayes Classifier as Data Mining techniques in determining diabetes in women. The main objective is to forecast if the patient has been affected by diabetes using the data mining tools by using the medical data available.

The classification type of data mining has been applied to the Pima Indians Diabetes Database of National Institute of Diabetes, Digestive, and Kidney Disease. Table 1 shows a brief description of the dataset that is being considered.

**Table 1.** Dataset Description

| Dataset | No. of attributes | No. of Instances |
|---|---|---|
| Pima Indians Diabetes Database of National Institute Diabetes and Digestive and Kidney Diseases | 8 | 768 |

The attributes descriptions are shown in Table 2 below.

**Table 2.** Attribute Description.

| Attribute | Relabeled values |
|---|---|
| 1. Number of times pregnant | Preg |
| 2. Plasma glucose concentration | Plas |
| 3. Diastolic blood pressure (mm Hg) | Pres |
| 4. Triceps skin fold thickness (mm) | Skin |
| 5. 2-Hour serum insulin | Insu |
| 6. Body mass index (kg/m$^2$) | Mass |
| 7. Diabetes pedigree function | Pedi |
| 8. Age (years) | Age |
| 9. Class Variable (0 or 1) | Class |

Pre-processing and transformation of the dataset are done using WEKA tools [11].

The latter makes it easy to use the dataset, as the range of the variable is restricted from 0 to 1. Feature selection has been employed using the CfsSubsetEval class, and the attributes obtained after execution are as follows:

1. Plasma glucose concentration
2. Body mass index (kg/m2)
3. Diabetes pedigree function
4. Age (years)
5. Class Variable (nominal) - Determines if the person has diabetes or not

The descriptive statistics of the dataset are presented in Table 3. Since the parameters are normalized the range of all are in the range 0 to 1.

**Table 3.** Descriptive Statistics of Transformed Dataset

| Parameter | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Plas | 0 | 1 | 0.608 | 0.161 |
| Mass | 0 | 1 | 0.477 | 0.117 |
| Pedi | 0 | 1 | 0.168 | 0.141 |
| Age | 0 | 1 | 0.204 | 0.196 |

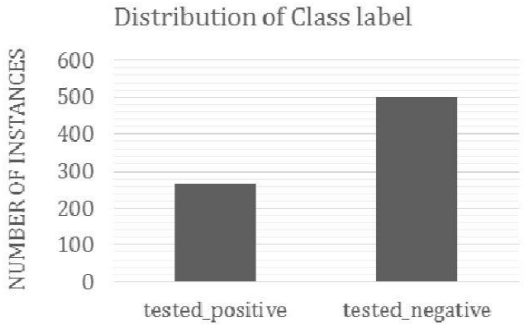


**Figure 5.** Class Attribute Distribution

### 4.2 Proposed Data Model

In this paper two algorithms namely, J48 (decision tree algorithm) and Naïve Bayes, have been used to create the model for diagnosis. The data was divided into training set and test set by the cross-validation technique and percentage split technique.

10-fold cross validation is used to prepare training and test data. After data pre-processing (CSV format), the J48 algorithm is employed on the dataset using WEKA (Java Toolkit for various data mining technique) after which data are divided into "tested-positive" or "tested-negative" depending on the final result of the decision tree that is constructed. The algorithm for conducting the procedure is as follows:

**ALGORITHM:** DIABETES_ALGO
**INPUT:** Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and
Kidney Diseases dataset pre-processed in CSV format.
**OUTPUT:** J48 Decision Tree Predictive Model with leaf node either tested-positive or tested negative and Naïve Bayes Prediction Results.

**PROCEDURE:**

1. The dataset is pre-processed using WEKA tools. Following operations are performed on the dataset
   a. Replace Missing Values and
   b. Normalization of values.
2. Processed dataset is passed through feature selection wherein sets of attributes are deleted from the dataset.
3. The final processed dataset is uploaded in WEKA
4. The J48 Decision Tree and Naïve Bayes algorithm are employed.
5. For purposes of the algorithms, Cross-Validation and Percentage Split techniques are Applied for model creation.

Both models analysed on the basis of correctly classified instances. Figure 4.shows the flow of the research conducted to construct the model

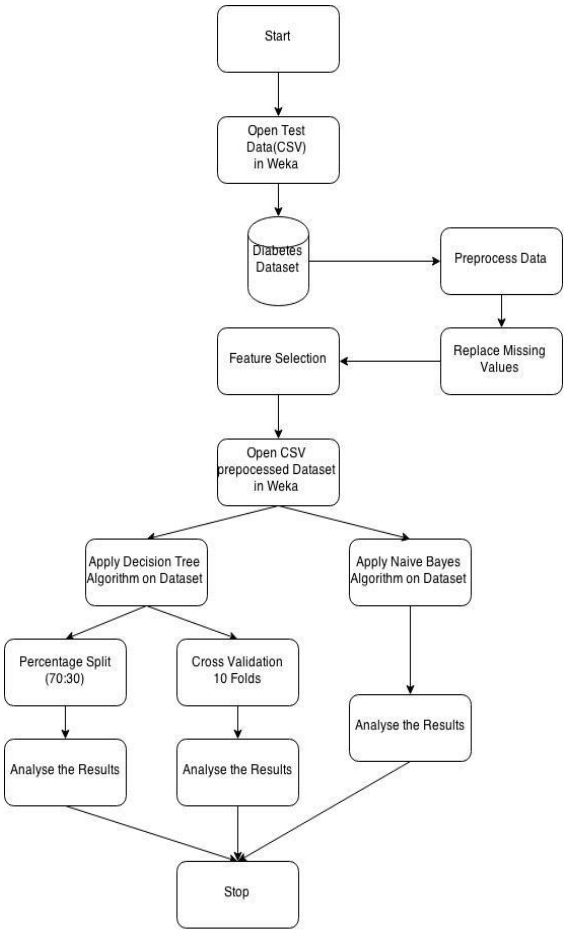


**Figure 6.** Flowchart depicting Model Creation

## V. IMPLEMENTATION

Figure 7 shows the Registration form in which the all fields are mandatory. It will take the patient detail and this detail gets stored in the database.

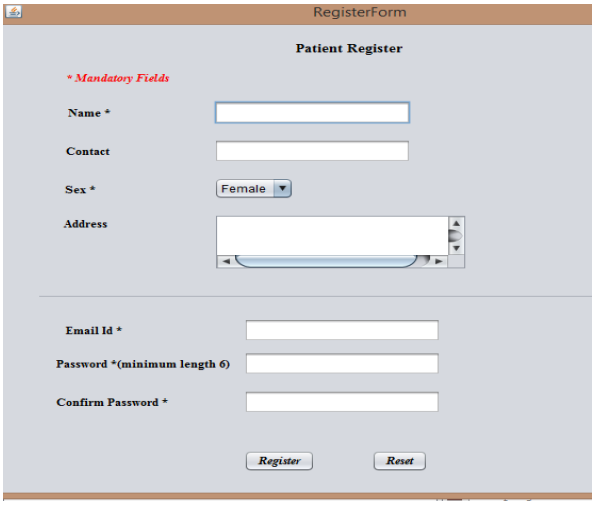


**Figure 7.** Registration form

Figure 8 Shows the login form in which it contain the email id and password field and permit user for getting access to their account.

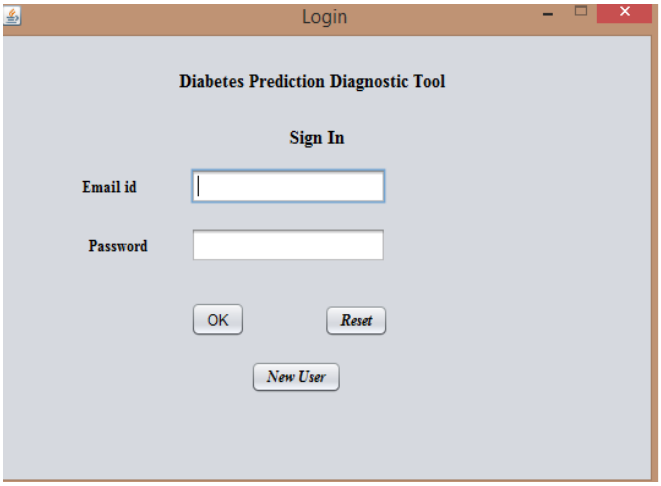


**Figure 8.** Login form

Figure 9 shows the Mandatory procedure it contain the reports which are necessary to analyse the result.

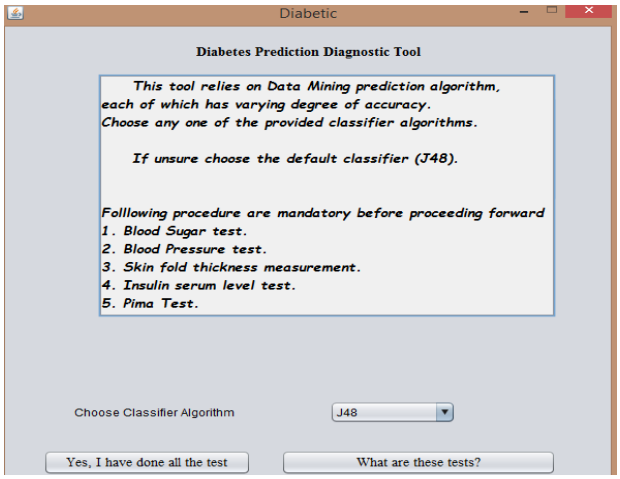


**Figure 9.** mandatory Reports

Figure 10 Contain the following fields from which we are taking the values of attributes using those attribute we are going to analyse the result whether person is having diabetes or not.

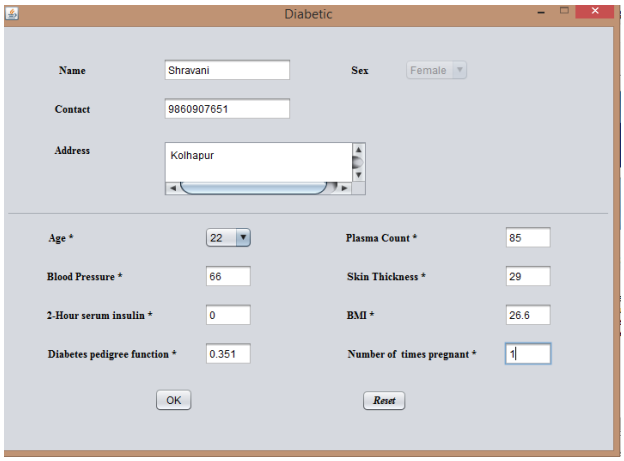


**Figure 10.** Data Filling Form

The Figure 5.5 shows the analysed result.

Figure 11 shows the account of the person, which shows the history of analysis and detail of that person, and here person can modify their details.

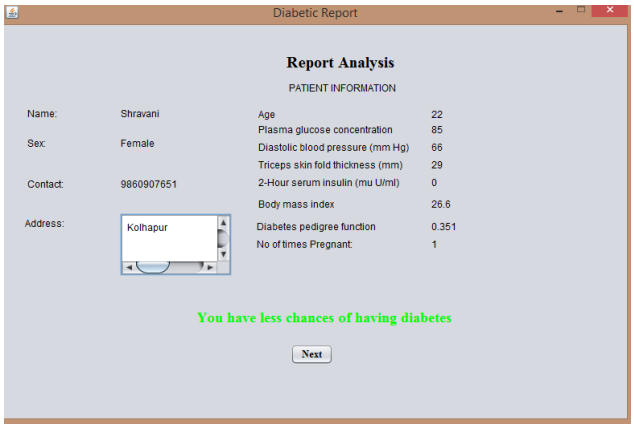


**Figure 11.** Analysed Result

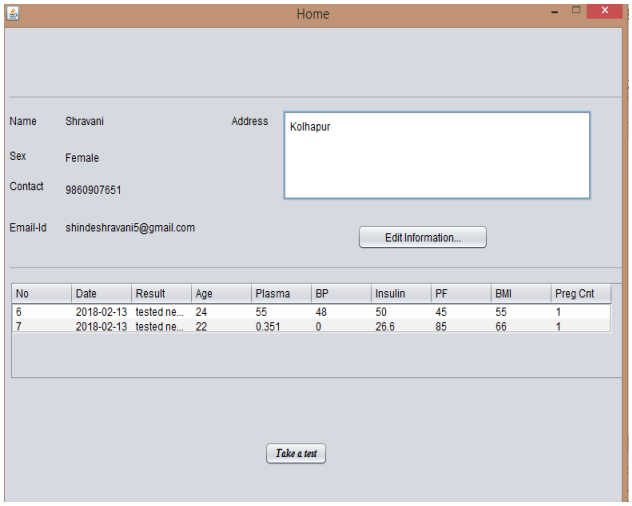


Figure 5.6 Result Review

### 5.1 WEKA Tool: -

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. To implement this project we used weka jar files to integrate weka tool with our JAVA code, which helps us to classify our test data to predict diabetes. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. From multiple prediction and mining algorithms, we choose J48 and Naïve baiyes Algorithm to classify data.

## VI. CONCLUSION

The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. This paper

shows how Decision Trees and Naïve Bayes are used to model actual diagnosis of diabetes for local and systematic treatment, along with presenting related work in the field. Experimental results show the effectiveness of the proposed model. The performance of the techniques was investigated for the diabetes diagnosis problem. Experimental results demonstrate the adequacy of the proposed model.

In future we have planned to gather the information from different locales. This can make a more precise model for diabetes prediction to discover new potential prognostic elements to be incorporated.

## VII. REFERENCES


[1]. National Diabetes Information Clearinghouse (NDIC), http://diabetes.niddk.nih.gov/dm/pubs/type1and2/#signs

[2]. Global Diabetes Community, http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

[3]. Jewie Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2001

[4]. S. Kumari and A. Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of Seventh international Conference on Intelligent Systems and Control, 2013, pp. 373-375

[5]. C. M. Velu and K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", 3rd IEEE International Advance Computing Conference (IACC), 2013

[6]. Sankaranarayanan.S and DrPramanandaPerumal.T, "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233

[7]. MostafaFathiGanji and Mohammad Saniee Abadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease", Proceedings of ICEE 2010, May 11-13, 2010

[8]. T.Jayalakshmi and Dr.A.Santhakumaran, "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", International Conference on Data Storage and Data Engineering, 2010, pp. 159-163

[9]. Sonu Kumari and Archana Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of71hlnternational Conference on Intelligent Systems and Control (ISCO 2013)

[10]. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer

[11]. Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann "Weka: Practical machine learning tools and techniques with Java implementations" Science and Software Engineering, Volume 3, Issue 6, June 2013.

[12]. White, A.P., Liu, and W.Z.: Technical note: Bias in information-based measures in decision tree induction. Machine Learning 15(3), 321–329 (1994)

[13]. https://webcache.googleusercontent.com/search?q=cache:csnTI4lJwxkJ:https://www.niddk.nih.gov//media/19473DA6136D401FA38C3DE767005D0F.ashx+&cd=3&hl=mr&ct=clnk&gl=in

[14]. S R Priyanka Shetty, Sujata Joshi, "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique" I.J. Information Technology and Computer Science, 2016, DOI: 10.5815/ijitcs.2016.11.04