IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, Mar 2019 E-ISSN: 2347-2693

# Prediction of Women's Diabetic Disorder Using R Tool

**G.Kanimozhi[1*], S.Nalini[2]**

[1,2]Dept. of Computer Applications, University College of Engineering, Anna University (BIT Campus), Tiruchirappalli, India

[*]*Corresponding Author: kanimozhig96@gmail.com, Tel. 7094020659*

DOI: https://doi.org/10.26438/ijcse/v7i3.10081011 | Available online at: www.ijcseonline.org

Accepted: 12/Mar/2019, Published: 31/Mar/2019

***Abstract***— This Project is Modern Medicine generates a great deal of information which is deserted into the medical dataset. The proper analysis of such information may reveal some interesting facts, which may otherwise be hidden or go dissipate data analytics is one such field which tries to extract some interesting facts from a huge dataset. In this project, an attempt is made to analyze the diabetic dataset and drive some interesting facts from it which can be a prediction model. Random forest algorithm builds in multiple decision trees and merges them to get a more accurate and stable prediction. A huge medical dataset accessible in different data repositories used in the real world application. This aim of this project is to give a detailed version predictive models from base to state-of-art, describing predictive models, steps to develop a predictive model for determining diabetic disorder.

***Keywords***— Diabetic disorder, Classification, Prediction, And Random Forest

## I. INTRODUCTION

Computer Aided Diagnosis is a rapidly growing dynamic area of research in medical industry. The recent researchers in machine learning machine learning promise the improved accuracy of perception and diagnosis of disease. Here the computers are enabled to think by developing intelligence by learning. There are many types of Machine Learning Techniques and which are used to classify the data sets [678,8]. They are Supervised, Unsupervised Semi-Supervised, and Machine learning algorithms [3]. Advantages of machine learning are it is used in so many industries of application such as banking financial sector, healthcare, retail , publishing and social media. Due to machine learning there are tools available to provide continues quality improvement in large a computer process on environment.

The diabetic disorder for pregnancies women is a group of metabolic disorders the blood sugar level is low, higher than and normal period's time. To increasing rate of diabetes and prediabetes, the health care industry to rightly identify the factors that contribute to the occurrence of diabetes in pregnancies women. From secondary research, factors such as BMI, Blood pressure, cholesterol and Glucose levels are important factors causes' diabetes. To validate the above hypotheses, identify additional risk factors and build tools that can predict the occurrence of diabetes particularly in women. The PIMA Indian's diabetes dataset was chosen. The diabetes data containing information about PIMA Indian's females. It contains information of 768 females, in which 268 females were diagnosed with diabetes. Information available includes 8 variables. Such as Age, Number of pregnancies, Glucose, insulin. A more detailed description of the variables is listed in the table below. The response variable in the dataset is a binary classifier, outcome, that indicates if the person was diagnosed with diabetes or not. The prediction and accuracy level.

## II. RELATED WORK

Statistical models for estimation that are not capable to produce good performance results have flooded the assessment area. Statistical models are unsuccessful to hold categorical data, deal with missing values and large data points. All these reasons arise the importance of MLT. ML plays a vital role in many applications, e.g. image detection, data mining, natural language processing, and disease diagnostics. In all these domains, ML offers possible solutions. This paper provides the survey of different machine learning techniques for diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue and hepatitis disease. Many algorithms have shown good results because they identify the attribute accurately. From previous study, it is observed that for the detection of heart disease, SVM provides improved accuracy of 94.60%. Diabetes disease is accurately diagnosed by Naive Bayes. It offers the highest classification accuracy of 95%. FT provides 97.10% of correctness for the liver disease diagnosis. For dengue disease detection, 100% accuracy is

achieved by RS theory. The feed forward neural network correctly classifies hepatitis disease as it provides 98% accuracy. Survey highlights the advantages and disadvantages of these algorithms. Improvement graphs of machine learning algorithms for prediction of diseases are presented in detail. From analysis, it can be clearly observed that these algorithms provide enhanced accuracy on different diseases. This survey paper also provides a suite of tools that are developed in community of AI. These tools are very useful for the analysis of such problems and also provide opportunity for the improved decision making process. [1]

Diabetes mellitus has been defined as a clinical syndrome characterized by hyperglycaemia, due to deficiency or diminished effectiveness of insulin [5]. Diabetes mellitus has become a global menace. The world health organization has estimated the number of diabetics in the world by 2025 may reach up to 60 million and India's contribution to it would be 30 million. Hence this is a major issue and an awareness regarding this disease is essential. A huge amount of data gets accumulated in the hospitals, most of them just get stored in some form of files which are never touched back, and if these data are analyzed properly they help in deriving some interesting facts. A small touch of data mining will help in generating interesting facts which remained unrevealed otherwise, hence taking into consideration the diabetes mellitus a detailed analysis of diabetic data set is performed using data mining technique [2].

Big Data Analytics in Hadoop's implementation provides a systematic way for achieving better outcomes like availability and affordability of healthcare service to all population. Non-Communicable Diseases like diabetes is one of a major health hazard in India. By transforming various health records of diabetic patients to useful analyzed result, this analysis will make the patient understand the complications to occur. The goal of this research deals with the study of diabetic treatment in the healthcare industry using big data analytics [3].

Various data mining techniques and its application were studied or reviewed .application of machine learning algorithm were applied in different medical data sets Machine learning methods to have different power in different data set.

The single algorithm provided less accuracy than the ensemble one.in most study decision tree provided high accuracy.in this study hybrid system, Weka and java are the tools to predict diabetes dataset [4].

Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy of study. In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree, and neural network, but random forests are obviously better than other classifiers in some methods. The best result for Liuzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important. Due to the data, we cannot predict the type of diabetes, so in future, we aim to predict the type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes [5].

Machine learning is one of the key methods used in modern day analysis. With the explosion of the IT industry, and the rise of big data; it became necessary to analyze and predict trends. Slowly over the years machine learning has branched out into almost every major industry, and performs functions that were almost unheard, compared to that mere few years ago. In this work we have concentrated exclusively upon the supervised algorithms. Supervised algorithms can be broadly classified into two sub divisions- Regression algorithms- Regression algorithms are used to predict continuous values. For example, Naive Bayes and KNN algorithms. Classification algorithms- Classification algorithms are used in order to predict discrete values. For example, linear regression and K Means algorithms. For this project we have extensively used scikit Learn which is a module that is built on top of sci py library in python version 2 onwards. Scikit Learn is a python library that exclusively focuses on data science and the various classifications, regression and clustering algorithms including support vector machines, k-NN, random forests, Logistic Regression, gradient boosting, Naive Bayes, k-means, and Decision Tree, and is designed to operate within the python script for the Python numerical and scientific libraries NumPy and SciPy[9].

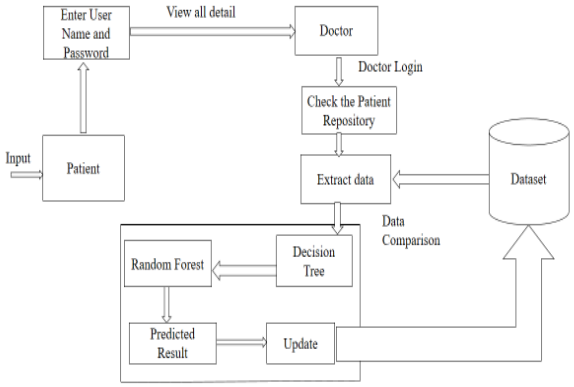
## III. METHODOLOGY

This Proposed System consists of eight input parameter value and one of the two possible outcomes. Namely, whether the patient is tested positive for diabetes (indicate by output one) or not (indicated by zero).

The data set is analyzed by R software. By using eight attributes as an input and determine the diabetic disorder. Random Forest algorithms help to analyze the dataset which gives tested as Maximum and Minimum. Then determining the accuracy level.

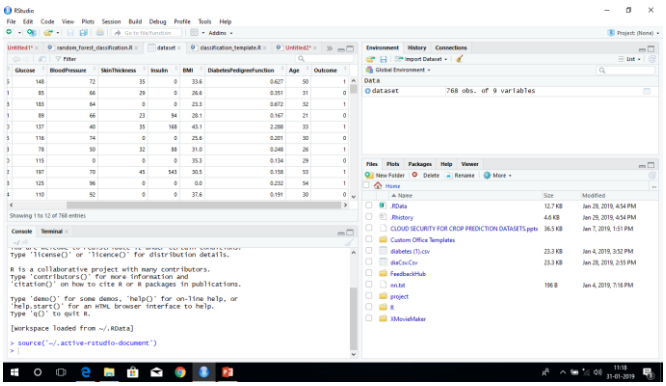
PROPOSED ARCHITECTURE DIAGRAM:




## IV. RESULTS AND DISCUSSION

5.1 Loading the dataset:

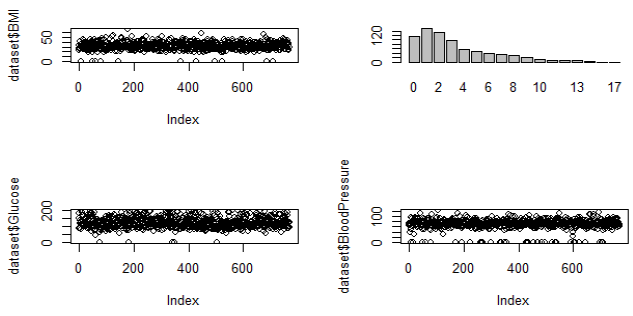
The dataset is importing the data set. Loading the data set .then view the dataset.



5.1 Loading the Dataset

5.2 Input Plotting:

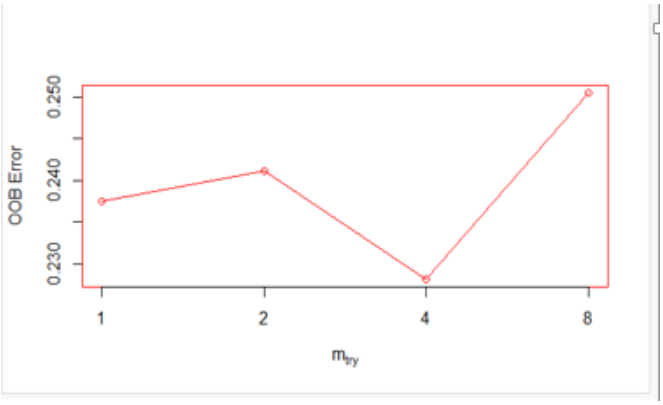
The dataset is ncluded into the eight attributes.eight attributes comparision to the dataset.




5.2. input plotting

5.3 Random Forest:

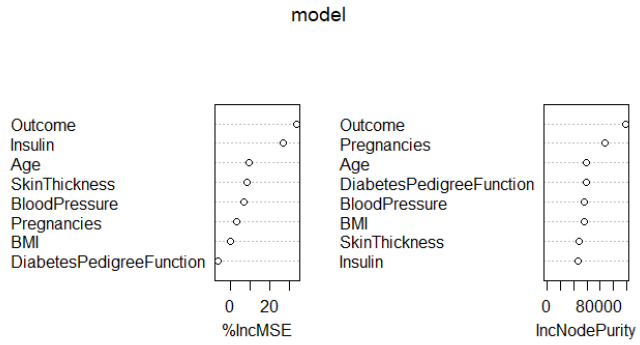
Create a Random Forest model with default outcome.then only mtry parameter using tuneRF() and OOB Error.

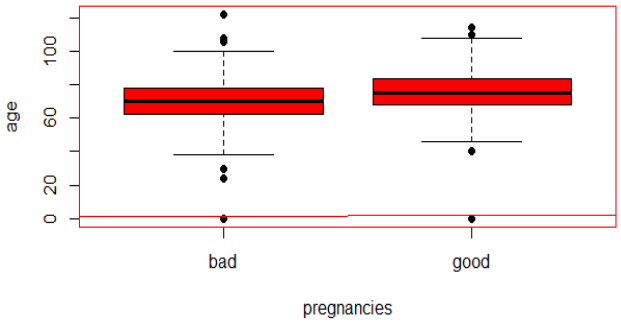



5.3 Random Forest Output

5.4 Accuracy Level:

Check the number of unique values. Into align the eight attributes condition checking in prediction for the accuracy level. View the report in good and bad.




5.5 Accuracy level Bad and Good

*Tables:*

The input parameter ranges in less than and greater than using in the value. Prediction of the accuracy level in the dataset.]

INPUT PARAMETER RANGE ANALYSIS:

| S . No | Attributes | Minimum | Maximum |
|---|---|---|---|
| 1 | Pregnancies | 4.0 - 5.5 mmol /L (95mg/dl) | 7.00 mmol/L (140mg/dl) |
| 2 | Glucose | 60 to 99 mg/dl | 100 to 129 mg/dl |
| 3 | Blood Pressure | Level:120/80 mm Hg is considered normal | |
| 4 | Skin Thickness | 66 IDDM patients aged 24–38 year | |
| 5 | BMI | BMI of 18.5-24.9 is considered normal | 30-39.9 is classified as obese |
| 6 | Pedigree Function | 0.5 | 2.329 in Accuracy (72.3%). |
| 7 | Insulin | < 25 mIU/L | < 174 pmol/L |
| | | 30-230 mIU/L | 208-1597 pmol/L |
| | | 18-276 mIU/L | 125-1917 pmol/L |
| | | 16-166 mIU/L | 111-1153 pmol/L |

## V. CONCLUSION AND FUTURE SCOPE

A detailed analysis of the diabetic data set was carried out efficiently with the help of R. The facts which were revealed during the process can be used for developing some prediction models. In this work only the analysis is carried out but the information which was revealed can be further used to develop efficient prediction models. There are prediction accuracy level 95.2%s. The outcome is diabetic disorder yes or no in pregnancies women. Predicting for the accuracy level in using random forest algorithm.by the future work the additional process of we going to update monthly check up -diabetic patient details will store automatically in the dataset. It will help to view dataset of blood sugar level as per monthly check up both doctor and patient.

.

## REFERENCES


[1] Meherwar Fatima1, Maruf Pasha2, "Survey of Machine Learning Algorithms for Disease Diagnostic" , Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16.

[2] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R ", International Journal of Emerging Technology and Advanced Engineering, 2014.

[3] Dr. Saravana Kumar, Eswari, Sampath & Lavanya," Predictive Methodology for Diabetic Data Analysis in Big Data", ELSEVIER, 2015.

[4] Rahul Joshi, Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach" International Research Journal of Engineering and Technology (IRJET) , 2017.

[5] Quan Zou, Kaiyang Qu , Yamei Luo , Dehui Yin, Ying Ju, and Hua Tang," Predicting Diabetes Mellitus With Machine Learning Techniques" Frontiers in Genetics,2018.

[6] Abdullah A. Aljumah, Mohammed Gulam Ahmad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients" in Journal of King Saud University – Computer and Information Sciences (2013).

[7] Allen Daniel Sunny1, Sajal Kulshreshtha2, Satyam Singh3, Srinabh4, Mr. Mohan Ba5, Dr. Sarojadevi H.6," Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology (IJIET), Volume 10 Issue 2 May 2018

[8] http://www.patient.co.uk/doctor/diabetes-mellitus

[9] http://www.idf.org/diabetesatlas/introduction

[10] https://data.world/data-society/pima-indians-diabetes-database/workspace/project-summary