Machine Learning In Predicting Diabetes In The Early Stages

SEMINAR REPORT

Submitted by,

Niranj Rajesh
(RIE18CS023)

to

APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of B-Tech Degree in Computer Science & Engineering.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RAJADHANI INSTITUTE OF ENGINEERING & TECHNOLOGY
THIRUVANANTHAPURAM 695102

DECEMBER 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that this report entitled "Machine Learning in Predicting Diabetes in the Early Stages" is the report of seminar presented by **Niranj Rajesh, RIE18CS023** during 2021-2022 in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science & Engineering of the APJ Abdul Kalam Technological University.

GUIDE
Mrs. Vishagini V
Assistant Professor
Dept. of Computer Science & Engineering
Rajadhani Institute of Engineering & Technology

SEMINAR COORDINATOR
Mrs.Sinciya
Assistant Professor
Dept. of Computer Science & Engineering
Rajadhani Institute of Engineering & Technology

HEAD OF THE DEPT.
Ms. SANGEETHA SHIBU
Associate Professor
Dept. of Computer Science & Engineering

DECLARATION

I, **Niranj Rajesh** hereby declare that, this seminar report entitled **"Machine Learning in Predicting Diabetes in the Early Stages**" is the bonafide work of mine carried out under the supervision of **Mrs.VISHAGINI V**, Assistant Professor, Rajadhani Institute of Engineering and Technology. I declare that, to the best of my knowledge, the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occa-sion to any other candidate. The content of this report is not being presented by any other student to this or any other University for the award of a degree

Signature:
Name of the Student: NIRAN RAJESH
Uni. Register No: RIE18CS023 of year 2018

Signature(s):
Name of Guide(s): VISHAGINI V

Countersigned with Name:
Dept. of Computer Science & Engineering
Rajadhani Institute of Engineering & Technology Date:

ACKNOWLEDGEMENTS

I take this opportunity to express my deep sense of gratitud~ and sincere thanks to all who helped me to complete the project successfully. I am deeply indebted to my guide Mrs. **VISHAGINI V**, Assistant Professor, Department of Computer Science & Engineering for her excellent guidance, positive criticism and valuable comments. I am greatly thankful to **Mrs. SANGEETHA SHIBU** Head of Computer Science & Engineering Department for _her support and cooperation. Finally, I thank everyone who directly and indirectly contributed to the successful completion of my seminar.

NIRANJ RAJESH

Place:

Date:

ABSTRACT

Diabetes is a common disease and its early symptoms are not very noticeable, so an efficient method of prediction will help patients make a self-diagnosis. However, the conventional method to identify diabetes is to make a blood glucose test by doctors and the medical resource is limited. The process of Machine Learning is to train a computational algorithm for prediction based on a big dataset. It is popular for its efficiency and accuracy. Also, it has the advantage of dealing with tons of data, so we can make diagnoses for plenty of patients in a short time and the result will be more accurate. Our data was from **UCI Machine Learning Repository**, which was collected by direct questionnaires from the patients of the Sylhet Diabetes Hospital. **Random Forest algorithms** are often used for each classification and regression tasks and also it is a type of ensemble learning method. The accuracy level is greater when compared to other algorithms. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Keywords: Diabetes, Predictive models, Machine Learning, Random Forest.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Diabetes is a very common disease. According to the
National Diabetes Statics Report , as of 2015, 30.3 million people in the United States had diabetes, which means one of ten people in the United States is suffering from diabetes. In addition, one in ten of them do not even know they have the disease. It is also a chronic disease affecting the quality of life negatively, since most patients must deal with diabetes every day and it can lead to problems affecting almost every body systems. Hence, it is necessary to prevent and diagnose the disease.
An accurate and timely diagnosis will help patients prevent the diabetes, and it helps the patients find out whether they get diabetes in the early stage. However, the medical resource is limited, and doctors can only make diagnoses for certain number of patients in the limited time. Therefore, most people make an assessment based on their experience and symptoms. However, most patients lack professional medical knowledge, and they are just based on what they know and what they hear so it is inaccurate for patients to make diagnoses for themselves. Hence, it is necessary to make an efficient prediction model, which can save medical resources and help patients make a self-test accurately.

Several researchers square measure conducting experiments for identification the diseases exploitation machine learning approaches. This research work focuses on accuracy rate of diabetes which affects the people.The objective of our study is to make machine learning models to predict diabetes. The process of machine learning is like the computer algorithms are learning through experience instead of human, which means that they are much more efficient than humans. Also, machine learning is more and more popular nowadays. It can easily handle problems in high dimensional space. Decision tree is a tree-like structure. Each branch represents different outcomes. Decision tree can handle the non-linear relationship. Random forest is a representative ensemble learning method. Random forest is a way to average multiple decision trees and reduce its variance. In this work, we use the Random Forest rule.

1.2 DATASET

The data set we used is from UCI machine learning repository, which is public available at http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. It is collected from the patients of Sylhet Diabetes Hospital in Sylhet. There are17 attributes and 520 instances totally. Detailed information about the variables are listed in Table 1. Only age is a numerical variable, and others are categorical variables. 320 samples are positive cases and 200 samples are negative cases. We partition the whole data set into 2 subsets, the train set containing 364 cases and the test set containing 156 cases. We normalize the age to 0~1 by min-max normalization.

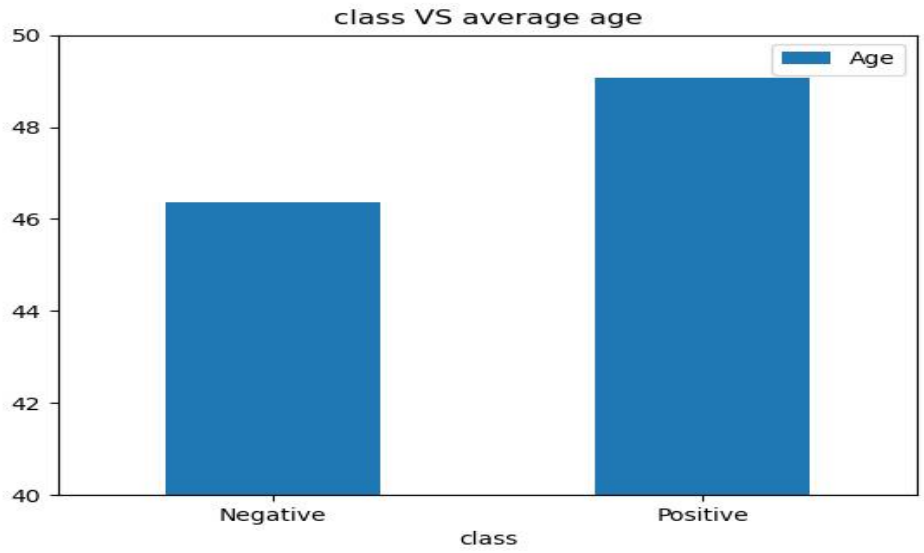
In our study, we choose the attributes such as sudden weight loss, obesity to construct our prediction model, which are more understandable and accessible. The patients do not need to do some medical tests, which makes our model more understandable and applicable.

TABLE I. THE VARIABLES IN THE DATA SET

Age	**Polyphagia**	**partial paresis**
Sex	**Polyuria**	**Genital thrush**
muscle stiffness	**Polydipsia**	**weakness**
sudden weight loss	**Itching**	**Obesity**
delayed healing	**Alopecia**	**visual blurring**
Irritability	**Class**	

We conduct the descriptive statistics for each attribute to investigate the relation between each attribute and diabetes. First, we plot the average age for each class in Figure 1, and we can see that the average age of people with diabetes is higher. For the categorical variables, we calculated morbidity for each class of every attribute and the results are summarized in Table 2.

FIG. 1 THE AVERAGE AGE FOR POSITIVE AND NEGATIVE SAMPLES



People have symptoms like delayed healing, Obesity, muscle stiffness, Itching, Polyphagia, Polyuria, sudden weight loss, Genital thrush, Irritability, Polydipsia, partial paresis, weakness, visual blurring are more likely to get diabetes, while symptoms like Alopecia , Itching imply lower probability.

TABLE II. MORBIDITY FOR EACH CLASS OF EVERY ATTRIBUTE

	Alopecia	**delayed healing**	**Gender**	**Itching**	**muscle stiffness**
No	71.0%	59.4%	90.1%	62.2%	56.9%
Yes	43.6%	64.0%	44.8%	60.9%	69.2%
	sudden weight loss	**Genital thrush**	**Irritability**	**partial paresis**	**Polydipsia**
No	43.6%	58.7%	53.3%	43.2%	33.1%
Yes	86.6%	71.6%	87.3%	85.7%	96.6%
	Obesity	**Polyphagia**	**Polyuria**	**visual blurring**	**weakness**
No	60.0%	46.3%	29.4%	50.5%	47.4%
Yes	69.3%	79.7%	94.2%	75.1%	1.5%

CHAPTER 2

LITERATURE SURVEY

2.1 DIABETES PREDICTION USING NEURAL NETWORKS

Nesreen Samer El-Jerjawi, and **Samy S. Abu-Naser** proposed an **ANN model** to predict diabetes that can be useful and helpful for doctors and practitioners. In this research, they used the following attributes: Number of pregnancies, PG Concentration (Plasma glucose at 2 hours in an oral glucose tolerancetest), Diastolic BP (Diastolic Blood Pressure (mm Hg))), Tri Fold Thick (Triceps Skin Fold Thickness (mm)), Serum Ins(2-Hour Serum Insulin (muU/ml)), BMI (Body MassIndex: (weight in kg/ (height in m)^2)), DP Function(Diabetes Pedigree Function), Age (years), Diabetes (Whether or not the person has diabetes).

So in this paper, they used artificial neural networks to predict whether a person is diabetic or not. The criterion was to minimize the errorfunction in neural network training using a neural network model. After training the ANN model, the average error function of the neural network was equal to **0.01** and the accuracy of the prediction of whether a person is diabetics or not was **87.3%**.

The Objectives of the Study

- To predict and categorize the state of health.
- To identify some appropriate factors that affect health conditions.
- To design an artificial neural network that can be used to predict health performance based on certain pre-defined data for a particular health condition

2.2 PREDICTION OF DIABETES USING DATA MINING TECHNIQUES

Fikirte Girma Wolde Michael and **Sumitra Menaria** in this study proposed to predict diabetes using data mining techniques. **Back propagation** algorithm is used to predict whether the person has diabetic or not. And also **J48**, **naïve bayes** and **support vector machine** were used to predict diabetes. These neural networks were having an input layer with having 8 parameters, one hidden layer having 6 neurons and produce one output layer.5 fold cross-validation technique and large value learning rate was used to improve the performance of the model. PIMA Indian dataset used to conduct this study. The study implemented in RStudio using R programming language.The performance of Back propagation algorithm is used to predict diabetes diseases gave **83.11%** accuracy, **86.53%** sensitivity and **76%** specificity, the result shows improvement from previous work. The obtained result is also compared with J48, naïve bayes and support vector machine algorithm.

The Objectives of the Study

- To explore and investigate the data mining based diagnosis and prediction solutions for diabetes.
- Provide comprehensive classification and comparison of the techniques that have been frequently used for diagnosis and prediction of diabetes based on important key metrics.
- To highlight the challenges and future research directions in this area that can be considered in order to develop optimized solutions for diabetes detection and prediction.

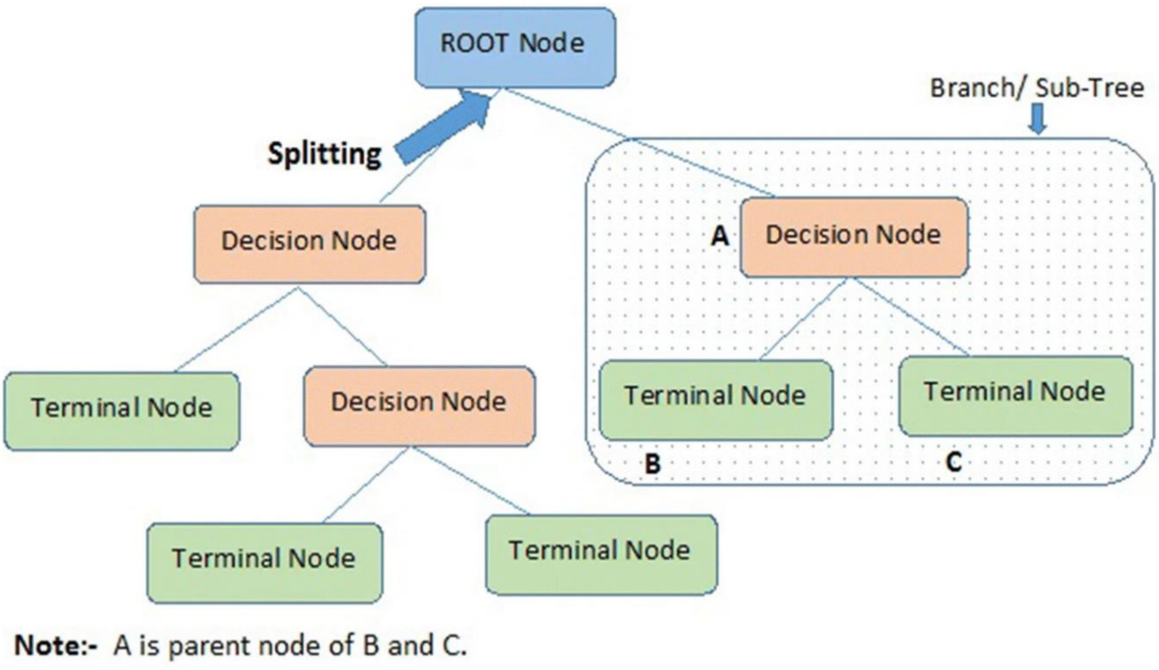
CHAPTER 3
EXISTING SYSTEM

Decision tree is a basic classification and regression method. Decision tree model has a tree structure, which can describe the process of classification instances based on features. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space.Decision tree uses tree structure and the tree begins with a single node representing the training samples. If the samples are all in the same class, the node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree.

According to the value of the current decision node attribute, the training samples are divided into serval subsets, each of which forms a branch, and there are serval values that form serval branches. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples.

There are many ways to choose the best attribute to be as the root node, based on the degree of impurity of the child nodes.The Performance measures are Entropy, Giniindex, classification error. These measures are done for all attributes and comparison is done, to select the best spilt.

FIG. 2 DECISION TREE MODEL



In this model, we use min_samples_split which means the minimum number of samples required to split an internal node to control the complexity of the model. As the minimum number becomes smaller, the model becomes more complex.

We use different parameter from 2 to 29 to build the model and calculate train error and test error which is presented in the following figure.
The graph is plotted against minimum number of samples required for best split and the error.

FIG. 3 DECISION TREE TRAIN AND TEST ERROR

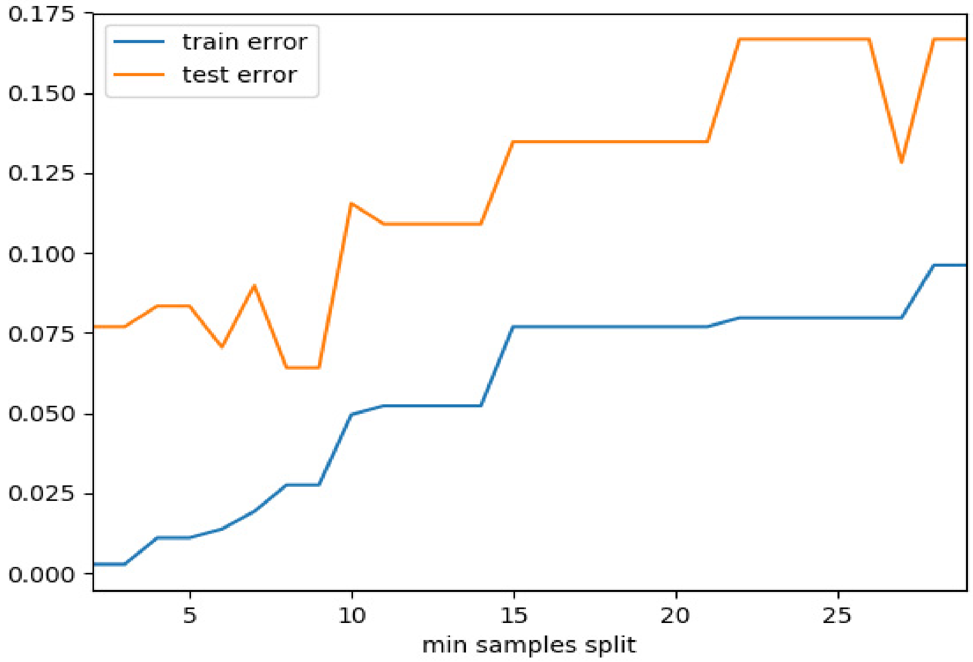


TABLE III. THE CONFUSION MATRIX FOR DECISION TREE

	Train	
	Negative Predication	***Positive Predication***
Negative True	90.60%	9.40%
Positive True	4.30%	95.70%
	Test	
	Negative Predication	***Positive Predication***
Negative True	77.60%	22.40%
Positive True	20.30%	79.70%

CHAPTER 4

PROPOSED SYSTEM

Due to the instability of single decision tree, it comes out the ensemble learning which combines multiple models to improve the prediction accuracy overall and reduce the variance. Random forest is a typical ensemble learning method, and the number of decision trees controls the complexity of the model, so we try different number of trees and calculate their errors .

4.1 SYSTEM ARCHITECTURE

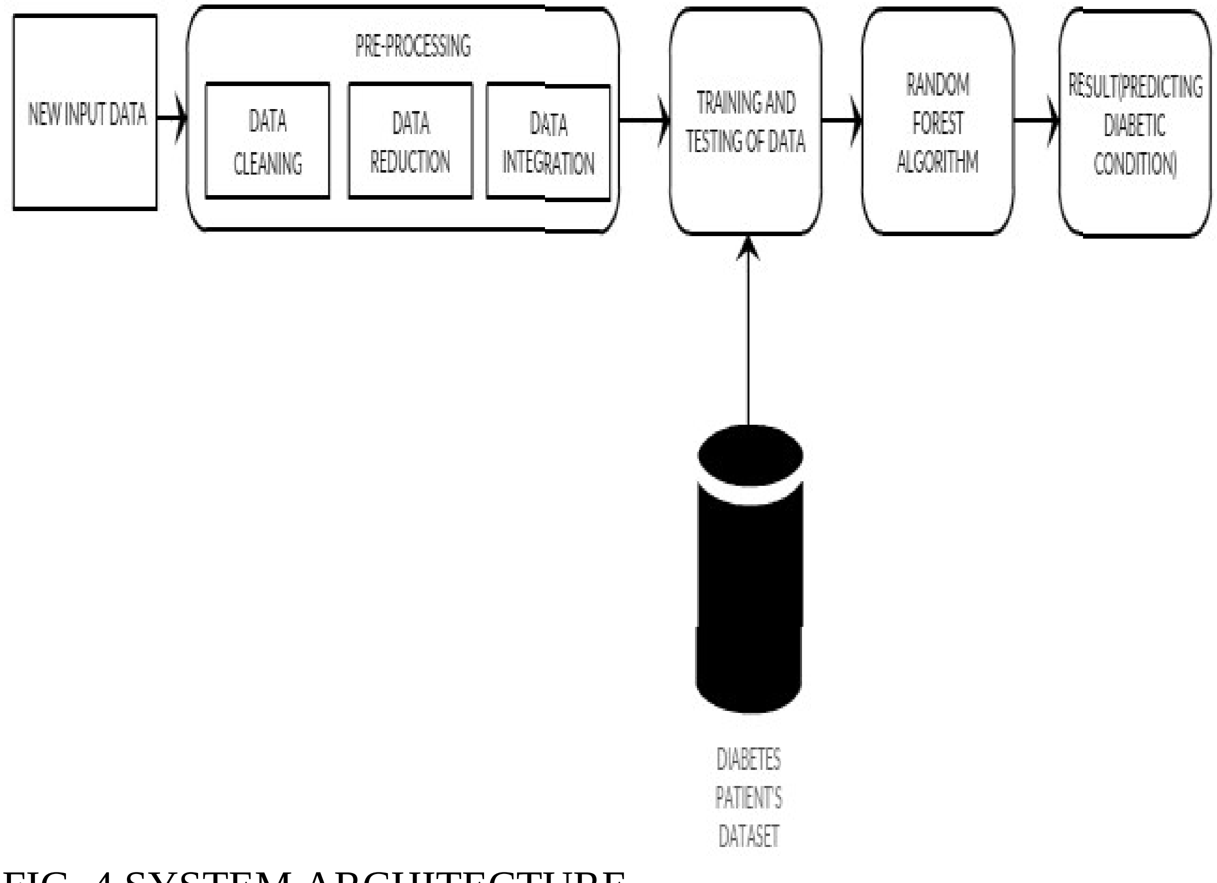


FIG. 4 SYSTEM ARCHITECTURE

The datasets are collected from the database. In phase two the data will be pre-processed which will include data cleaning, integration and transformation. By using Random Forest algorithm we can find better accuracy when compared to other algorithms.

The New Input Data is given and goes through various phases. The new input data will be pre-processed which will include data cleaning, integration and transformation.Then the data is finally tested with the model.

4.2 DATA PRE-PROCESSING

Data pre-processing is one vital step in data discovery methodology. Most health care information contain missing value, wheezy and inconsistency information.

The most common form of predictive modeling project involves so-called structured data or tabular data. This is data as it looks in a spreadsheet or a matrix, with rows of examples and columns of features for each example.

We cannot fit and evaluate machine learning algorithms on raw data; instead, we must transform the data to meet the requirements of individual machine learning algorithms. More than that, we must choose a representation for the data that best exposes the unknown underlying structure of the prediction problem to the learning algorithms in order to get the best performance given our available resources on a predictive modeling project.

Given that we have standard implementations of highly parameterized machine learning algorithms in open source libraries, fitting models has become routine. As such, the most challenging part of each predictive modeling project is how to prepare the one thing that is unique to the project: the data used for modeling.

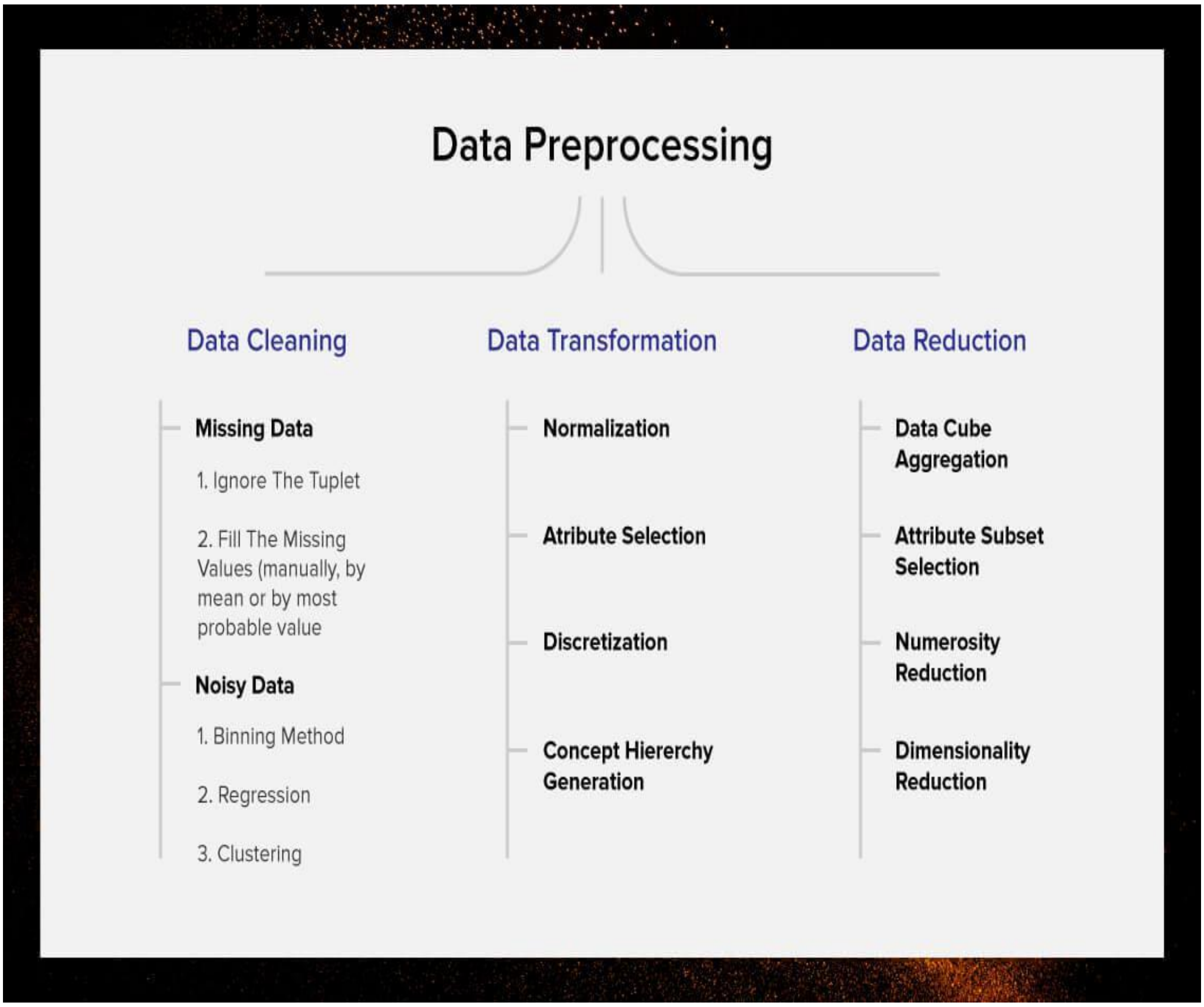
Data cleaning is that the tactic of detection and correcting (or removing) corrupt or inaccurate records from a record set, table, or data and refers to distinguishing incomplete, incorrect, inaccurate or tangential parts of the knowledge some substitution, modifying, or deleting the dirty or coarse data.Data cleansing is additionally performed interactively with data twenty five haggle tools, or as execution through scripting. Information cleansing is in addition said as information clean-up or information cleansing.

eg: drop the duplicates, check for null values and replace with mean values.

Data integration could be a method within which heterogeneous knowledge is retrieved Associate in Nursing combined as an incorporated kind and structure. Knowledge integration permits fully completely different information kinds (such as information sets, documents and tables) to be integrated by users, organizations and applications, to be used as personal or business processes and or functions.

Data reduction is that the transformation of numerical or alphabetical digital data derived through empirical observation or by experimentation into a corrected, ordered, and simplified type. The fundamental construct is that the reduction of undeterminable amounts of data all the means right down to the purposeful components.

FIG. 5 DATA PREPROCESSING



Data Preprocessing

Data Cleaning

- Missing Data
  1. Ignore The Tuplet
  2. Fill The Missing Values (manually, by mean or by most probable value
- Noisy Data
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

- Normalization
- Atribute Selection
- Discretization
- Concept Hiererchy Generation

Data Reduction

- Data Cube Aggregation
- Attribute Subset Selection
- Numerosity Reduction
- Dimensionality Reduction

4.3 RANDOM FOREST ALGORITHM

Random Forest was developed by Leo Bremen. Random Forest rule may well be a supervised classification rule, there square measure two stages in Random Forest rule, one is random forest creation, and thus the choice is to make a prediction from the random forest classifier created among the first stage , the pseudo code for Random Forest is

a. The first step is to select the "**R**" features from the total features "**m**" where **R<<m**.

b. Among the "**R**" features, the node using the best split point.

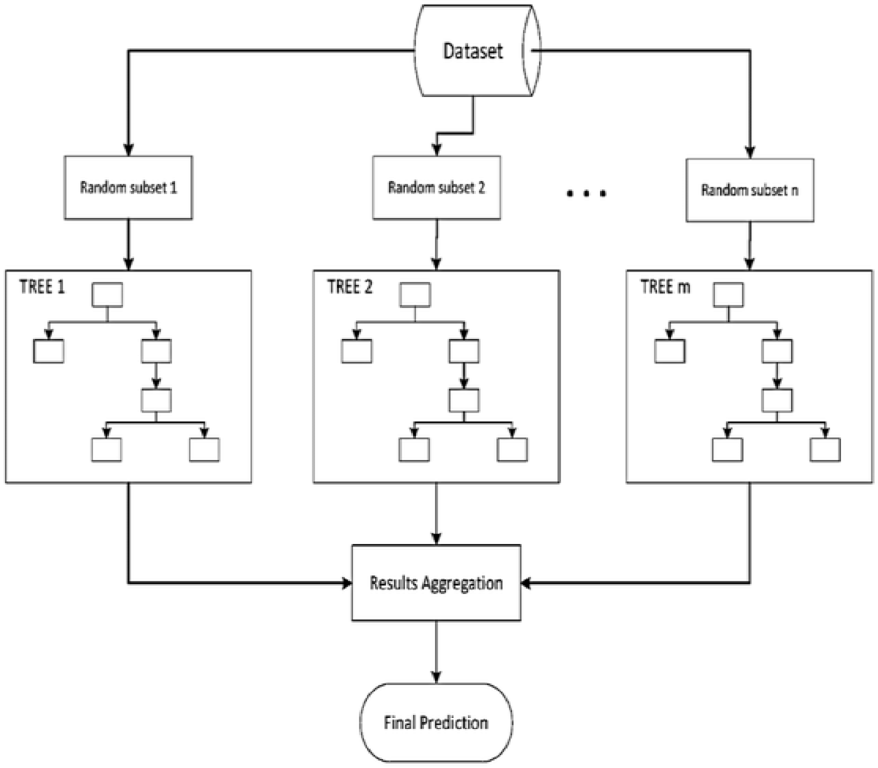
c. Split the node into daughter nodes using the best split.

d. Repeat a to c steps until ”**l**” number of nodes has been reached.

e. Built forest by repeating steps a to d for ”**a**” number of times to create "**n**" number of trees.

The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. The Random Forest simplified diagram is given below

FIG. 6 RANDOM FOREST SIMPLIFIED



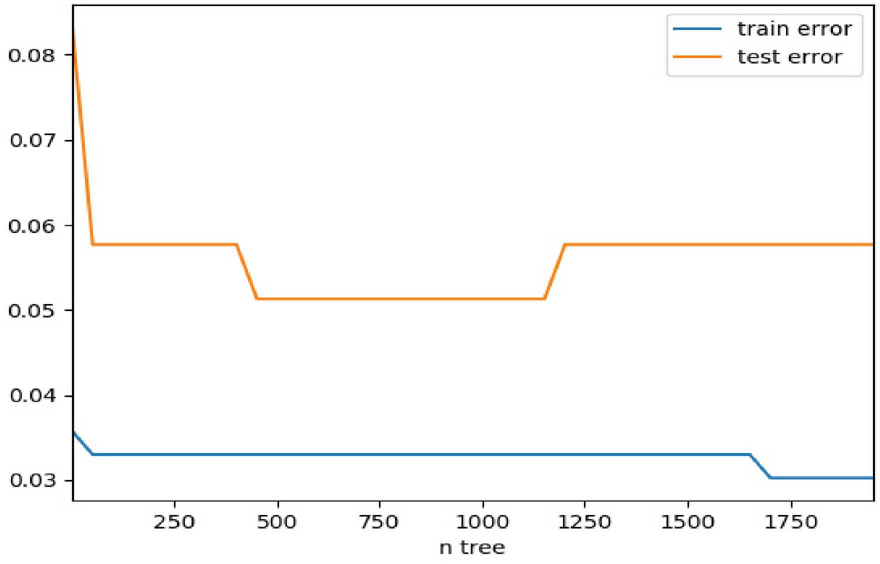
Some of the options of Random Forest does correct predictions result for a spread of applications areoffered. Through model coaching the importance of every feature may be measured. and therefore the trained model will live the pair-wise proximity between the samples.The Advantages of victimization Random Forest algorithmic rule ar for each the classification and regression task, identical random forest algorithmic rule may be used. For applications in classification issues, it'll avoid the over fitting downside, and it may be used for distinctive the foremost vital options from the coaching dataset.

CHAPTER 5

RESULTS

we try different number of trees and calculate their errors which is plotted in the graph given below.

FIG. 7 THE TRAINING AND TESTING ERROR FOR RANDOM FOREST



testing error of random forest is not likely going up with larger number of trees, which implies random forest is less likely to over fit. Even though the training error continues to go down, the testing error keeps stable after falling, which means it is not likely to be an overfitting model relatively.

The confusion matrix is summarized in Table 4.

TABLE IV. THE CONFUSION MATRIX FOR RANDOM FOREST

	Train	
	Negative Predication	***Positive Predication***
Negative True	100.00%	0.00%
Positive True	5.30%	94.70%
	Test	
	Negative Predication	***Positive Predication***
Negative True	93.80%	6.30%
Positive True	4.30%	95.70%

Compared with single decision tree, the testing accuracy of random forest, which is 94.9%, increased near 16.8 percent.

CHAPTER 6

CONCLUSION

One of the required real-world medical problems is that the detection of genetic defect at its early stage. Throughout this study, systematic efforts area unit created in coming up with a system that finally ends up among the prediction of illness like genetic defect.An efficient diabetes prediction model will help doctors make accurate diagnoses and help patients get timely treatment. We conduct descriptive statistics for diabetes risk prediction dataset to investigate the variables which influence the diabetes.

Throughout this work Random Forest algorithms area unit studied and evaluated on varied measures. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using machine learning technique which provides advance support for predicting the accuracy rate of diabetes.

The aim of this project is to develop a system which can perform early prediction of diabetes with higher accuracy and provide advance support for predicting the accuracy rate of diabetes.

CHAPTER 7

REFERENCES

1. Fikirte Girma Wolde Michael, Sumitra Menaria, “ Prediction of Diabetes using Data Mining Techniques, Dept.of ComputerScience and Engineering, Sumitra.Menaria@paruluniversity.ac.in, Proceedings of the 2nd International conference on Trends in Electronics and Informatics(ICOEI 2018).

2. El_Jerjawi, Nesreen Samer & Abu-Naser, Samy S. (2018). Diabetes Prediction Using Artificial Neural Network. International Journal of Advanced Science and Technology 121:54-64.

3. Deepti Sisodiaa, Dilip Singh Sisodiab,Prediction Diabetes and Classification Algorithm A National Institute of Technology, G.E Road, Raipur and 492001, India, International Conference on Computational Intelligence and Data Science.

4. A.M.Rajeswari,M.Sumaiya Sidhika,M.Kalaivani C.Deisy,Prediction of Pre-Diabetes using Fuzzy Logic Based Association Classification�, Thiagarajar College of Engineering,Madurai, India Proceedings of the (ICICCT 2018), cdcse@tce.edu.

5.Alpayd, Ethem. Introduction to Machine Learning, 2nd ed.[J]. Methods Mol Biol, 2014, 1107(1107):105-128.

6.Shalev-Shwartz, Shai; Ben-David, Shai (2014). "18. Decision Trees". Understanding Machine Learning. Cambridge University Press.

7.Krogh A, Sollich P. Statistical mechanics of ensemble learning[J]. Phys.rev.e, 1997, 55(1).

8.Centers for Disease Control and Prevention. National diabetes statistics report, 2017. Centers for Disease Control and Prevention website. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetesstatistics-report.pdf External link (PDF, 1.3 MB) . Updated July, 18 2017. Accessed August 1, 2017.