PREDICTIVE ANALYSIS OF DIABETIC PATIENTS USING MACHINE LEARNING TECHNIQUES AND R

Rifat Ameena, R[1] and Ashadevi,B[2]

[1]Research Scholar, Department of Computer science, M.V. Muthiah Government Arts College for Women, Dindigul & India, rifat.ameena7869@gmail.com

[2]Assistant Professor, Department of Computer science, M.V. Muthiah Government Arts College for Women, Dindigul & India, asharajish2005@gmail.com

---

Abstract

Today diabetes is a very common disease with all age groups which leads to heart disease as well as increases the risks of developing Nephropathy, Neuropathy, Retinopathy, Polycystic Ovarian Syndrome, Gastro paresis and depression. Diabetes Mellitus is a chronic disease associated with abnormally high levels of the blood sugar levels over a prolonged period. It is one of the most serious health challenges even in developed countries. Early prediction and continuous monitoring of diabetic patients can help a person live a normal life. Big data refers to massive volume of both structured and unstructured datasets with complex structures that are difficult to capture, store, format, extract, cure, integrate, analyse and visualize using traditional methods and tools. The main objective of our thesis is Detecting Diabetes with PIMA Indian Diabetes Data set using R tool. PIMA are people of Indian American origin. In this research different classifying algorithms such as K Nearest Neighbour and Neural Network were applied to the dataset. The results obtained from our experiments indicates that K Nearest Neighbour model gives higher accuracy of 80%. This paper concludes the evaluation of K nearest neighbour and Neural Network in order to predict the diabetic patients with statistical implication using R.

Keywords: PIMA, R framework, Machine Learning, K Nearest Neighbour, Neural Network.

INTRODUCTION

This work aims at both detecting Diabetes as well as predicting the risk of diabetes in PIMA Indian Women data set. PIMA are people of Indian American origin. The framework used here is R Studio with R programming language. The selection of R framework is done with keeping in mind that it forms an important aspect of Data Analytics and Visualization Studio provides a statistical tool with support of machine learning and visualization language is easy to learn, provides high code density, open source, freely available, easy to install and provides sophisticated results also. It has huge web support also. One of the striking features of R Studio is that it can be combined with Spark, Hadoop that is mainly required to handle

big data sets. Hence, power of cause of many associated health issues like heart attacks, liver failure, kidney failures, nerve damage both big-data as well as analytics can be handled.

Present day life-style has instilled very serious problem of Diabetes. It is the situation in which there is high blood sugar levels and poor circulation. Women with diabetes has poor immunity which reduces body's ability to fight infections. This is the major cause for many health problems like heart attacks, obesity, nerve damages, kidney failures, liver failure, high blood pressure, vision loss and Polycystic Ovarian Syndrome (PCOD). This PCOD is frequently occurring in women nowadays because of high resistance towards insulin. As a result, even at teen age chances of being diabetic has increased a lot. It also causes problems during pregnancy. Hence, diabetes detection and prediction is an important concern to provide better health care services especially for Women. There are three types of diabetes namely,

A. Type-1 Diabetes

It frequently occurs in the children where no insulin is produced in the body. Pancreatic cells are destroyed due to this no glucose is formed in the body. It is generally known as juvenile diabetes. Common symptoms are-weight loss, dehydration and damage to body parts like liver, kidney, vision loss, Urinary infections etc.

B. Type- 2 Diabetes [1]

Type 2 diabetes (also called diabetes mellitus type 2) is the most common form of diabetes since it accounts for 90% of diabetes cases. It is a long term metabolic disorder that is characterised by high blood glucose and insulin resistance. In addition, it results from the body's ineffective use of insulin [3]. There are two main causes of type 2 diabetes, namely an increase in body weight and a lack of physical activity [3, 4]. Rates of this type of diabetes have increased considerably since 1960 in conjunction with increasing rates of obesity [5]. The number of type 2 diabetic patients increased from approximately 30 million in 1985 to around 368 million in 2013 [6, 7]. Until recently, type 2 diabetes was seen only in adults, but is now becoming increasingly common in young people [3].

C. Gestational Diabetes

This type of diabetes is specific to pregnant women. Gestational diabetes occurs later during pregnancy. Women with glucose tolerance test get gestational diabetes. In most women, gestational diabetes goes away after pregnancy.

LITERATURE SURVEY

The dataset was originally published by National Institute of Diabetes and Digestive and Kidney Diseases. This chapter contains literature review related with supervised learning model and unsupervised learning model. various classification algorithms like

Logistic regression, decision tree, Random forest, support vector machine, k nearest neighbor, Naïve bayes and artificial neural network are used. This chapter also refers about the related research works carried out in the field of diabetic research using various machine learning tools and techniques.

Many people have developed various prediction models using data mining to predict diabetes [6]. Some of the models developed using data mining are as follows: Abdulla et al. [1] worked on predictive analysis of diabetic treatment using a regression based data mining technique Support Vector Machine. The datasets of Non Communicable Diseases (NCD) was analyzed for finding out the effectiveness of different treatment types for different age groups. The arrived at a conclusion that drug treatment for patients in the young age group can be delayed whereas; patients in the old age group should be prescribed drug treatment immediately.

K. Rajesh et al [6] carried out a research to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes. The training dataset used for data mining classification was the Pima Indians Diabetes Database they applied Different classification techniques and found out that c4.5 classification algorithm was the best algorithm to classify the da Yunsheng et al. [5] used KNN algorithm by removing the outlier/OOB (out of bag) and in this study the storage space was minimized. After removing a parameter which have less effect the researchers got better accuracy.

Nilashi et al. [3] used CART (classification and Regression Tree), clustering Algorithm (principal component Analysis (PCA) and Expectation maximization (EM) techniques. They found that Some fuzzy rules generated by CART by removing noise was effective in prediction purpose.

Velide Phani Kumar et al. [4] analysed diabetes data using various data mining techniques such as Naive Bayes, J48(C4.5) JRip, Neural networks, Decision trees, KNN, Fuzzy logic and Genetic Algorithms based on accuracy and time. They found that out of various data mining techniques J48 (C4.5) took least time.

Rupa Bagdi et al. [5] developed a decision support system which combined the strengths of both OLAP and data mining. This system would predict the future state and generate useful information for effective decision making. They also compared the result of the ID3 and C4.5 decision tree algorithms. The system could discover hidden patterns in the data and it also enhanced real-time indicators and discovered obstacle and it improved information visualization.

Arwa Al-Rofiyee et al. [9] in their research paper have focused on predictive analysis of diabetes diagnose using artificial neural network as a data mining technique. The Pima Indian diabetes database was obtained from UCI server and used for analysis. The technique that is applied is Multi-layer perception which is a classifier that uses back propagation to

classify instances. The WEKA software was employed as mining tool for diagnosing diabetes. This paper used Pima Indians Diabetes Dataset for their analysis. They have used data mining tool WEKA and Back propagation technique to predict diabetes. All the above researchers have been successful in analysing the diabetic data set and developing good prediction models. But most of them used tools like weka, Tanagra Rapid Miner and oracle data miner. In this paper an attempt is made to make analysis of diabetic data set using R.

PROPOSED WORK

The data set used for the purpose of this study is PIMA Indians Diabetes Database for women published by National Institute of Diabetes and Digestive and Kidney Diseases. It contains diagnostic information of women whose age is greater than 20. Information available includes 768 females, of which 268 females are diagnosed with diabetes. The samples consist of examples with 8 attribute values and one of the two possible outcomes, namely whether the patient is tested positive for diabetes (indicated by output one) or not (indicated by zero). This data set is evaluated using R.

The diabetic data set is given as input to the system, which is loaded in to R. The raw data is just a CSV file consisting of comma separated values, it just looks like a clutter of data. But a proper evaluation of this data set will reveal some interesting facts. The raw input is given as input to R, the data set is analysed and partitioned based on different attribute the output which is obtained from R-is well formatted data. R is one of the best languages which is used for statistical computing as well as for generating graphs. As we all know that pictures speak more than words, after evaluating the graphs are generated for each data set using R and then the data is plotted. The analysis of proposed work is shown in figure1. The proposed work is discussed in the following section.

3.1 ATTRIBUTES

The proposed work makes use of PIMA Indian Diabetes Data-set. Dataset of diabetic patients with minimum twenty-one-year age of Pima Indian population has been taken from UCI machine learning repository. This dataset is originally owned by the National institute of diabetes and digestive and kidney diseases. In this dataset there are total 768 instances classified into two classes: diabetic and non-diabetic with eight different risk factors: number of times pregnant, plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age as in Table1.

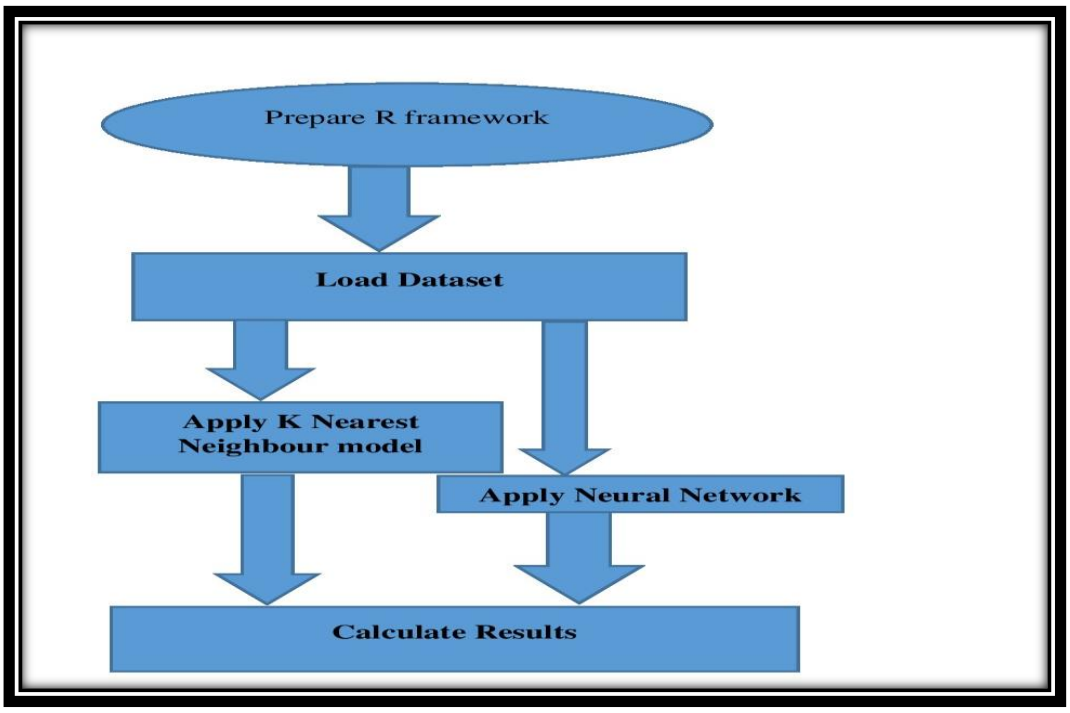

Figure 1:Proposed work

The 8 attributes that are defined here are-

Attribute ID	Attribute Definition
1	number of times pregnant
2	plasma glucose concentration after two hours in an oral glucose tolerance test
3	diastolic blood pressure (mm Hg)
4	triceps skin fold thickness (mm)
5	two-hour serum insulin (mu U/ml)
6	body mass index (weight in kg/ (height in m)^2)
7	diabetes pedigree function (It is a measure of the expected genetic influence of relatives on the subject's eventual diabetes risk)
8	age (years)
9	class variable (0 or 1)

Table 1: Diabetes Dataset Attributes.

3.1 PREDICTIVE MODELS

3.1.1 K Nearest Neighbour

The k-nearest neighbours (KNN) algorithm is a simple machine learning method used for both classification and regression. K-nearest Neighbour is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure [11]. The KNN algorithm predicts the outcome of a new observation by comparing it to k similar cases in the training data set, where k is defined by the analyst. KNN is a method which is used for classifying objects based on closest training examples in the feature space. A distance measure is needed to determine the "closeness" of instances. KNN classifies an instance by finding its nearest neighbours and picking the most popular class among the neighbours. The k nearest neighbour algorithm is simplest of all machine learning algorithms and it is analytically tractable. The model is build based on the relationship between predictors and outcome of the training set, then model specification is used to predict Dataset of female patients with minimum twenty-year age of Pima Indian population has been taken from UCI machine learning repository. This dataset is originally owned by the National institute of diabetes and digestive and kidney diseases. In this dataset there are total 768 instances classified into two classes: diabetic and non-diabetic with eight different risk factors: number of times pregnant, plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age as in Table1.

In KNN, the training samples are mainly described by n-dimensional numeric attributes. The training samples are stored in an n dimensional space. When a test sample (unknown class label) is given, k-nearest neighbor classifier starts searching the 'k' training samples which are closest to the unknown sample or test sample. KNN has been used in statistical estimation and pattern recognition. many distance measure is used like Euclidean distance, Manhattan distance, Minkowski distance.It is given by the equations

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Some of the Features of KNN are

a) All instances of the data correspond to the points in an n-dimensional Euclidean space.

b) Classification is delayed till a new instance arrives.

c) The Classification is done by comparing feature vectors of the different points in a space region.

d) The target function may be discrete or real valued.

For K Nearest Neighbor we compute the outcome for each test case by comparing that case to the "nearest neighbors" in the training set. When fitting the KNN algorithm, the analyst needs to specify the number of neighbors (k) to be considered in the KNN algorithm for predicting the outcome of an observation. To ensure we use a number for k that gives better model performance, for diabetes prediction in PIMA Indian Diabetic Dataset a two-part cross-validation is performed by varying the possible values for k from 2 to 10; second, the process is repeated by splitting the data into training and test sets 100 times to ensure a robust estimate of model performance for each k. finally the knn function is used within the class package and computed model accuracy on the test set for each fold. There are many libraries to do KNN analysis. for this dataset e1071 library is used.

In this model, we have used the following code:

```
library(tidyverse)
library(caret)
pima.train.norm <- decostand(pima.train.data, "normalize")
pima.test.norm <- decostand(pima.test.data, "normalize")
```

Implementing the K Nearest Neighbour Model

```
cv <- trainControl(method = "repeatedcv", number = 10, repeats = 3, classProbs = T, summaryFunction = twoClassSummary)

(knn.pima <- train(test ~ ., data = pima.train, method = "knn", preProcess = c("center", "scale"), trControl = cv, metric = "ROC", tuneLength = 10))

knn.pima <- knn(train = pima.train.norm, test = pima.test.norm, cl = pima.train.lab, k = 23, prob = TRUE)
```

This model is implemented using R studio and the result is shown in the figure 2.

3.2.2 Neural Network

Neural network is one the most popular machine learning algorithm, with wide area applications in predictive modelling and building classifiers. Presently, many advanced models of Neural Networks like Convolutional Neural Network, Deep learning models are

popular in the domain of Computer vision, Network security, Artificial intelligence, Robotics applications, Health care and many more advanced technologies. Few exciting facts which drive data scientists to use Artificial Neural Networks are: -

- Adapts and trains itself to complex non-linear problems.
- Flexible to various kinds of problem sets.
- Fundamentally compatible with real-time learning (Online Learning).

Within the field of machine learning n neural networks are a subset of algorithms built around a model of artificial neurons spread across three or more layers [26]. There are plenty of other machine learning model which is notable for being adaptive in nature. Every node of neural network has their own sphere of knowledge about rules and functionalities to develop it-self through experiences learned from previous techniques that don't rely on neural networks. Neural networks are well-suited to identifying non-linear patterns, as in patterns where there isn't a direct, one-to-one relationship between the input and output [27]. This is a learning training. Neural networks are characterized by containing adaptive weights along paths between neurons that can be tuned by a learning algorithm that learns from observed data in order to improve model. One must choose an appropriate cost function. The cost function is what is used to learn the optimal solution to the problem being solved [26]. In a nutshell, it can adjust itself to the changing environment as it learns from initial training and subsequent runs provide more information about the world. The number of neurons in hidden layers should be similar to the input neurons. If the number of neurons is large enough, that may increase performance but also may increase complexity. A trade-off is to be maintained for the same. Use of Momentum with backpropagation can help in convergence of solution, and achieve global optima. for diabetes prediction data in the PIMA Indian Diabetes Dataset is normalized to fit the neural network. There is no fixed rule to choose the number of hidden nodes, but as general rule, it should be 2/3 of the input nodes. model performs is to calculate the correlation between predictions and actual data. This model for predicting diabetes using R is implemented by using the following code:

```
pima.norm <- pima2 %>%
mutate_all(scale01)
```

Implementing Neural Network model

```
pima.size <- floor(0.75 * nrow(pima.norm))
train <- sample(seq_len(nrow(pima.norm)), size = pima.size)
pima.nn <- neuralnet(test ~ pregnant + glucose + diastolic + triceps + insulin + bmi + diabetes + age, hidden = 4, data = pima.train.n, linear.output = TRUE)
plot(pima.nn2)
predict.nn2<-compute(pima.nn2,pima.test.n[,1:8])$net.result
cor(predict.nn2, pima.test.n$test)
```

This model is implemented in R and result is shown in the figure 3.

EXPERIMENTAL ANALYSIS

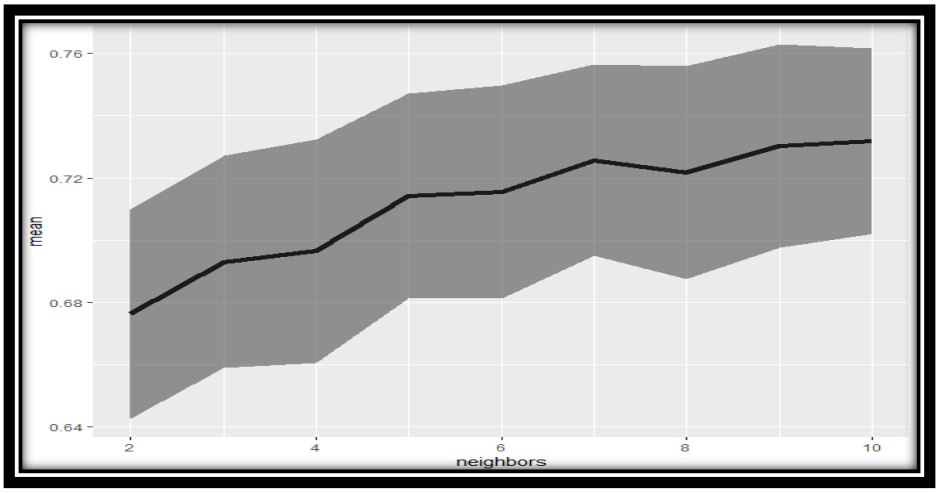

Figure 2: Plot using K Nearest Neighbour Model in R

Figure 2. KNN model performance accuracy for varying values of k. Black line indicates mean of all 100 folds for each value of k; grey ribbon indicates standard deviation. From this plot, we can see that k-nearest neighbours performs better for somewhat larger values of k, with performance reaching a maximum of about 73% classification accuracy. Though there is still some variance on the exact data split, using 9 or 10 neighbours seems to yield fairly stable model estimates on the test set.

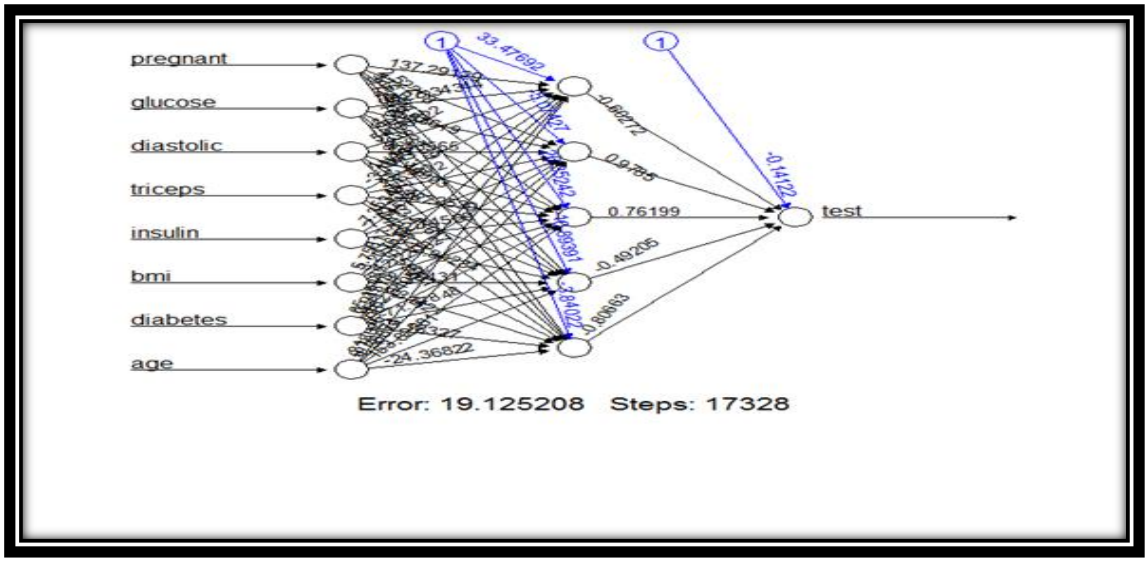

Figure 3: Neural Network generated to Predict Chances of diabetes in R

Figure 3 shows that by applying single layer architecture and by increasing more number of hidden nodes this model yields a saturated accuracy of 72% and an error rate of 19.125%.

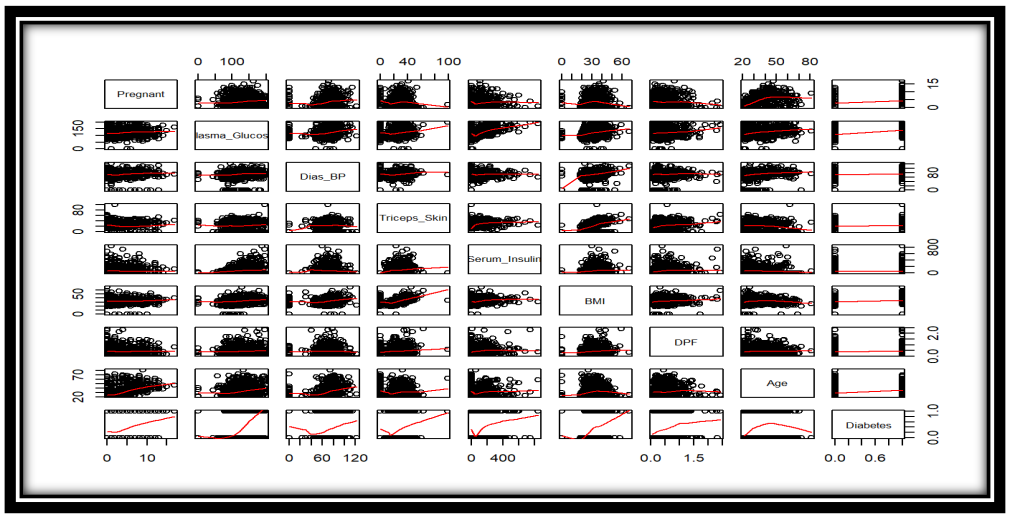

Figure 4: Matrix of Scatterplots

Figure 4 shows the matrix of scatterplots; it shows that there are no missing values in the data. sapply function is used to find the number of missing values in each columns.

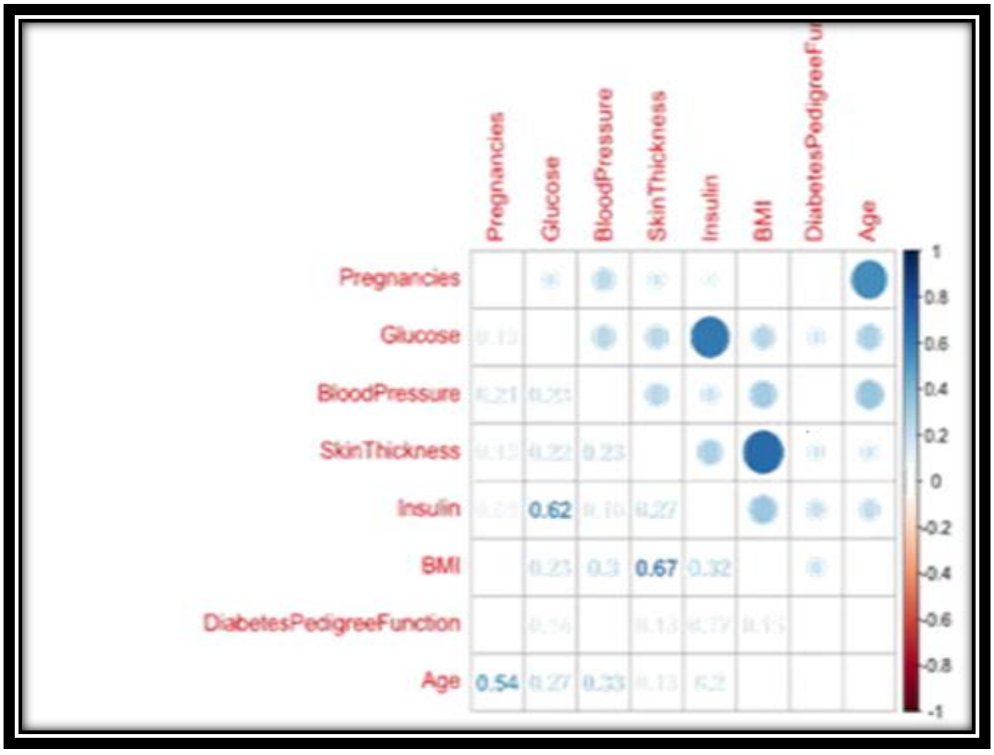

Figure 5: Matrix of correlation between variables

Figure 5 shows the matrix of correlation between variables. A correlation matrix is used to summarize data, as an input for more advanced analysis. for this analysis package corr is used in R tool.

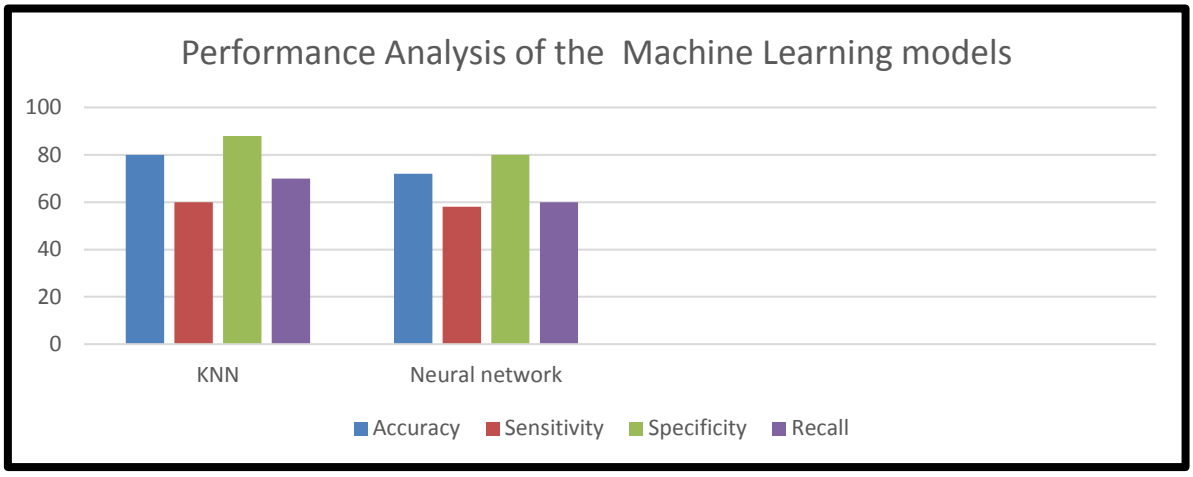

Figure 6: Performance Analysis of the Machine Learning models

MODELS	*ACCURACY*
K Nearest Neighbor	80%
Neural Network	72%

Table 2: Classification Accuracy of the Machine learning models

CONCLUSION AND FUTURE ENHANCEMENT

This paper focuses on analysis of diabetes in women through comparing various prediction models and find their accuracy with statistical implication using R. we have compared the classification of results using K Nearest Neighbor and Neural network. The classification results showed that K Nearest Neighbor gave the best results. The K Nearest Neighbor with increased classification performance also overcame the overfitting problem generated due to missing values in the datasets. Various data mining techniques and its application were studied or reviewed. application of machine learning algorithm is applied in different medical data sets. Machine learning methods have different power in different data set. In future some more additional parameters such as thirst, fatigue, frequency of urination can be added for improvement. The facts which were revealed during the process can be used for developing some prediction models and some other datasets can be added for the prediction of diseases.

ACKNOWLEDGMENT

We would like to thank god and acknowledge all the authors that provide significant help in the research of Diabetes.

REFERENCES

1. Abdullah A. Aljumah, Mohammed Gulam Ahamad, 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) 25, 127– 136.*
2. Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). *An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26-34.*
3. *Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering.*
4. Velide Phani Kumar, Lakshmi Velide*, "A data mining approach for prediction and treatment of diabetes disease" in international journal of science inventions today Volume 3, Issue 1, January,February 2014.*
5. Rupa Bagdi, Prof. Pramod Patil*, "Diagnosis of Diabetes Using OLAP and Data Mining Integration" in International Journal of Computer Science & Communication Networks, Vol 2(3), 314322.*
6. K. Rajesh, V. Sangeetha*, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.*
7. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). *Machine learning AND data mining methods in diabetes research. Computational and structural biotechnology journal.*
8. Mohammed Imran, Alhanouf M. AlAbdullatif, Bushra S. AlAwwad, Mzoon M. Alwalmani, Sarah A. Al- Suhaibani, and Shahad A. Al-Sayah*," Towards Early Detection of Diabetic Retinopathy Using Extended Fuzzy Logic" ,International Journal of Pharma Medicine and Biological Sciences Vol. 5, No. 2, April 2016.*
9. Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufad, Dr. Mohammed Abdullah AL-Hagery, "Using Prediction Methods in Data mining for Diabetes Diagnosis", *http://www.psu.edu.sa/megdam/sdma/Downloads/Posters.*
10. Polatkemal, SalihGüne*," An expert system approach bas ed on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", Digital Signal 17 (2007).*
11. *Han J, KanberM. Pei J, "Data Mining: Concepts and Techniques", 3rd ed. USA: Morgan Kaufman; 2012.*
12. ChandrakarOmprakash,Dr.kumarJatinder,SainiR.*,"Development of Indian Weighted Diabetic Risk Score (IW- DRS) using Machine Learning Techniques for Type-2 Diabetes",ACM COMPUTE '16, October 21-23, 2016.*
13. ButwallMani,kumarShradha*," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications (0975 – 8887) Volume 120 – No.8, June 2015.*

14. VeenaVijayan V.C.Anjali*," Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum.*
15. *"Analysis of a Population of Diabetic Patients Databases with Classifiers",* Murat Koklu and Yavuz Unal, *World Academy of Science, Engineering and Technology International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol:7 No:8, 2013.*
16. VarmaKamadiV.S.R.P,RaoAllamAppab,ThummalaSitaMahalakshmia,*"A Computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach", Applied Soft Computing.*
17. http://www.who.int/mediacentre/factsheets/fs312/en/.
18. Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). *Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.*
19. Balpande, V. R., & Wajgi, R. D. (2017, February). *Prediction and severity estimation of diabetes using data mining technique. In Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on (pp. 576580). IEEE.*
20. Quinlan JR, "Induction of decision tree". Machine Learning 1, Khuwer Academic Publisher, pp. 81-106,1986.
21. Saravananathan, K., & Velmurugan, T. (2016*). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).*
22. Pavate, A., & Ansari, N. (2015, September). *Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques. In Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on(pp. 371-375). IEE.*
23. Vijayan, V. V., & Anjali, C. (2015, December). *Prediction and diagnosis of diabetes mellitus—A machine learning approach. In Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in (pp. 122-127). IEEE.*
24. Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S. J., & Yu, K. S. (2015). *An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Systems with Applications, 42(9), 4265-4273.*
25. Al Jarullah, A. A. (2011, April). *Decision tree discovery for the diagnosis of type II diabetes. In Innovations in Information Technology (IIT), 2011 International Conference on (pp. 303-307). IEEE.*
26. *Saimadhu P. How the Random Forest Algorithm Works in Machine Learning. Published on May 22, 2017.*
27. *Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507.*