International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Diagnosis of Diabetes by using Data Mining Techniques

Sachin Kumar, Narender Kumar

Abstract: ***There are many classifiers that are used for diagnosis of diabetes but the result of this paper shows that how logistic regression having best accuracy among the other classifiers. Logistic regression removes the disadvantages of linear regression. There are different classifiers that are used for prediction. In the worldwide millions of peoples are suffering from diabetes according to WHO report. In the medical region, many researches have done with the help of data mining. The aim of this paper is to diagnosis of diabetes by using the best classifiers and providing best parameter tuning. The study helps to find whether a patient is enduring from diabetes or not using classification methods and it further investigate and evaluates the functioning of different classification in relations of precision, accuracy, recall & roc.***

Keywords : Diabetes, KNN, Classification, Decision tree, logistic regression, SVM.

I. INTRODUCTION

Data mining [1] is meant to collect the useful data from a huge amount of dataset which means that extracting useful information or pattern from a large amount of dataset. There are many alternative names for the data mining like KDD (Knowledge discovery from data), extraction of knowledge, knowledge mining from data, pattern discovery etc.

Decision tree [1] is the simplest one classification technique used for the prediction. In past years the decision tree is the widely used technique for prediction but in today worldwide it is one of the least used technique. The decision tree contains the number of nodes, branches and a root node. There are two types of models:-supervised and unsupervised model. For splitting the best tree there is a formation of selecting the best attribute as compared to all other attributes using Gini-index, information gain and gain ratio.

KNN (K-Nearest Neighbour) is a classification technique. By using KNN, finding the distance between the two or more point by using the Euclidian distance formula. Target value used in KNN is categorical type in which the target value is already categorized into predefined classes or category like income can be categorized into high income, low income, or moderate income [1].

Revised Manuscript Received on November 30, 2019.
* Correspondence Author
Sachin kumar*, M.Tech, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana India.
Mr. Narender Kumar, Assistant Professor, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana India.

SVM (SUPPORT VECTOR MACHINE) support vector machine [1] is helpful for solving the complex problems. SVM is used for real-world problems, handwritten recognition, bio sequence analysis and some other classification. Data preparation is not found essential, in most of the cases and parameter tuning are also available by default.

SVM is known to be a supervised Machine Learning (ML) algorithm that could be applied in classification and regression techniques of data mining. For the most part utilized for classification issue.

NEURAL NETWORK [1] is actually based on neurons. The Artificial neural network is used to work as a human brain and it produces the artificial neuron as in human. Approximately near to 100 billion of neurons works in the human brain to instruct or to control the human body. Every neuron has a connection with rest of other neurons.

In artificial neural network commonly three neuron layers works to become it like human brain. These layers are input layer, hidden layer, and output layer.

BAYESIAN CLASSIFICATION in the Bayesian classification, we are predicting the class membership probability which decides the given tuple belonging to which class. Bayesian classification depends on Bayes theorem [1].

In byes theorem:

X- Data tuple $\mathbf{P}\left(H/_X\right) = \frac{\mathbf{P}\left(H/_X\right)\boldsymbol{P}\left(H\right)}{\boldsymbol{P}(X)}$

H- Hypothesis

Where H is known to be hypothesis s.t. data tuple X ∈ specified class (C)

As identified p(X/H).

II. OVERVIEW ON DIABETES

According to WHO report diabetes in worldwide is common diseases and millions of people are suffering from diabetes and in future, there is number of peoples will suffer from diabetes. In diabetes, when our body is incapable to produce insulin then glucose in our body will increase and generally we eat that part of the food is converted into glucose. This whole amount of glucose is incapable to convert into insulin so our body contains more amount of glucose and it is dangerous for our body [15].

Diabetes is of 2 types:

Type 1

When we say that the person is suffering from type1 diabetes it means that the person body does not properly produce insulin and the type 1 diabetes is found in the adult age of the patient. According to the WHO report, only 10% of the patient are the victim of this type diabetes [15].

Type 2
In type 2 diabetes disease, the human body has not capable enough to convert glucose into insulin but not in sufficient amount. Type2 diabetes causes the patient to die due to less amount of insulin in the patient body. Generally, patients are suffering from type2 diabetes [15].
Gestational diabetes it is also a type of diabetes. Only female patients are suffering from that type of diabetes. It is the temporary state of diabetes and mainly occurs during pregnancy. If we are not caring about this type of diabetes then the several problems occur in the patient body like damages of eyes, kidney failure etc [15].

III. RIVEW OF LITERATURE

C4.5 is a decision tree variant. Decision nodes[1] are used to test the attributes of dataset. Output produced by the decision tree is in the form of 'yes' or 'no'. It is the simplest one classifier used for prediction and easy to understand. According to the study of Felix Tamin[6] et al. in their study dataset is collected from the UCI ML data set. Before implementation data set first applied to the pre-processing phase. A) Removal of missing values: for numeric data type and nominal data type they are using mean and mode value respectively. B) Attribute selection: unwanted and useless attributes were supposed to be deleted. They took five category of the data sample and used 5 various type of pre-processing on them and find the results. Based on their analysis and experiment C4.5 approach could be used to expect re-admission rate of diabetes patient with the accuracy of 74.5%. mira kania et al.[7] in their examination, they are concentrating on early recognition of type two diabetes, mellitus by using tree regression, random forest and classification methods. They are utilizing the investigation of the combination of classifiers which enhance the accuracy of the single classifiers. Single classifier gives 77% exactness and after combination with random forest, it will give 83.8%average exactness. Asha A Aljarullah et al. [9] in their investigation, they are concentrating on type 2 diabetes. They are dealing with a two-step process. In the initial step pre-processing was done to enhance the accuracy of the classifiers and in the second stage connected the classifier to the enhanced dataset. The accuracy get through the modified dataset for J48 decision tree was 78.17%. K. Rajesh et al. [11] they started their research to predict about patient that in a given dataset they are suffering from diabetes or not. For this, they collect the dataset from the UCI machine learning repository which contains 768 tuples and nearly eight various persistent attributes with various classification algorithms used, RAND will give maximum accuracy but the rules are applied on RAND algorithms are very large due to this it suffers from over fitting problem. After this C4.5 has come back with most extreme accuracy, almost about 91%. But when feature relevance technique was utilized then the accuracy was dropped to 88%. So the analysts prescribed C4.5 as the best classifier, while the RND tree has problem of over fitting issue. Purushottam et al. [12] in their work or research they are using 2 phase process. In the 1st phase, they efficiently predict the risk level of patient and in 2nd phase compare the evaluated model to C4.5, partial tree. Measure the performance in terms of accuracy and classification error. The dataset was collected from the Pima Indian Diabetes Database which contains 768 tuples and each tuple having 9 attributes. The rule is made up and tested under the proposed model and finally get 81.27% accuracy. Yoichi Hayashi et al.[17] recursive - rule extraction with J48 joined with sampling selection techniques for analysis of type2 diabetes utilizing Pima Indian dataset . Finally, the accuracy comes is 82.80%.
K-Nearest Neighbour [1] is an instance-based learning. K stands for how many nearest neighbour are used for prediction. The k of value varies from 1 to k-1. According to knn theory we just comparing the unclassified data to already classified data by measuring the distance between them like Dr. Zubair et al.[4] used k-nearest neighbour for diagnosis of diabetes. The dataset is collected from Stanford.edu. In their proposed work they have divide dataset into training dataset which consists of 100 rows and 11columns. Out of 100 records, 50 records are utilized as test information and after that applied Knn. For various k values, different accuracy comes. For the estimation of K=3 and K=5 finds the error rate by using MATLAB and they find that for K=3 accuracy comes under 70% and for K=5 it will come under 75%.so they give a conclusion that if we increase the value of k accuracy also increase KNN is used to find the closeness between the points. Emrana Kabir Hashi et al. [16] in their examination they are they are focusing on prediction of disease by using classification technique. The classifiers utilized by them are C4.5, and KNN and partition the dataset into 2 sections trained and test dataset into the proportion of 70:30. For the testing part, C4.5 and KNN having accuracy are 90.43% and 76.96% respectively.
Support Vector Machine [1] is a supervised learning model. It means first we trained the dataset and then we apply for the test dataset. In some cases, the dataset is not normally distributed means that skewed data was present. This is the disadvantage of SVM. Many of researchers apply their dataset on SVM like Bradley A. P. et al. [5] in their research study they are using SVM classifier and in their result, they find that the combined model gives the maximum accuracy as compared with other classifiers. The dataset was gathered from the UCI machine learning repository and some different medical clinics.
The Artificial neural network [1] is used as a human brain. It is one of the simplest definition and building blocks are neurons. There are 100 billions of neurons are available in the human mind. Every neuron has an association with rest of different neurons that are available in the human cerebrum. There are three neuron layers present in the neural system that are the input, hidden and output layer. Sonu kumari et al. [2] they are focusing on a data mining approach for diagnosis of diabetes.in their study neural network with back propagation classifier is used. In their dataset around 100 tuples are present. A Neural network consists of 28 nodes in which 13 input node, 13 hidden nodes and 1 output node. The result shows that the accuracy is 92.8%. Tanja Dujic et al. [3] in their proposed work artificial neural network was used to study the classification. Their main focus was on Type2 diabetes and pre-diabetes. How can control or detect these disease. For that, they are using feed forward artificial neural network which contains 2 layered architecture. And in the artificial neural network, there will be a hidden layer also.

So they additionally deal with the hidden layer and fix the estimation of neurons present in the hidden layer and the value comes out is 15. Neural network tested 2 important parts of the disease one is glucose level and another one is an HbA1c test and finally get 94.1% accuracy for pre-diabetic and 93.3% for Type2 diabetes. Dr.B.L.Shivakumar et al. [8] in their examination, they are concentrating on the overview of technologies used for diagnosis and prediction of diabetes. Procedures used for diagnosis are the neural network, decision tree, clustering, association rules are utilized.

Veena Vijayan et al. [13] they propose the decision support system and the decision stumps in his theory. For this study, he took his data set from the UCI Machine Learning Repository, which contains 768 pieces and every piece has 9 attributes and used some local datasets from Kerala. By using the Adaboost-decision stumps, they get 80.72% of Accuracy, and for Future Work, they said that the Accuracy of the Decision Stump can be improved by using other classifiers. The other classifiers use the same kind as Neural Network, K Nearest neighbour and some ensembles of the classifiers.

loannis et al.[14]Their study shows that machine learning algorithms are the best algorithm in diagnosing any disease. He used SVM, LOGISTIC REGRESSION, NAIVE BAYES his study. Using the cross validation, he kept the value of the fold 10. After comparing the accuracy of all the classifiers, it came to know that SVM is the best of all the classifiers.

IV. EXPRIMENT SETUP FOR LOGISTIC REGRESSION

Logistic Regression [1] is named for the function used at the core of the method, logistic function. Logistic function, also known as sigmoid function. It's an S-shaped curve graph that can take all real-valued numbers and map it into a value between 0 and 1. Data Preparation for logistic regression:

Binary Output Variable: It is knowing for binary classification problem. Gaussian distribution: Logistic regression known as linear algorithm. Expecting a direct connection between the input factors with an output. By using log, root, Box-Cox or other univariate transforms. Remove highly correlated variables: overfitting problem arise when the highly correlated input are present.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
Pregnancies	1.000	0.129	0.209	0.083	0.056	0.022
Glucose	0.129	1.000	0.218	0.182	0.408	0.218
BloodPressure	0.209	0.218	1.000	0.193	0.073	0.281
SkinThickness	0.083	0.182	0.193	1.000	0.158	0.542
Insulin	0.056	0.408	0.073	0.158	1.000	0.167
BMI	0.022	0.218	0.281	0.542	0.167	1.000
DiabetesPedigreeFunction	-0.034	0.137	-0.003	0.101	0.099	0.153
Age	0.544	0.264	0.325	0.128	0.137	0.026
Outcome	0.222	0.467	0.166	0.215	0.214	0.312

Table- II: Correlated value of attributes

Performing the pairwise connections between every inputs and expelling extremely correlated inputs. We find that the age and pregnancy is highly correlated. So we can drop the pregnancy variable, otherwise over fitting can occur.

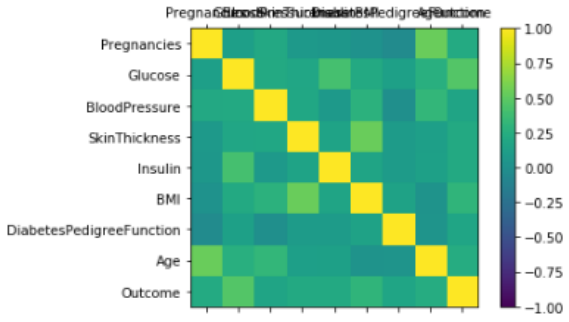

Fig. 1. Correlated values

Since Age, Insulin are not normally distributed and are right skewed distribution because the mean is right hand side of the median/peak Transform Age, Insulin and Diabetes into normal.

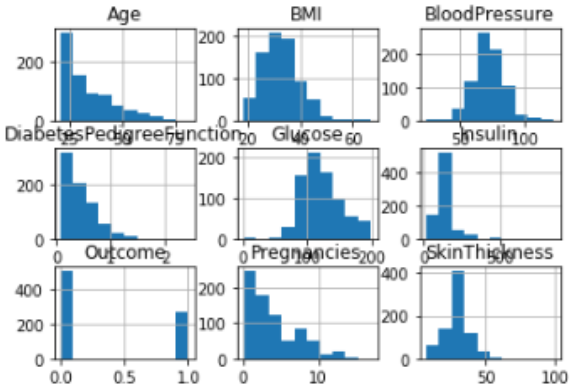

Fig. 2. Before normal distribution .

Since the transformation to Age is not showing normal distribution so perform box-cox transformation on that variable.

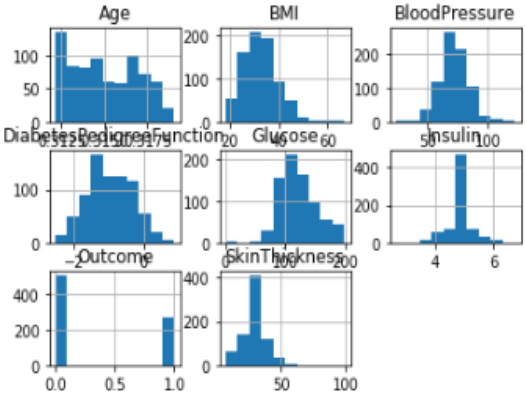

Fig. 3.After normal distributiion.

Now our data is ready for use and applying the logistic regression model.

A. *KNN CLASSIFICATION*

After applying the algorithm KNN on the given dataset we finds that during the parameter tuning the many parameters are having like:

KNN algorithm gives Tuning hyper-parameters for precision

Weight= 'uniform', neighbour= '5', leafsize = '10', algorithm = 'balltree', numberofjobs ='-1'.

KNN algorithm gives Tuning hyper-parameters for recall

Weight= 'distance', neighbour= '9', leafsize = '10', algorithm = 'balltree', numberofjobs ='-1'.

KNN algorithm gives Tuning hyper-parameters for accuracy

Weight= 'uniform', neighbour= '8', leafsize = '10', algorithm = 'balltree', numberofjobs ='-1'.

KNN algorithm gives Tuning hyper-parameters for roc_curve

Weight= 'distance', neighbour= '9', leafsize = '10', algorithm = 'balltree', numberofjobs ='-1'.

After applying KNN, the Accuracy is 76.7%. The accuracy of KNN depends on the distance between its dataset. Due to the fact that there is a lot of values missing in the dataset, its accuracy has decreased. Apply KNN to real dataset, then the accuracy can be quite good. Mixture of unalike parameters gives the best result after missing values from the dataset.

B. ***Decision tree algorithm gives Tuning hyper-parameters for precision***

Max_samples_split = '2', max_depth = '1'.

Decision tree algorithm gives Tuning hyper-parameters for recall

Max_samples_split = '4', max_depth = '9'.

Decision tree algorithm gives Tuning hyper-parameters for accuracy

Max_samples_split = '2', max_depth = '6'.

Decision tree algorithm gives Tuning hyper-parameters for roc_curve

Max_samples_split = '8', max_depth = '4'.

In the decision tree having less accuracy as compared to the benchmark result, the reason for the decrease in the accuracy of the decision tree is that in this study there has used of the CART decision tree variant classifier. Maximum accuracy of CART comes under when the maximum depth of the decision tree is 6 and maximum split size is 2. The reason for considering the split size and depth of the dissection tree was that the over fitting did not come.

Accuracy comes under SVM is 76.3%. Parameter selection should be best otherwise the performance of the SVM is not so good and the dataset also responsible for the performance. SVM is used for complex problems so that for the complex problems it works better than KNN and other classifiers. Size of training data also influences the performance of the classifier. Accuracy comes under naïve Bayes is 76.7% reason being for that is the assumptions are independent of class. Easy to implement and gives better result in some cases. But in most of the cases the performance will decrease due to independence in the attributes.

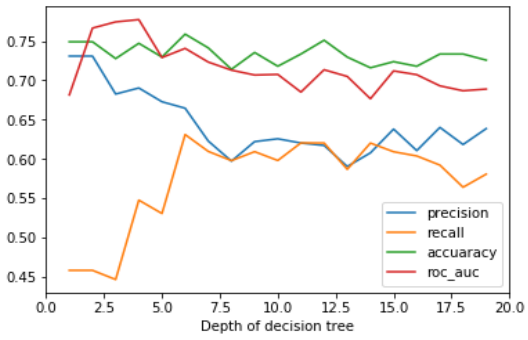

Fig. 4.Decision tree line graph

In fig4 the four terms compared are precision, recall, accuracy and roc for decision tree. When we increase the depth of decision tree means the number of nodes are increasing in the decision tree then the accuracy will become the best term among the remaining terms. Average value of accuracy will come out is 76.5%.

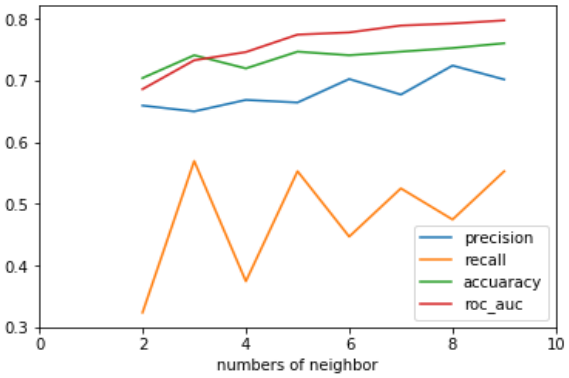

Fig. 5.Knn line graph

Fig5 is used for knn classifier. The result shows that when the number of neighbour are increasing then the accuracy will also increase but the roc_auc curve will give the best result among all the remaining terms. The neighbour value is starting from 2.

V. EXPERIMENTAL RESULT

Accuracy of any model is calculated by:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

Recall of any model is calculated by:

Recall = TP÷(TP+FN)

Precision of any model is calculated by:

Precision = TP÷(TP+FP)

TP (TRUE POSITIVE): true case and predicted correctly. TN (TRUE NEGATIVE): negative case and predicted incorrectly. FP (FALSE POSITIVE): negative case and predicted correctly. FN (FALSE NEGATIVE): positive case and predicted incorrectly.

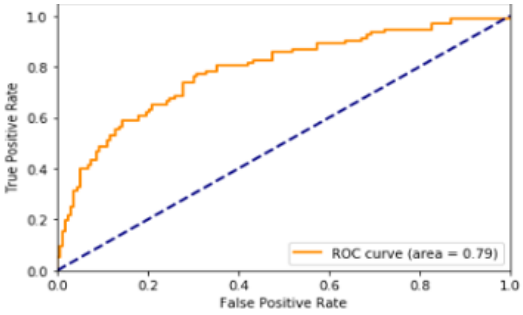

Fig. 6.Logistic regression line graph

Experimental result				
Type of algorithm	**Accuracy**	**Recall**	**Precision**	**Roc**
Decision tree	0.765(+/-0.70)	0.632(+/-0.146)	0.731(+/-0.158)	.0781(+/-.084)
KNN	0.767(+/-0.11)	0.720(+/-0.076)	0.761(+/-0.119)	0.803(+/-0.66)
Logistic regression	0.818(+/-0.09)	0.804(+/-0.07)	0.815(+/-0.12)	.794(+/-0.11)
Naïve Bayes	.767(+/-0.12)	.755(+/-0.76)	.776(+/-0.79)	.798(+/-0.05)
SVM	0.763(+/-0.11)	.743(+/-0.45)	.757(+/-0.79)	.747(+/-0.34)

VI. CONCLUSION AND FUTURE WORK

In the study, use of decision tree, logistic regression, naïve Bayes, SVM and KNN to diagnose diabetes. For this, the UCI machine learning repository data set was used. If feature selection in pre-processing and tuning parameters was done in decision tree and KNN, then our result becomes good. If we compare the proposed logistic regression to other classifier then we find that the logistic regression having maximum accuracy which was 81.8%. On the other hand the other classifiers that are used in the classification problem are having good accuracy but there are many problems that are facing during the classification. SVM having lowest accuracy which was 76.3%. For the complex classification problems SVM is used. Value of recall and precision for logistic regression is highest.

In futures work may include removal of disadvantages/demerits of logistic regression. Logistic regression covers the disadvantages of regression model. We can also increase the value of classifiers. Different types of disease dataset will be used on logistic regression. Distinct classifiers will be joined and furthermore parallel calculation will be researched assist for future investigation.

REFERENCES

1. "Data Mining and Predictive analytics, 2nd edition" Daniel T.Larose, Chantal D.Larose.
2. Sonu Kumari et al. "A Data Mining Approach for the Diagnosis of Diabetes Mellitus" Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013), pp.373-375, 2013.
3. Tanja Dujic , Dijana Sejdinovic et.al "Classification of Prediabetes and Type 2 Diabetes using Artificial Neural Network", CMBEBIH springer singapore ,pp.685-689, 2107.
4. Dr. Zubair Khan, Shefali Singh, Krati Saxena, "Diagnosis of Diabetes
5. Mellitus using K Nearest Neighbour Algorithm" ,International Journal of Computer Science *Trends and Technology(IJCST)* vol. 2, no. 4, 2014.
6. Barakat N., Bradley A. P., & Barakat M. N. H. "Intelligible support vector machines for diagnosis of diabetes mellitus". *IEEE transactions on information technology in biomedicine*, vol.14, no.4, pp.1114-1120, 2010.
7. Ramzan, M. "Comparing and evaluating the performance of WEKA classifiers on critical diseases" *In Information Processing (IICIP), 2016 1st India International Conference on*, vol.5, no10, pp. 1-4. IEEE.
8. Mira Kania Sabariah et al. , "Early Detection of Type II Diabetes Mellitus with Random Forest and Classification and Regression Tree (CART)", *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pp.238-242, 2014.
9. Dr.B.L.Shivakumar, S. Alby, "A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes", *2014 International Conference on Intelligent Computing Applications*, pp.167-173, 2014.
10. Asha A Aljarullah, "Decision Tree Discovery of the Diagnosis of Type II Diabetes". *Proc. of the International Conference on Innovations in Information Technology*, pp. 303 -307, 2011.
11. C M Velu, K R Kashwan, "Visual Data Mining Techniques for Classification of Diabetes Patients". Proc. of the IEEE 3rd International Advance Computing Conference, pp. 1070-1075, 2013.
12. K.Rajesh and V.Sangeetha. "Application of Data Mining Methods and Techniques for Diagnosis" *International Journal of Engineering and Innovative Technology (IJEIT)*, vol.2, no.3, pp.115-121, 2012.
13. Purushottam , Dr.Kanak Saxena , Richa Sharma, " Diabetes Mellitus Prediction System Evaluation Using C4.5 rule and Partial Tree" , IEEE.
14. V.Veena, Vijayan, and C.Anjali. "Prediction and Diagnosis of Diabetes Mellitus-A machine LearningApproach"*IntelligentComputational System(RAICS),IEEE Recent Advance in* , pp.122-127,IEEE 2015.
15. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, "Machine learning and data mining methods in diabetes research" *Computational and structural biotechnology journal.* , vol.9, no.2, pp.1232-1242, 2017.
16. https://www.medicalnewstoday.com/info/diabetes.
17. Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques",International Conference on Electrical, Computer and Communication Engineering (ECCE),pp.396-400,IEEE, 2017.
18. Yoichi Hayashi, Shonosuke Yukita, " Rule ectraction usingRecursive-Rule ectraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type2 diabetes mellitus in the Pima indian dataset", Informatic in Medicine Unlocked, pp.92-104, elsevier, 2016.

AUTHOR PROFILE

Sachin kumar received his M.Tech degree in department of Computer Science and Engineering from Guru Jambheshwar University of Science and Technology, Hisar, Haryana(India).

Mr. Narender Kumar is an Assistant Professor in department of Computer Science and Engineering in Guru Jambheshwar University of Science and Technology, Hisar, Haryana(India). He has received his B Tech degree in Computer Science and Engineering from National Institute of Technology Hamirpur, H.P. (India) in 2005 and M Tech degree from Deenbandhu Chhotu Ram University of Science and Technology Murthal, Sonepat, Haryana (India) in 2011. His area of interest are Data Mining and Machine Learning..