

BLOOD SUGAR DETECTION USING DIFFERENT MACHINE LEARNING TECHNIQUES

Jeevan Kumar, Rajesh Kumar Tiwari and Vijay Pandey

ABSTRACT:

Blood sugar, or glucose, is the main sugar found in your blood. It comes from the food you eat, and is your body's main source of energy. Your blood carries glucose to all of your body's cells to use for energy. The presence of glucose helps ensure that this primary source of nutrition for infants is palatable and acceptable. There are several machine learning algorithms in order to predict blood-sugar. Blood-sugar prediction is a challenging task. The proposed algorithm may help doctors for decision making. This will also help for knowing about health and next treatment of the patient. The predictive analytics in healthcare is discussed and six different machine learning algorithms are used. Comparison and discussion of the accuracies and performances of the proposed algorithms is mentioned in this paper.

Keywords: Machine learning, big data, blood-sugar, predictive analytics, health care, deep-learning, regression, classification.

Reference to this paper should be made as follows:

Jeevan Kumar, Rajesh Kumar Tiwari and Vijay Pandey, (2021), "BLOOD SUGAR DETECTION USING DIFFERENT MACHINE LEARNING TECHNIQUES" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 23-33, DOI 10.30696/IJEEA.IX.III.2021.23-33

Biographical notes:

Jeevan Kumar has completed B. tech and M. Tech from NIT Patna. He is pursuing PhD from Jharkhand University of Technology, Ranchi, Jharkhand. He is Asst. Professor in CSE Dept. R.V.S. College of Engineering and Technology, Jamshedpur. Jharkhand, India.

Rajesh Kumar Tiwari completed his PhD from BIT Mesra, Ranchi in 2010. He is Professor in CSE Dept. R. V. S. College of Engineering and Technology, Jamshedpur, Jharkhand, India.

Vijay Pandey completed his PhD from BIT Mesra, Ranchi in 2008. He is Director(Curriculum), at Jharkhand University of Technology, Ranchi, Jharkhand.

1. INTRODUCTION

A chronic illness blood-sugar is a group of metabolic diseases due to high sugar content in blood over long periods. There are two types of disorders owing to blood-sugar. Blood-sugar Mellitus being the metabolic disorder where Type-1 being the case in which pancreas can't produce insulin and Type-2 in which the body don't respond to the insulin, both of which lead to high blood sugar. According to World Health Organization (WHO) data, we can summarize the facts as,

- The number of people suffering from diabetes have increased from 108 million in 1980 to 422 million in 2014.
- The disease is increasing in low- and medium-income countries.
- Diabetes is the major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation.
- In 2016, approximately 1.6 million deaths were due to diabetes and this approximation is estimated to rise up to 2.2 million for the year 2022 due to high blood glucose levels.
- Diabetic retinopathy is an important cause of blindness, and occurs as a result of long-term accumulated damage to the small blood vessels in the retina. 2.6% of global blindness can be attributed to diabetes. This serious illness can be reduced if it can be predicted early. This paper uses six classification algorithms to predict the diabetes.
-

Machine learning (ML) is a type of artificial intelligence (AI) that allows authors to become more accurate at predicting outputs without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. There are different ways an algorithm can model a problem based on the given input data. Input data can be used for training the machine.

Algorithm which is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. We may categorize these algorithms as supervised machine learning algorithms.

Algorithms which are prepared through Input data that is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity. We may categorize these algorithms as unsupervised machine learning algorithms.

Many algorithms which are prepared through Input data that is in both labeled and unlabelled.

2. BACKGROUND

Machine Learning is a most active domain of Artificial Intelligence. Research scholars are working in medical, finance, defense, social networking and other area of research using machine learning. In this proposed research, we have used different algorithms of machine learning for the blood-sugar disease.

Velijayan, V.V. et al. (2015) proposed an application for tracking the location i.e. SensTrack which is used with smart phones embedded with Wi-Fi facility in order to reduce the usage of GPS due to its availability at high cost with negative impacts on battery in very short period. SensTrack operated the GPS sample by using stored information and can switch the location and to re-build the track route from recorded location Gaussian Process Regression approach is followed.

Fatima, M. and Pasha, M (2017) forecast students' performances using machine learning techniques (e.g., C4.5, sequential minimal optimization (SMO), Naïve bayes, 1-NN (1-Nearest Neighborhood), and MLP (multi-layer perceptron) with input features (e.g., gender, income, board marks and attendance). They applied correlation-based feature selection (CBFS) techniques to improve the model performances and determined that SMO achieves a higher effective average testing accuracy (66%) than do other methods.

K Sowjanya et al. (2015) employed artificial neural networks (ANNs) to predict student's performance. These models achieved high accuracy (85%) using input features such as grades, periods of study and school scores.

Huang and Fang (2013) performed a study that used machine learning techniques to predict student academic performance in engineering courses. In this study, the input features included course grades from all semesters and the output variable was exam scores. The researchers observed that SVMs are suitable for predicting an individual student's performance and that multilinear regression is suitable for forecasting the performance of all students in a course

Alberto M. C. Soza and Jose` R. Amazonas (2015) has been implemented as Principal Component Analysis (PCA) based clustering algorithm for fault detection that used Hadoop Framework and Mahout implementation. This algorithm integrated with IOT architecture implemented by the LinkSmart middleware. Proposed implementation and architecture increased the potential and functionality of IOT LinkSmart middleware [12].

Hardi Desai and et al. (2017) has proposed a vision to implement an affordable and compatible IOT based wireless sensor network in order to monitoring and analyzing the grocery levels at supermarkets as well as at homes. This system also provides an immense to use as future scope in the kitchens and to monitor the different storage places to manage the commodities in smooth manner [13].

Vahdat et al. (2015) used process mining (PM) and complexity matrix (CM) methods to analyze the relationship between grades and students' learning processes using DEEDS data. They concluded that the average student grades are positively correlated with the CM and that difficulty is negatively correlated with the CM. In addition, they determined that process discovery using PM and CM models provides valuable information regarding student learning processes.

Recently, an early predictive model was developed using student demographic, LMS data, and aptitude-related features. The authors developed a learning analytic system with an applied LR model that sent emails to high-risk students (Jayaprakash et al. 2014).

3. METHODOLOGY

Whichever Machine Learning classification algorithms is used to predict diabetes, the first step is to preprocess the data. Thereby on application of every classification algorithm, accuracy is noted and compared to predict which algorithm performs better for prediction of diabetes.

3.1 ANALYSIS OF DATASET

A Large data set has to be considered to get good prediction. In this paper PIMA Indians diabetes dataset, which has 781 records is taken. Figure:1. shows the actual dataset used for this model in csv file. The dataset is split into 70-30 ratio. The 70 percent of the training data is fed to the classification algorithms and the remaining 30 percent data is testing data.

The description of the columns of the dataset is below:

- Pregnancies: In the lifetime of a woman number of pregnancies.
- Glucose: In an oral glucose tolerance test, it is Plasma glucose concentration after 2 hours.
- Blood Pressure: It is Diastolic blood pressure in mm Hg.
- Skin Thickness: It is Triceps skin fold thickness in mm.
- Insulin: It is 2-Hour serum insulin in mu U/ml.
- BMI: It is calculated as $(\text{weight in kg}/(\text{height in m})^2)$
- Diabetes Pedigree Function: History of the family that reveals how many people having this diabetes.
- Age: Age of the person in years.

A	B	C	D	E	F	G	H	I
Pregnancies	Glucose(mg/dl)	BloodPressure (mmHg)	SkinThickness(mm)	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1

Figure: 1. Screenshot showing structure of Dataset in Excel Sheet.

3.2 DATA PREPROCESSING

Preprocessing of the data used to transform the raw data into useful and efficient format in the field of Data Mining. Actually, around us the data is incomplete with missing and lacking some values. In lacking certain attributes or containing only aggregate data. The data is noisy if it is containing errors or outliers and leads to inconsistency.

In order to get better result and accuracy, the dataset has been cleaned. A check has been done to identify if there is any correlation among the attributes of dataset. The outcome ‘true’ is replaced with ‘1’ and the outcome ‘false’ is replaced with ‘0’. If there is any missing and zero values; they are replaced with their mean. The data now is preprocessed.

3.3 SYSTEM ARCHITECTURE

Jupyter notebook is used in the system which sets the path to read the data set. Before preprocessing is done, the data set is read from the data base. After preprocessing, in the form of training and testing data it is splitted. Now to train the model, this training data is used. On applying the test data on trained model of each algorithm, accuracies of each algorithm are recorded. By comparing accuracies of each algorithm, the best algorithm to detect diabetes is found.

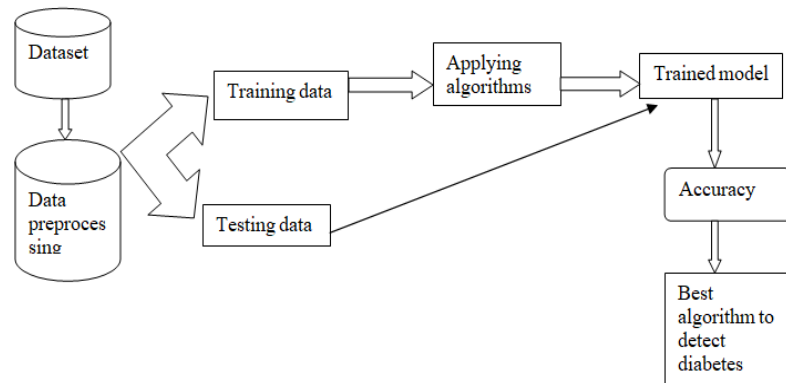


Figure: 2. System Architecture of the proposed system

3.4 CLASSIFICATION ALGORITHMS

The process in order to recognize, understand, and group the ideas and objects into different classes or sub-populations is called Classifications. Machine learning algorithms are used to classify future datasets into categories using pre-categorized training datasets. Using these algorithms, things like sentiment analysis to categorize unstructured data by polarity of opinion can be performed.

3.4.1 DECISION TREE ALGORITHM

This algorithm can be used for both Classification and Regression problems which is supervised learning algorithm. This Decision tree algorithm is used mostly in Classification problems. Here, internal nodes represent the test of an attribute, branches represent the outcome of an attribute and each leaf node in it represents the class label. So, it is a hierarchical-structured classifier. For predicting the class of each example, the algorithm starts from the root node of the tree. Decision tree follows the below steps:

Step-1: Let root node be S. Start from it, which contains the complete given dataset.

Step-2: Using Attribute Selection Measure (ASM) find the best attribute in the dataset to take it as root node.

Step-3: For the next branches, S has to be divided into subsets that contains possible values for the best attributes.

Step-4: In order to continue tree building, generate the decision tree node, which contains the best attribute.

Step-5: Repeat this process recursively until the leaf node cannot be further classified as a node.

3.4.2 NAIVE BAYES

This algorithm is based on Bayes theorem and is used for solving classification problems and is a supervised learning technique. In high-dimensional training dataset, it is mostly used for data classification. In making quick predictions Naive Bayes Classifier is one of the most effective

Classification algorithms. Some of the examples of Naive Bayes Algorithm are Spam filtration, Sentimental Analysis and classifying articles. Gaussian, multinomial and binomial are three types of naïve bayes models.

3.4.3 LOGISTIC REGRESSION ALGORITHM

Using numerical and categorical predictors, the Logistic Regression Algorithm is a generalized linear model which is used to model a binary categorical variable. It is one of the models which is used to predict the probability of a certain class or event as binary, such as pass or fail, win or lose, one or zero.

3.4.4 SUPPORT VECTOR MACHINE

One of the most popular Supervised Learning algorithms is Support Vector machine which is used to solve not only Classification but also Regression problems. But mostly this algorithm is used for solving Classification problems in Machine Learning. The aim of the SVM algorithm is to create the decision boundary. The main idea behind that is to divide n-dimensional space into classes. Now in the future one can correctly put the new data record into the correct category. Hyperplane is the name given to this best decision boundary.

3.4.5 ARTIFICIAL NEURAL NETWORK

The term "Artificial neural network" basically refers to a biologically inspired sub class of artificial intelligence modeled after the brain. An Artificial neural network is based on neurons and is usually a computational network that constructs the structure of the human brain. Artificial neural networks have neurons that are linked to each other in various layers of the networks similar to a human brain has neurons interconnected to each other. These neurons in ANN are known as nodes. ANN is of two types-feed forward and feedback networks.

3.4.6 RANDOM FOREST CLASSIFIER

Random Forest is a classification algorithm. It is a supervised learning technique. In Machine learning Both Classification and Regression problems can be solved using it. As the name itself suggests, "Random Forest classifier contains a number of decision trees on various sub-datasets of the given dataset and takes the prediction or average of all the decision trees to improve the accuracy of that dataset."

4. TESTING AND RESULTS

Accuracy: In any classification model, one of the metrics is accuracy for the performance analysis. By dividing the number of correct predictions with the total number of predictions is used to calculate accuracy. The Fraction of prediction is accuracy. The ratio of number of predictions made correctly to the overall predictions made by the model is determined. The exact formula of accuracy is calculated as:

$$\text{Accuracy} = \text{No: of correct predictions} / \text{Total no: of predictions}$$

Precision: When the algorithm predicts yes, how often it is actually correct

Recall: When it is actually a yes, how often does the algorithm predicts yes.

F1- score: It is weighted average of the true positive rate and the precision.

Support: Support is the number of actual occurrences of the true response class.

Decision Tree: It has been observed that the number of correct predictions of Decision Tree model are one sixty-five (165) out of two thirty one cases (231).

Accuracy = $165/231$

Accuracy = 0.714

Naive bayes: It has been observed that the number of correct predictions of Decision Tree model is One hundred seventy (170) out of two thirty one cases (231)

Accuracy = $170/231$

Accuracy = 0.7359.

Logistic Regression: It has been observed that the number of correct predictions of Decision Tree model are One sixty Three (163) out of two thirty one cases (231).

Accuracy = $163/231$

Accuracy = 0.7056.

Support Vector Machine: It has been observed that the numbers of correct predictions of Decision Tree model are One hundred seventy (170) out of two thirty one cases (231).

Accuracy = $170/231$

Accuracy = 0.7359.

Artificial Neural Network: It has been observed that the number of correct predictions of Decision Tree model are One sixty Two (162) out of two thirty one cases (231).

Accuracy = $162/231$

Accuracy = 0.7013

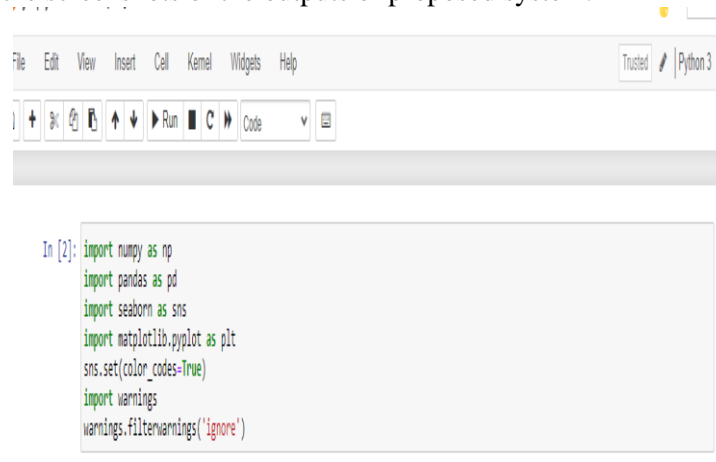
Random Forest Classifier: It has been observed that the number of correct predictions of Decision Tree model are One twenty six(126) out of three hundred eight cases (154).

Accuracy = $126/154$

Accuracy=0.8181

Output Screens:

The following are the screenshots of the outputs of proposed system.

A screenshot of a Jupyter Notebook interface. The top menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. On the right, it shows 'Trusted' and 'Python 3'. Below the menu is a toolbar with icons for undo, redo, run, and other actions. The main area contains a code cell with the following Python code:

```
In [2]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(color_codes=True)
import warnings
warnings.filterwarnings('ignore')
```

Figure: 3. Screenshot Showing the Libraries Imported in Proposed System

In the Figure: 4. the output for Decision tree model in python using Jupyter platform is shown.

```

decision tree
Accuracy of our DT model on trained data is : 1.0000
Accuracy of DT model on test data: 0.7143
Confusion Matrix for DT
[[ 52 28]
 [ 38 113]]

Classification Report

              precision    recall  f1-score   support

     1         0.58        0.65        0.61         80
     0         0.80        0.75        0.77        151

 accuracy         0.69
 macro avg         0.70
 weighted avg         0.71
    
```

Figure: 4. we can observe the accuracy of Decision tree model as 71.43

In the Figure: 5. the output for Naive bayes model in python using Jupyter platform is shown.

```

naive bayes
Accuracy of our naive bayes model on training data is : 0.7542
Accuracy of our naive bayes model on testing data is: 0.7359
Confusion Matrix
[[ 52 28]
 [ 33 118]]
Classification Report
              precision    recall  f1-score   support

     1         0.61        0.65        0.63         80
     0         0.81        0.78        0.79        151

 accuracy         0.71
 macro avg         0.72
 weighted avg         0.74
    
```

Figure: 5. accuracy of naive bayes model as 73.59% is observed

In the Figure: 6. the output for Logistic Regression model in python using Jupyter platform is shown.

```

Accuracy of Logistic Regression: 0.7056
[[ 54 26]
 [ 42 109]]

Classification Report
              precision    recall  f1-score   support

     1         0.56        0.68        0.61         80
     0         0.81        0.72        0.76        151

 accuracy         0.68
 macro avg         0.70
 weighted avg         0.71
    
```

Figure: 6. The accuracy of Logistic Regression model as 70.56% is observe

In the Figure: 7. the output for Support Vector Machine model in python using Jupyter platform is shown.

```

Accuracy of our SVM model on trained data is : 0.7840
Accuracy of SVM model on tested data: 0.7359
Confusion Matrix for Support Vector machine
[[ 46 34]
 [ 27 124]]

Classification Report
              precision    recall  f1-score   support

     1         0.63        0.57        0.60         80
     0         0.78        0.82        0.80        151

 accuracy         0.71
 macro avg         0.70
 weighted avg         0.73
    
```

Figure: 7. The accuracy of Support Vector Machine model as 73.59% is observed

In the Figure: 8. the output for ANN model in python using Jupyter notebook is shown


```

Accuracy of our ANN model on trained model is : 0.8045
Accuracy of our ANN model on test data: 0.7013
Confusion Matrix for Artificial Neural Network
[[ 40  40]
 [ 29 122]]

Classification Report

              precision    recall  f1-score   support

     1         0.58      0.50      0.54         80
     0         0.75      0.81      0.78        151

 accuracy         0.67
 macro avg         0.67      0.65      0.66
 weighted avg         0.69      0.70      0.70
    
```

Figure 8. The accuracy of A NN as 70.13% model is observed

In the Figure:9. the output for Random Forest Classifier model in python using Jupyter platform is shown.

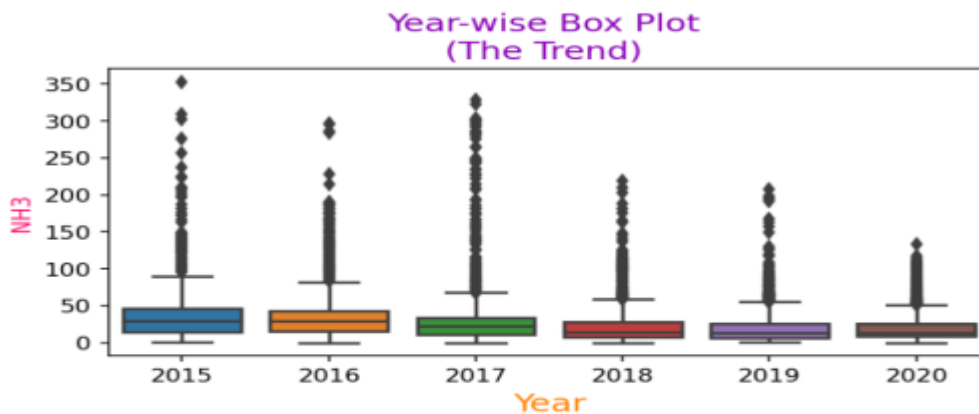
```

----- confusion_matrix -----
[[44  7]
 [ 7 19]]
----- accuracy_score -----
0.8181818181818182
----- classification_report -----
              precision    recall  f1-score   support

     0         0.86      0.86      0.86         51
     1         0.73      0.73      0.73         26

 accuracy         0.82
 macro avg         0.80      0.80      0.80
 weighted avg         0.82      0.82      0.82
    
```

Figure 9. The accuracy of Random Forest Classifier model as 81.81% is observed



The accuracy of each algorithm as shown in above output screens are:

- Decision tree - 71.43%
- Naive bayes - 73.59%
- Logistic Regression - 70.56%
- Support Vector Machine - 73.59%
- Artificial Neural Network - 70.13%
- Random Forest Classifier - 81.81%

When the accuracies of all the Machine Learning Classification algorithms is compared, Random Forest Classifier is considered as best suitable one for the prediction of diabetes with the accuracy of 81.81%. So Random Forest Classifier should be used for prediction of diabetes.

5. FUTURE SCOPE

Missing the attribute values and the dataset size are the limitations of this Random Forest Algorithm. Not only these limitations but thousands of records are needed which includes zero missing values. So, for the better accuracy the entire focus is to tune the parameters by integrating other methods into the used model. Future work should be focused to build a prediction model for diabetes which leads to 99% accuracy. Apart from resulting with good accuracy, large dataset is to be considered with no missing attribute values which will reveal more outputs with good insights and the best prediction accuracy.

6. CONCLUSION

In this paper for predictive analytics various familiar machine learning algorithms are used. These algorithms include SVM, KNN, LR, DT, RF and NB. On the dataset named PIMA Indian consisting 768 records about the diabetes predictions were made. For training and testing the predictive model, 8 attributes were selected. The accuracies obtained from the above algorithms were considered. Out of these the algorithm that gives highest accuracy for predicting diabetes is Random Forest classifier. The accuracy provided by this algorithm is highest as compared to other five algorithms used which is 81.81%. So, it is considered as the appropriate algorithm in order to predict diabetes disease so early.

REFERENCES

- [1] Aishwarya, R., Gayathri, P. Jaisankar, N., 2013. "A Method for Classification Using Machine Learning Technique for Diabetes". *International Journal of Engineering and Technology (IJET)* 5, 2903–2908.
- [2] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. "Application of data mining: Diabetes health care in young and old patients". *Journal of King Saud University - Computer and Information Sciences* 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
- [3] Bamnote, M.P., G.R., 2014. "Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming". *Advances in Intelligent Systems and Computing* 1, 763–770. doi:10.1007/978-3-319-11933-5.
- [4] Dhomse Kanchan B., M.K.M., 2016. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", in: *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE*. pp. 5–10
- [5] Fatima, M., Pasha, M., 2017. "Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*" 09, 1–16. doi:10.4236/jilsa.2017.91001.
- [6] Krati Saxena, Dr. Zubair Khan, Shefali Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm", *International Journal of Computer Science Trends and Technology (IJCSST) – Volume 2 Issue 4, July-Aug 2014*
- [7] K Sowjanya , Ayush Singhal , Chaitali Choudhary , "A machine learning based system for predicting diabetes risk using mobile devices", *IEEE International Advance Computing Conference (IACC)* ,2015
- [8] Muhammad Azeem Sarwar, Nasir Kamal, Wjeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", *24 th International Conference on Automation and Computing (ICAC)*, 2018
- [9] Roxana Mirshahvalad, Nastaran Asadi Zanjani , "Diabetes prediction using ensemble perceptron algorithm" , *9th International Conference on Computational Intelligence and Communication Networks (CICN)* ,2017.
- [10] Vclijayan, V.V., Anjali, C., 2015. "Prediction and diagnosis of diabetes mellitus A machine learning approach". *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 122–127doi:10.1109/RAICS.2015.7488400.
- [11] Anoop Joyti Sahoo, and Rajesh Kumar Tiwari "A Novel Approach for Hiding Secret data in Program Files" *International Journal of Information and Computer Security*. Volume 8 Issue 1, March 2016,
- [12] Abu Salim, Sachin Tripathi and Rajesh Kumar Tiwari "A secure and timestamp-based communication scheme for cloud environment" Published in *International Journal of Electronic Security and Digital Forensics*, Volume 6, Issue 4, 319-332.
- [13] Rajesh Kumar Tiwari and G. Sahoo, "A Novel Watermark Scheme for Secure Relational Databases" *Information Security Journal: A Global Perspective*, Volume 22, Issue 3, July 2013