A Machine Learning Approach to Diabetes Prediction with Feature Selection

Rashmik Manchiraju
Birla Institute of Technology,
Mesra Ranchi

Vandana Bhattacharjee
Birla Institute of Technology,
Mesra Ranchi

***Abstract*:- Diabetes is a constant sickness that happens when the pancreas neglects to create sufficient insulin or when the body's insulin is incapably utilized. Insulin is a chemical that assists with holding glucose levels under tight restraints. Uncontrolled diabetes causes hyperglycemia, or high glucose, which makes disastrous harm to a significant number of the body's frameworks, including the neurons and veins, after some time. Machine learning techniques have been applied for many health care problems with good results. The goal of this paper is to analyse different classification algorithm such as Support Vector Machines, Decision tree, K-Nearest Neighbour, Naïve Bayes and Random Forest to identify diabetes at beginning phase.**

1. INTRODUCTION

Diabetes is one of the world's most frequent diseases. If a person leads a stressful life or is obese, and carries additional weight in the belly area of the body, insulin activity is hampered, resulting in diabetes. In [1] it was stated that as indicated by (WHO) World Health Organization around 422 million individuals experiencing diabetes particu-larly from low or inactive pay nations. And this could be increased to 490 billion up to the year of 2030. The causes of diabetes as mentioned in [2] are Genetic factors. It is brought about by somewhere around two freak qualities in the chromosome 6, the chromosome that influences the reaction of the body to different antigens. Viral disease may likewise impact the event of type 1 and type 2 diabetes. Diabetes as stated in [3] is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay away from the complications. Application of machine learning algorithm were applied in different medical data sets including machine Diabetes dataset [4-5]. Machine Learning (ML) has been a magnificent help for making expectation of a specific framework via preparing. ML is tied in with gaining structures from the information which is given. ML lately has been the advancing, solid and supporting apparatus in clinical space. Programmed learning has gotten a more prominent measure of interest in clinical space because of less measure of time for identification and less communication with patient, saving time for patients care. Medical care areas have huge volume data sets [7]. Such data sets might contain organized, semi-organized or unstructured information. Big Data Analytics is the cycle which investigations enormous informational indexes and uncovers stowed away data. These days, there is a developing requirement for Internet of Things (IoT)- based versatile medical services applications that assistance to anticipate sicknesses [8].

2. METHODS

2.1 K-Nearest Neighbour Algorithm

The k-nearest neighbors (KNN) calculation is a machine learning algorithm used to take care of classification and regression issues. KNN utilizes working out distance between two focuses on a diagram. Choosing the right K is our errand and for that we run our calculation on different occasions for decreasing the quantity of blunders with the end goal that a more exact outcome is gotten.

The K-NN working can be made sense of based on the underneath algorithm:

Step-1: Input the number K of the neighbours

Step-2: Calculation of the Euclidean distance of the test data from all the data points is done.

Step-3: Take the K closest neighbors according to the determined Euclidean distance.

Step-4: Among these K neighbors, the quantity of the data of interest in every classification is counted.

Step-5: The new information point is relegated to that classification for which the quantity of the neighbors is greatest.

2.2 SVM Algorithm

SVM represents Support Vector Machine and is one of the most generally involved Supervised Learning calculations for Classification and Regression issues. Nonetheless, it is generally used in Machine Learning for Classification troubles. The SVM calculation's motivation is to track down the ideal line or choice limit for classifying n-layered space into classes so extra

information focuses can be promptly positioned in the right classification later on. A hyperplane is the name for the ideal decision limit.
The outrageous focuses/vectors that assist in making the hyperplane are picked by SVM.These outrageous vectors are realized by the term called support vectors and consequently the name of the calculation is called Support Vector Machine.

2.3 Decision Tree Algorithm

Decision Tree is a managed learning procedure that can be applied to grouping and relapse issues, but tackling characterization problems is generally ordinarily utilized. Inside hubs address dataset credits, branches address choice principles, and each leaf hub gives the end in this tree-organized classifier.

The Decision Node and the Leaf Node are the two hubs of a Decision tree. Leaf hubs are the result of those choices and contain no more branches, though Decision hubs are utilized to go with any choice and have a few branches.

The choices or tests depend on the qualities of the given dataset.It's a graphical portrayal for getting all doable answers for an issue/choice relying upon specific boundaries.

It's named a choice tree since, similar to a tree, it begins with a root hub and develops into a tree-like design with extra branches.We use the CART calculation, which represents Classification and Regression Tree calculation, to shape a tree.A choice tree essentially poses an inquiry and partitions the tree into subtrees in view of the response (Yes/No).

2.4 Naïve Bayes Neighbour Algorithm

The Naive Bayes strategy is a regulated learning method for tending to order gives that depends on the Bayes hypothesis. It is most regularly utilized in text order with a huge preparation dataset.The Naive Bayes Classifier is a basic and compelling grouping technique that guides in the improvement of quick ML models equipped for making fast expectations. It's a probabilistic classifier, and that implies it makes forecasts in view of an item's likelihood.

Baye's Theorem- It is used to determine the probability of hypothesis with prior knowledge.It depends on Conditional Probability.

$$\mathrm{P(A|B)}=\frac{P(B|A)P(A)}{P(B)}$$

2.5 Random Forest Algorithm

Random Forest is a prominent machine learning algo that utilizes supervised learning procedures. In ML, it very well may be used for both classification and regression issues. It depends on group realizing, which is a technique for incorporating various classifiers to tackle a complicated issue and increment the model's exhibition.

The random forest is shaped in two stages: the first is to consolidate N decision trees to assemble the random forest, and the second is to make expectations for each tree made in the primary stage. The most common way of working can be made sense of exhaustively through the accompanying advances

Step-1: Select irregular K pieces of information from the preparation set.

Step-2: Build the decision trees related with the chose data of interest (Subsets).

Step-3: Choose the number N for decision trees that you need to construct.

Step-4: Repeat Step 1 and 2

Step-5: For new points , find the expectations of every decision tree, and match the new data points with the category that has the highest votes.

2.6 Evaluation Parameters

Table 1. Confusion matrix

Predicted ***Actual***	True	False
True	True positive	False negative
False	False positive	True negative

The evaluation parameters used in this research work are precision, recall, f-measure and accuracy.

Precision estimates the quantity of positive class forecasts that have a place with the positive class.

Precision (P) = **TruePositives / (TruePositives + FalsePositives)**

Recall estimates the quantity of positive class expectations made from all certain models in the dataset.

Recall (R) = **TruePositives / (TruePositives + FalseNegatives)**

F-Measure offers a solitary score that adjusts both the worries of precision and recall in one number.

F-Measure (FM) = **(2 * Precision * Recall) / (Precision + Recall)**

Though, accuracy is the complete number of right expectations partitioned by the all out number of forecasts made for a dataset.

Accuracy (A) = **(TP+TN)/(TP+FP+FN+TN)**

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

3. EXPERIMENTS & RESULTS

Pima Indian Diabetes Database is a recognizable and ordinarily utilized informational collection for the expectation of diabetes. This informational index comprises of 768 rows and 9 columns. The characteristics remembered for the section are glucose, pregnancies, skin thickness, Blood Pressure, BMI, insulin, age, and results. The result variable predicts whether the patient is diabetic positive or diabetic-negative. Pandas capability is used to peruse CSV file where the informational index document is in succeed design [6]. Tables 1 – 5 present the results of experiments with different classifiers.

KNN

Table 1. Performance of KNN classifier with feature sets

Parameters	Precision	Recall	F-Measure	Accuracy
Glucose, BMI	0.6279	0.5744	0.6	0.7622
Glucose, BMI, Pregnancies, Age	0.6888	0.6595	0.6739	0.8051
Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin	0.6956	0.6808	0.6881	0.8116

SVM

Table 2. Performance of SVM classifier with feature sets

Parameters	Precision	Recall	F-Measure	Accuracy
Glucose, BMI	0.7222	0.5531	0.6265	0.7987
Glucose, BMI, Pregnancies, Age	0.6923	0.5744	0.6279	0.7922
Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin	0.7222	0.5531	0.6265	0.7987

DECISION TREE

Table 3. Performance of Decision Tree classifier with feature sets

Parameters	Precision	Recall	F-Measure	Accuracy
Glucose, BMI	0.4897	0.5106	0.5	0.6883
Glucose, BMI, Pregnancies, Age	0.5454	0.6382	0.5882	0.7272
Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin	0.5918	0.6170	0.6041	0.7532

NAIVE BAYES

Table 4. Performance of NAIVE BAYES classifier with feature sets

Parameters	Precision	Recall	F-Measure	Accuracy
Glucose, BMI	0.6829	0.5957	0.6363	0.7922
Glucose, BMI, Pregnancies, Age	0.625	0.6382	0.6315	0.7727
Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin	0.6458	0.6595	0.6526	0.7857

RANDOM FOREST

Table 5. Performance of Random Forest classifier with feature sets

Parameters	Precision	Recall	F-Measure	Accuracy
Glucose, BMI	0.5510	0.5744	0.5625	0.7272
Glucose, BMI, Pregnancies, Age	0.6530	0.6808	0.6666	0.7922
Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin	0.6862	0.7446	0.7142	0.8181

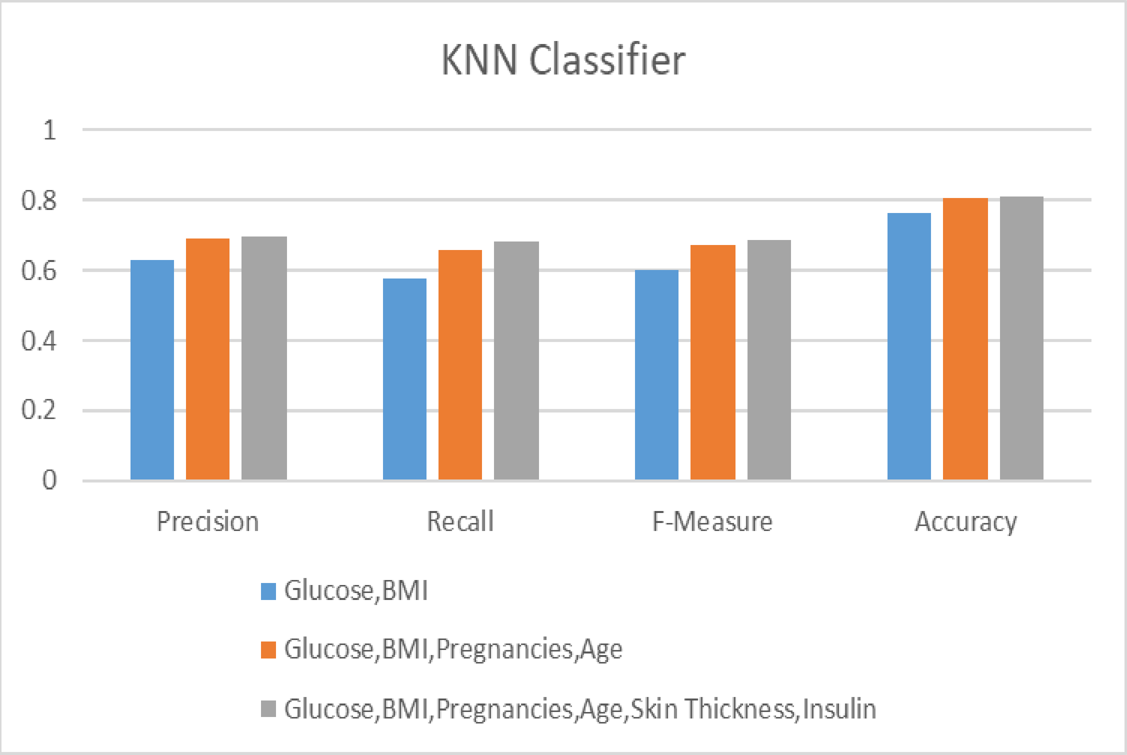

Figure 1. Bar Graph visualization for KNN classifier performance

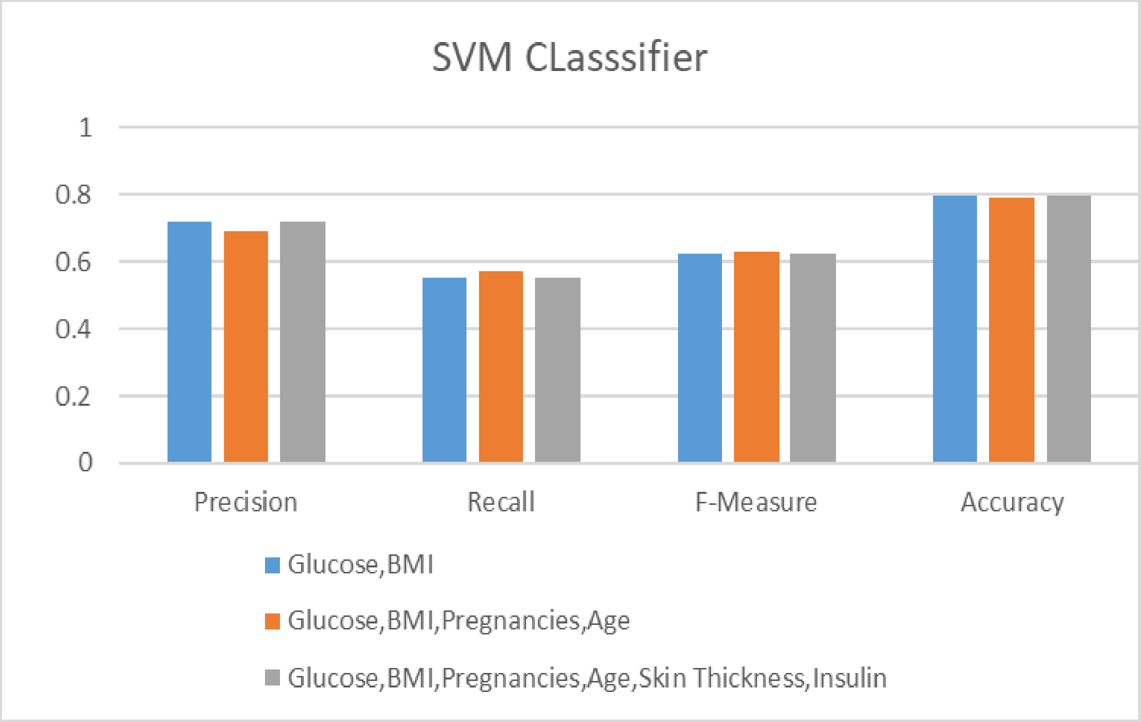

Figure 2. Bar Graph visualization for SVM classifier performance

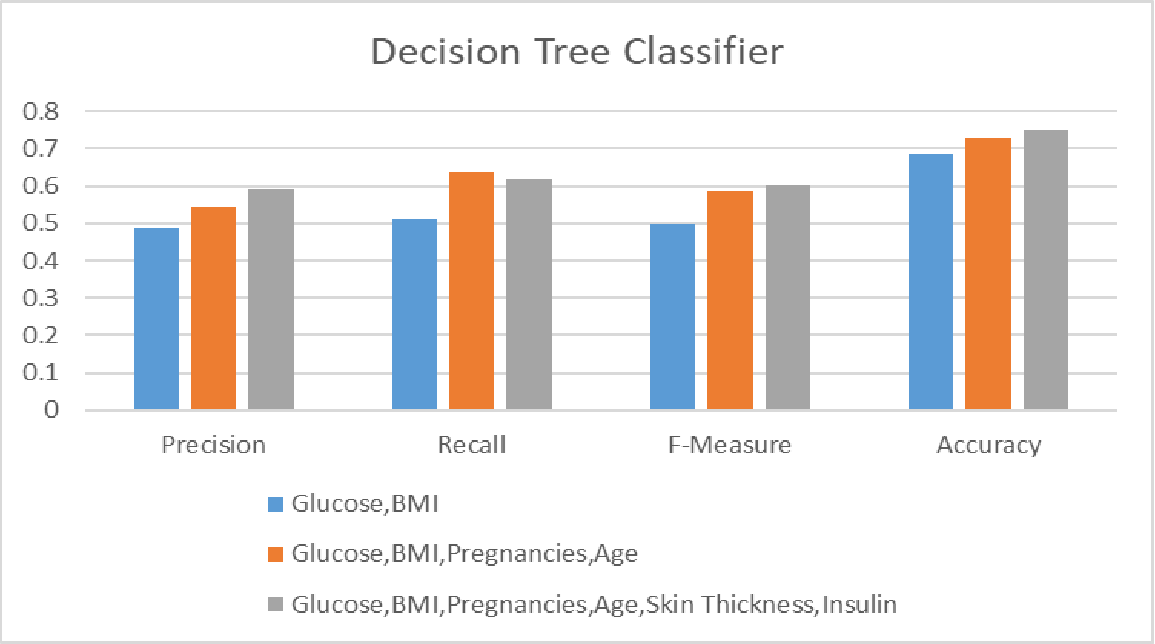

Figure 3. Bar Graph visualization for Decision Tree classifier performance

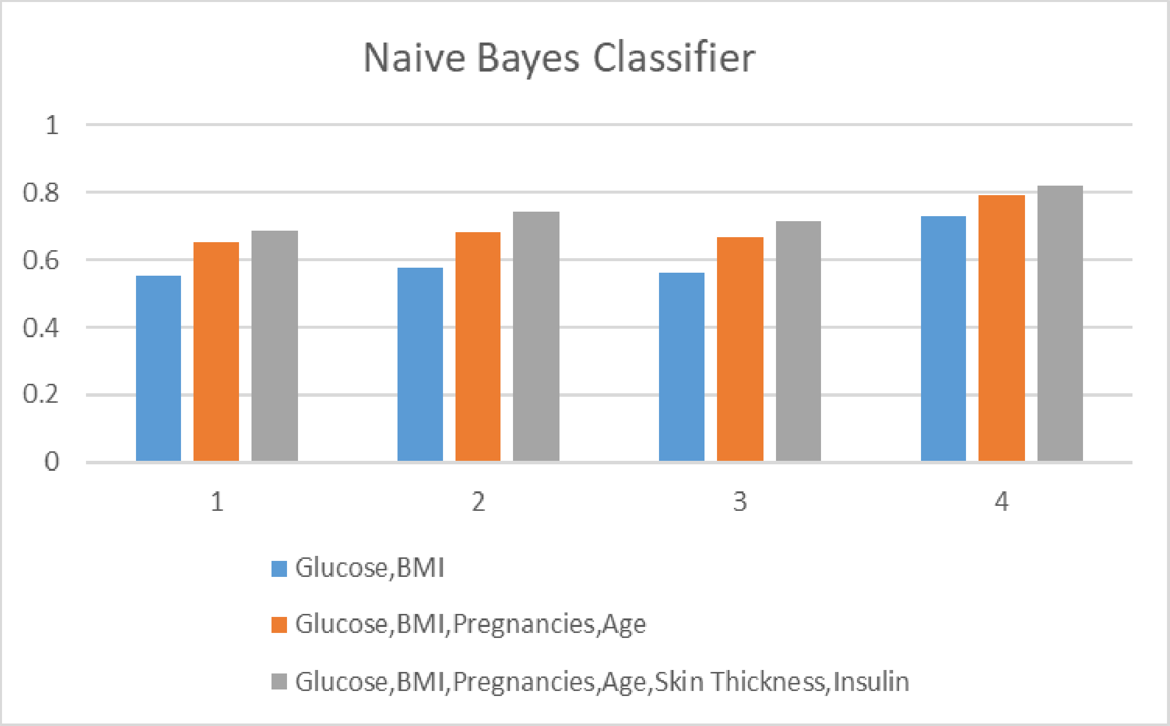

Figure 4. Bar Graph visualization for Naïve Bayes classifier performance

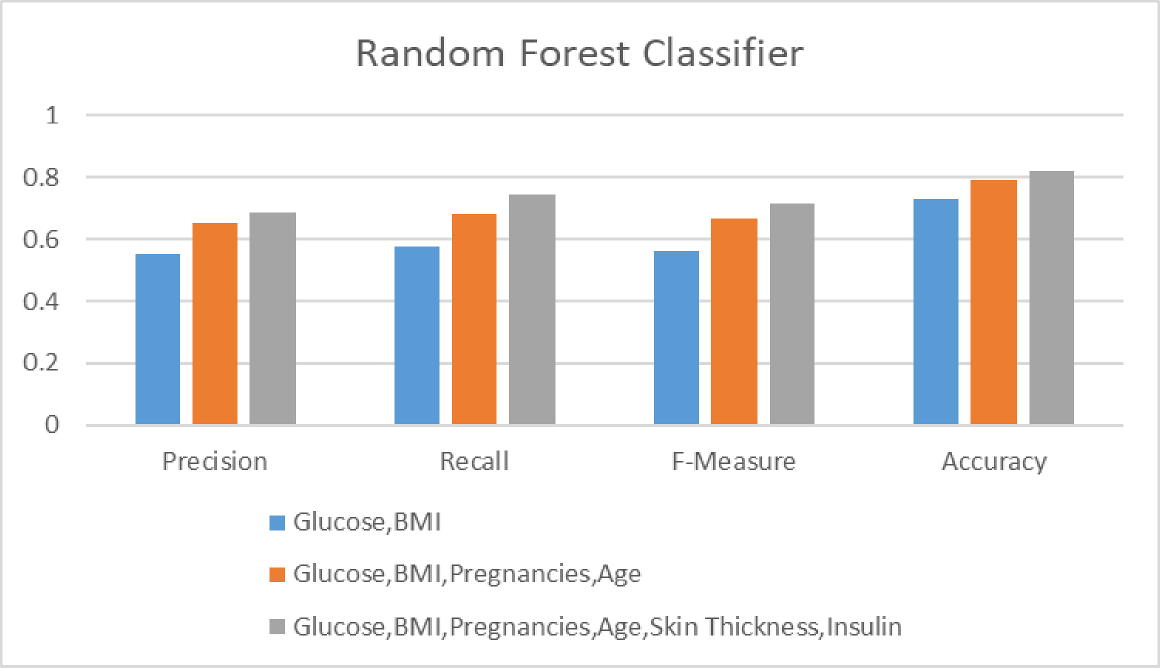

Figure 5. Bar Graph visualization for Random Forest classifier performance

4. CONCLUSION

Various machine learning classifiers have been applied on the diabetes dataset. Four evalution parameters were taken into consideration- Precision, Recall, F-measure, Accuracy. From Figures 1 – 5 it can be seen that, Precision was highest when SVM classifier was used with 72.22 % and lowest when Decision Tree was used with 48.97% . Recall and F-measure were highest when Random Forest was used with values of 74.46% and 71.42 %, lowest when Decision Tree was used with values of 51.16 % and 50 % respectively.

The accuracy value was highest for Random forest with 81.81 % slightly higher than the KNN value of 81.16 % and it was lowest for Decision Tree with a value of 68.83 %. For the SVM and Naïve Bayes classifier the feature set of {Glucose, BMI} gave the highest values for all parameters, however for all other classifiers the feature set {Glucose, BMI, Pregnancies, Age, Skin Thickness, Insulin}, gave the highest performance. It is important that we select the right classifier with the right feature set in order to get accurate solutions to real life problems.

REFERENCES

[1] Mitushi Soni, Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 09 (September 2020).

[2] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 294-305. 10.32628/CSEIT206463.

[3] Deepti Sisodia, Dilip Singh Sisodia,Prediction of Diabetes using Classification Algorithms,Procedia Computer Science,Volume 132,2018,Pages 1578-1585,

[4] Saru, S. and Subashree, S., Analysis and Prediction of Diabetes Using Machine Learning (April 2, 2019). International Journal of Emerging Technology and Innovative Engineering, Volume 5, Issue 4, April 2019

[5] Aishwarya R., Gayathri P., Jaisankar N, "A Method for Classification Using Machine Learning Technique for Diabetes '',International Journal of Engineering and Technology (IJET), 5 (2013), pp. 2903-2908

[6] Raja Krishnamoorthi, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, Basant Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Journal of Healthcare Engineering, vol. 2022, Article ID 1684017, 10 pages, 2022. https://doi.org/10.1155/2022/1684017

[7] Mujumdar, Aishwarya & Vaidehi, V.. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science. 165. 292-299. 10.1016/j.procs.2020.01.047.

[8] Sasmita Padhy, Sachikanta Dash, Sidheswar Routray, Sultan Ahmad, Jabeen Nazeer, Afroj Alam, "IoT-Based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction", Computational Intelligence and Neuroscience, vol. 2022, Article ID 2389636, 11 pages, 2022. https://doi.org/10.1155/2022/2389636