JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION IN HEALTHCARE SYSTEMS USING MACHINE LEARNING ALGORITHMS

K. Harini[1], K.S. Hemalatha[2], D. Kavi Priya[3], A.R. Nalini[4], C. Visali[5]

[1]Student, [2]Student, [3]student, [4] Students, [5] Assistant Professor

[1]Student, Department of information technology

Vivekanandha College of Engineering for Women (Autonomous), Tiruchengode, Namakkal-637205, Tamil Nadu, India.

ABSTRACT:

The remarkable advancements in public healthcare infrastructures have led to a momentous production of critical and sensitive healthcare data. By applying intelligent data analysis techniques, many interesting patterns are identified for the first onset detection and prevention of several fatal diseases.

Diabetes mellitus is a particularly life-threatening disease because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage. During this paper, a machine learning-based approach has been proposed for the classification, early-stage identification, and prediction of diabetes.

KEYWORDS: Diabetes mellitus, Algorithms used are Logistic Regression (LR), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Random Forest (RF), XG Boost, Gradient Boost.

1. INTRODUCTION

Public health could be a fundamental concern for shielding and preventing the community from hazardous diseases Governments are spending a substantial amount of their gross domestic product (GDP) for the welfare of the general public, and initiatives like vaccination have prolonged the life of individuals. However, for the last few years, there has been a substantial emergence of chronic and genetic diseases affecting public health diabetes is one all the extremely life-threatening diseases because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage.

Diabetes may be a disorder that impairs an individual's body to processing glucose, referred to as blood glucose. This disease is characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. An absolute deficiency of insulin secretion causes type 1 diabetes (T1D). Diabetes drastically spreads because of the patient's inability to use the produced insulin. It is called type 2 diabetes (T2D) Both types are increasing rapidly, but the ratio of increase in T2D is on top of T1D. 90 to 95% of cases of diabetes are of T2D.

Inadequate supervision of diabetes causes stroke, hypertension, and cardiovascular diseases. To avoid and reduce the complications due to diabetes, a monitoring method of BG level plays a prominent role. A mixture of biosensors and advanced information and communication technology (ICT) provides an efficient real-time monitoring management system for the health condition of diabetic patients by using an SMBG (self-monitoring of blood glucose) portable device. A patient can check the changes in glucose level in his blood by himself. Users can better understand BG changes by using CGM (continuous glucose monitoring) sensors.

Here six machine learning algorithms have been applied to predict diabetes. They are Logistic Regression (LR), Random Forest, Multilayer perceptron (MLP), Gradient Boosting, Long Short-Term Memory (LSTM), and Boost.

2. LITERATURE REVIEW:

The following chapters give an overview of the various methodologies used by various authors for disease prediction using machine learning methodologies. We can observe that there is a fine comparison made between 4 major machine learning algorithms whether they can predict the presence of the disease with greater accuracy, achieving optimal performance.

2.1 Diabetes Prediction Using Different Machine Learning Approaches

Author: Priyanka Sonar, Prof. K. JayaMalini

Published In: Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)

Diabetes is one of the most lethal diseases in the world. It is additionally an inventor of varied types of disorders for example blindness, viscus diseases, etc. In such a case the patient is required to go to a diagnostic center, to get their reports after consultation. Because each time they need to invest their time and currency. But with the expansion of Machine Learning methods, we've got the pliability to go looking for a solution to this issue, we've got advanced system mistreatment informatics that has the flexibility to forecast whether the patient has a polygenic illness or not. Furthermore, forecasting the sickness initially finishes up in providing the patients before it begins vital. Information withdrawal has the flexibility to get rid of unseen data from a large quantity of diabetes-associated information. This analysis aims to develop a system that could predict the diabetic risk level of a patient with much better accuracy. Model development is predicated on categorization methods such as Decision Tree, ANN, I Bayes, and LSTM algorithms. For Decision Tree, the models give precisions of 85%, for I Bayes 77%, and 77.3% for Long short-term memory. Outcomes show a big accuracy of the methods.

2.2 Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm,

Authors: Samrat Kumar Dey, Ashraf Hossain, and Md. Mahbubur Rahman

Published In: 2018 21st International Conference of Computer and data Information Technology (ICCIT)

Diabetes is caused due to the excessive amount of sugar condensed into the blood. Currently, it's considered together of the most lethal diseases in the world. People all around the globe are full of this severe disease knowingly or unknowingly. Other diseases like a coronary failure, paralyzed, renal disorder, blindness, etc., are also caused by diabetes. Numerous computer-based detection systems were designed and outlined for anticipating and analyzing diabetes. The usual identifying process for diabetic patients needs longer and more money. But with the increase of machine learning, we can develop an answer to the present intense issue. Therefore, we've developed an architecture that has the potential to predict whether the patient has diabetes or not. Our main aim of this exploration is to create an internet application that supported the upper prediction accuracy of some powerful machine learning algorithm. We've got used a benchmark dataset namely Pima Indian which may predict the onset of diabetes-supported diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a major improvement in accuracy which drives us to develop an Interactive Web Application for Diabetes Prediction.

2.3 Diabetes Disease Prediction Using Data Mining

Authors: Dheeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil

Published In: 2017 International Conference on Innovations in Information, Embedded, and Communication Systems (ICIIECS)

Data mining could be a subfield within the subject of software engineering. It's the methodical procedure of finding examples in huge data sets including techniques at the crossing point of manufactured intelligence, machine learning, insights, and database systems. The goal of the info data mining methodology is to think data from a knowledge set and alter it into an inexpensive structure for further use. Our examination concentrates on this a part of medical conclusion learning design through the gathered data of diabetes and to make smart therapeutic choices emotionally supportive network to assist the physicians. The first target of this examination is to assemble an Intelligent Diabetes Disease Prediction System that provides analysis of diabetes malady utilizing diabetes patient's database. During this system, we propose the utilization of algorithms like Bayesian and KNN (K- Nearest Neighbour) to use in diabetes patient's databases and analyze them by taking various attributes of diabetes for the prediction of diabetes disease.

2.4 An Intelligent System for Diabetes Prediction

Authors: Zhilbert Tafa, Nerxhivane Pervetica and Bertran Karahoda

Published In: 4thMediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro

With the emerging increase of diabetes, which recently affects around 346 million people, of which more than one-third go undetected in the early stage, a strong need for supporting the medical decision-making process is generated. Many researchers have focused either on using one of the algorithms or on the comparisons of the performances of algorithms on a given, usually predefined and static datasets that are accessible through the Internet. This paper focuses on the joint implementation of the Long short-term memory (LSTM) and Naive Bayes statistical analysis, in the dataset acquired from the medical examinations of 402 patients, to improve the computer-supported diagnosis reliability. The dataset contains some attributes that have not been previously used in computer-based evaluations. The results show that the joint implementation of two algorithms significantly improves the overall reliability of the system outcome, which is crucial in the computer-supported diabetes diagnostic process.

3. Role of physical activity in prevention and control of diabetes mellitus:

Generally, physical activity is the first prevention and control strategy suggested by healthcare professionals for diabetic or prediabetic patients. Among diet and medicine, exercise could be a fundamental component in diabetes, disorder, obesity, and lifestyle rescue programs. Nonetheless, managing all the fatal diseases incorporates a significant economic burden. However, diabetes emerged as a devastating problem for the health sector and economy of a rustic this century.

Recently, the international diabetes prevention and control federation predicts that diabetes can affect over 366 million people worldwide. The disease control and prevention center within the US alarmed the govt that diabetes can affect over 29 million people. While these alarming numbers are continuously increasing, they'll burden the economy around the globe. Therefore, researchers and healthcare professionals worldwide are researching and proposing guidelines to forestall and control this life-threatening disease.

Sato presented an intensive survey on the importance of exercise prescription for diabetes patients in Japan. He suggested that prolonged sitting should be avoided and physical activity should be performed every half-hour. Kirwan et al. emphasized regular exercise to manage and forestall type 2 diabetes. Particularly, they studied the metabolic effect on tissues of diabetic patients and located very significant improvements in individuals performing regular exercise. Moser et al. have also highlighted the importance of standard exercise in improving the functionality of varied organs of the body, as shown in Figure 1.

Yang et al. focused on exercise therapy which plays a big role in treating diabetes and its associated side effects. Specifically, they found cytokines that supply a unique insight into diabetes control, but the sequence continues to be an understudy. Kim and Jeon presented a scientific overview of the effect of various exercises on the metabolism improvement of diabetic young individuals. They distinguished out that several studies reported the importance of exercise on insulin, BP, and BG level improvement. However, none of those studies mentions the beta-cell improvement. Therefore, many challenges act diabetes prevention and control, which require serious attention from researchers worldwide.

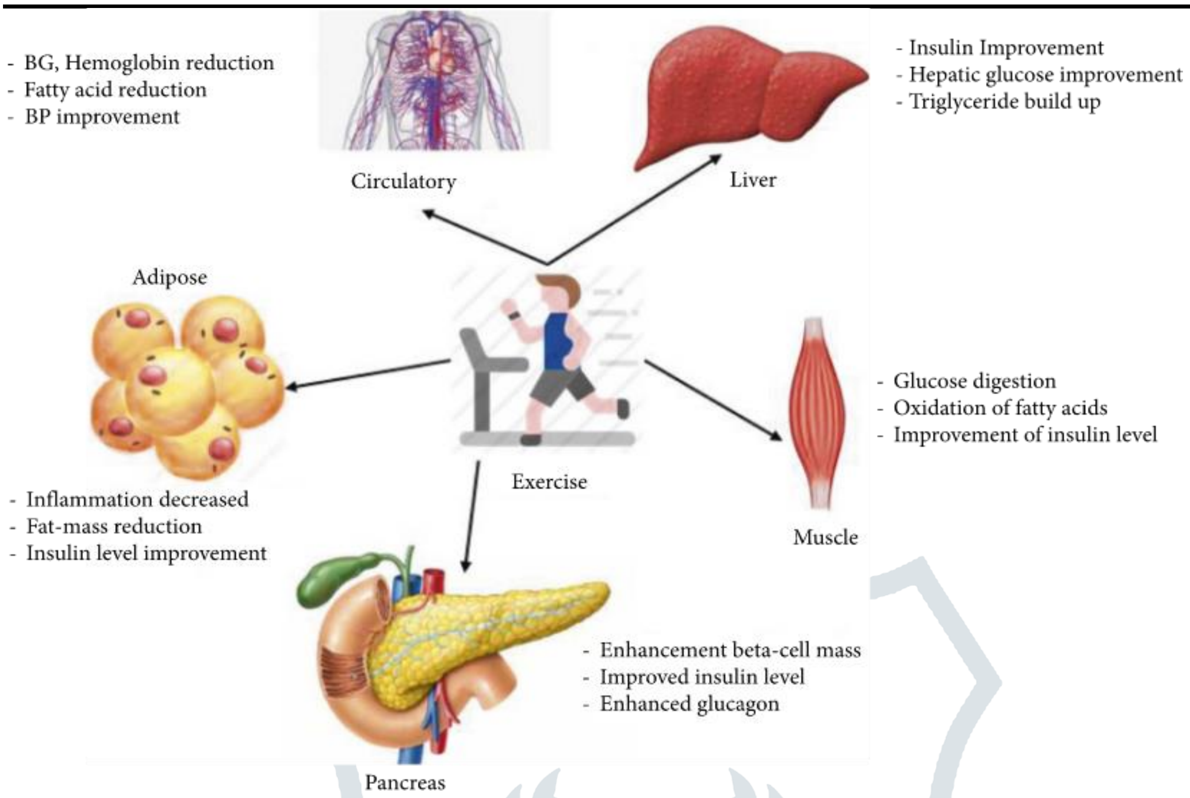


Figure 1: Result of well-ordered exercise on metabolism of diabetic patients

4. Proposed Diabetic Classification and Prediction System for Healthcare

The proposed diabetes classification and prediction system have exploited different machine learning algorithms. First, to classify diabetes, we utilized logistic regression, random forest, and MLP. Notably, we fine-tuned MLP for classification thanks to its promising performance in healthcare, specifically in diabetes prediction. The proposed MLP architecture and algorithm are shown in figure 2 and algorithm 1, respectively.

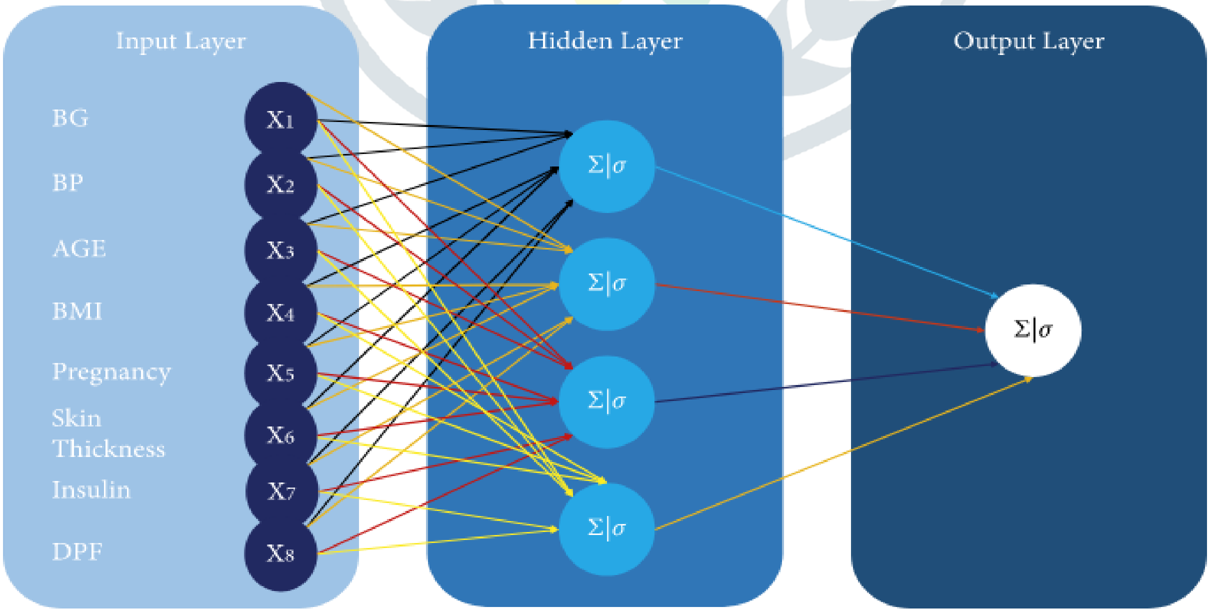


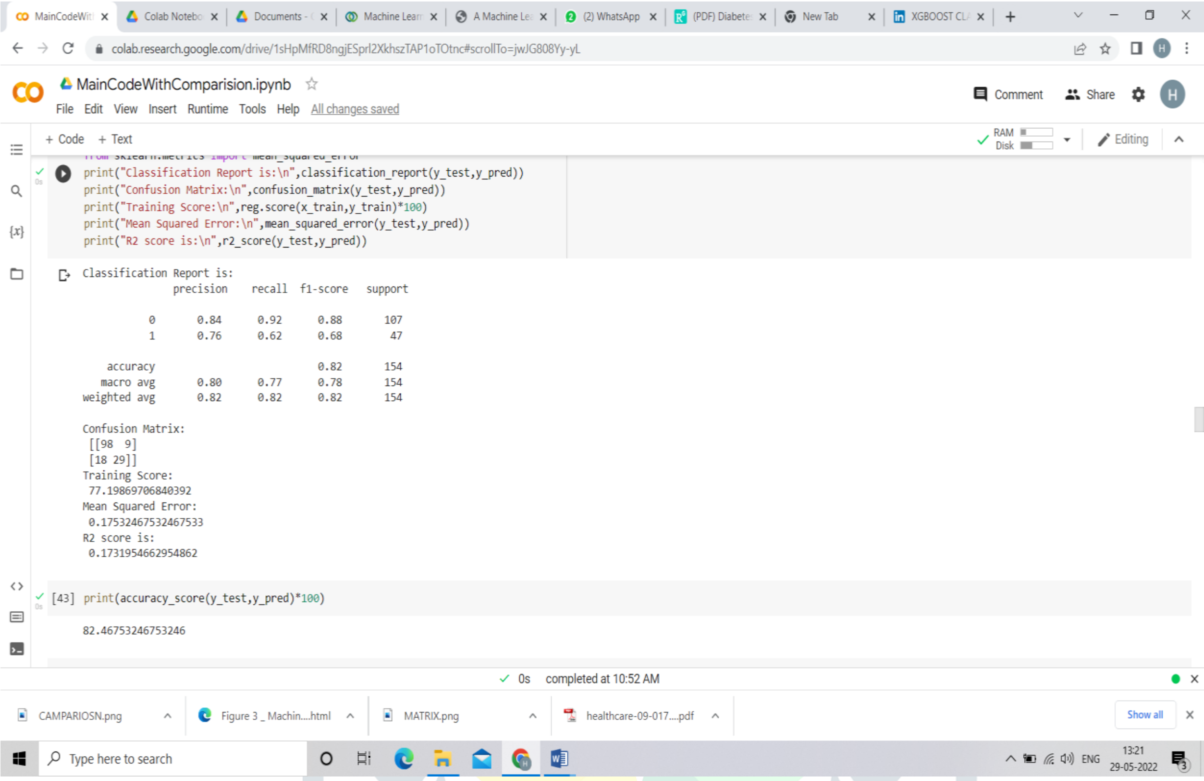
Figure 2: Suggested MLP architecture with eight variables as code for diabetes classification.

4.1. Diabetes Classification Techniques

For diabetic classification, we fine-tuned three widely used state-of-the-art techniques. Mainly, a comparative analysis is performed among the proposed techniques for classifying a person in either of the diabetes categories. The main points of the proposed diabetes techniques are as follows.

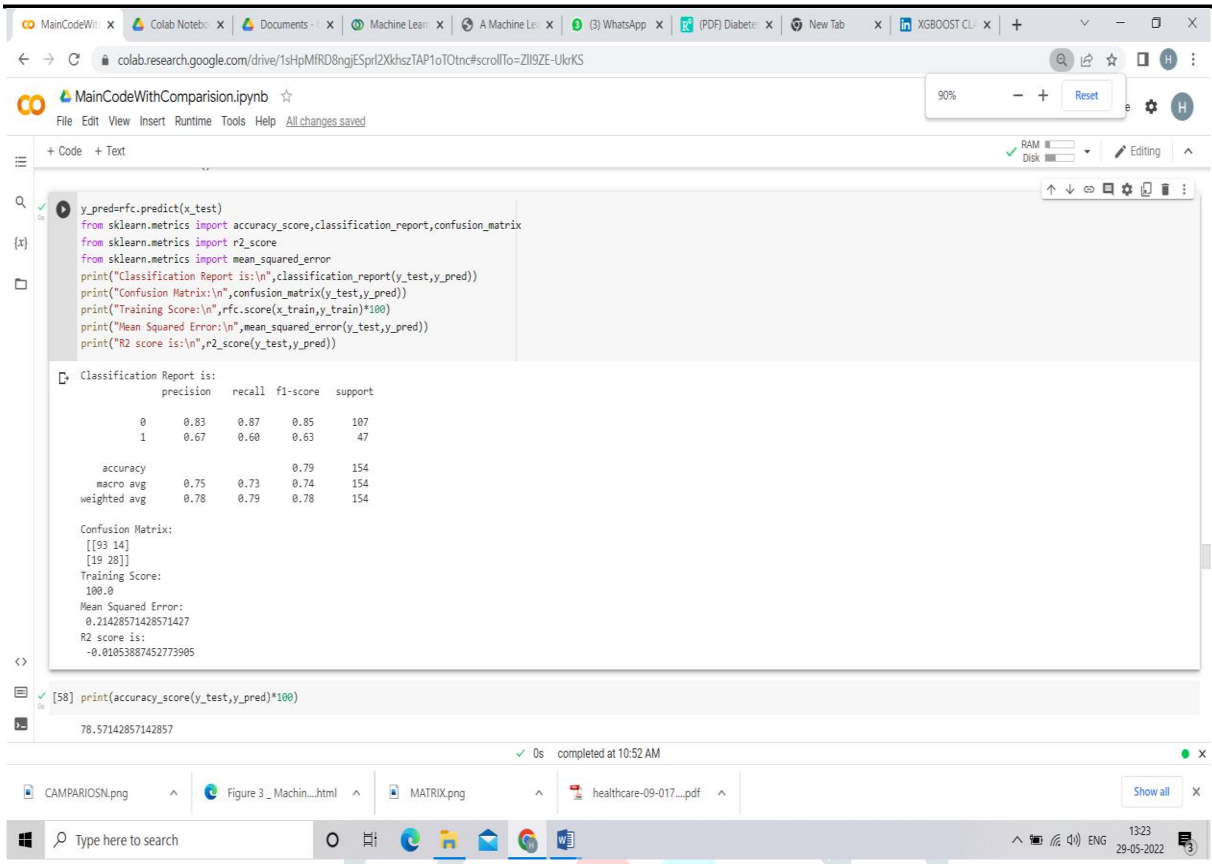
4.1.1. Logistic Regression

It is appropriate to use logistic regression when the variable is binary, as we've got to classify personnel in either type 1 or type 2 diabetes. Besides, it's used for predictive analysis and explains the connection between a variable and one or many independent variables, as shown in equation (1). Therefore, we used the sigmoid cost function as a hypothesis function (hθ(*x*)). The aim is to attenuate cost function J(θ). It always leads to classifying an example either in class 1 or class 2. **82.46%** accuracy is obtained by applying this algorithm.



4.1.2. Random Forest (RF)

As its name implies, it's a group of models that operate as an ensemble. The critical idea behind RF is that in the wisdom of the gang, each model predicts a result, and within the tip, the majority wins. It's been utilized within the literature for diabetic prediction and was found to be effective. Given a group of coaching examples $X = x_1, x_2, \ldots$ XM and their respective targets $Y = y_1, y_2, \ldots, y_m$, the RF classifier iterates B times by choosing samples with replacement by fitting a tree to the training examples. The training algorithm consists of the next steps depicted in equation (2). **78.57%** accuracy is obtained by applying this algorithm.

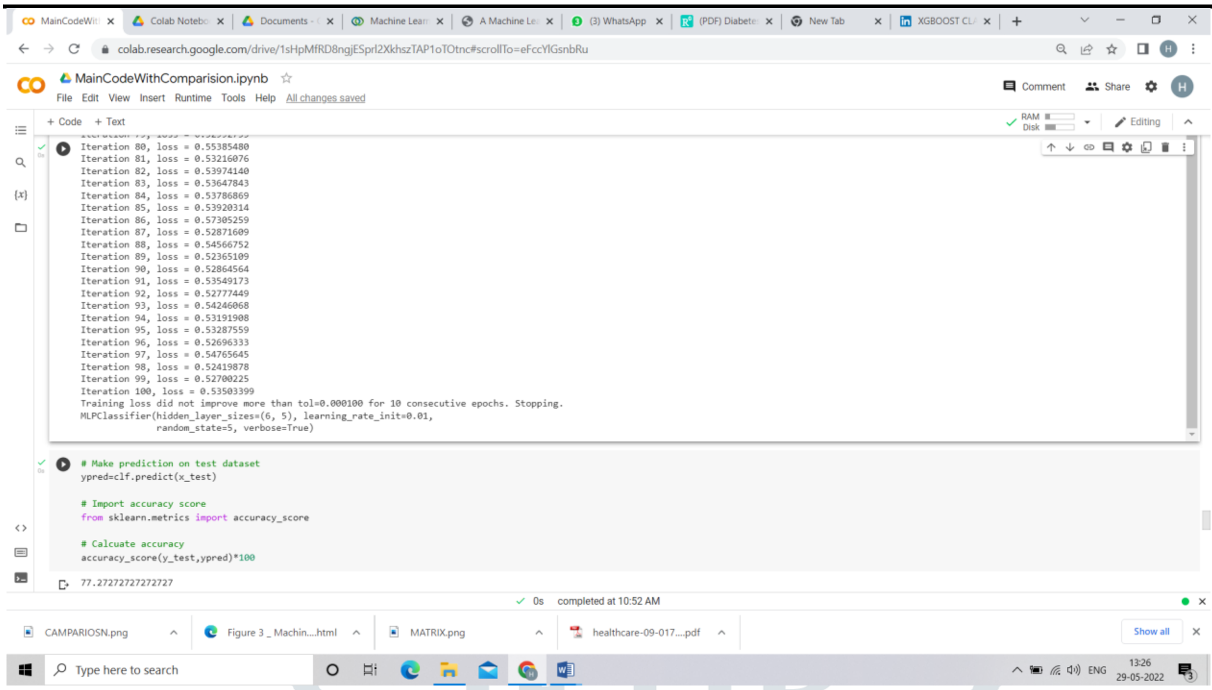


4.1.3. Multilayer Perceptron

For diabetes classification, we've fine-tuned multilayer perceptron in our experimental setup. It's a network where multiple layers are joined together to create a classification method, as shown in Figure 2. The building block of this model is perceptron, which could be a linear combination of input and weights. We used a sigmoid unit as an activation function shown in Algorithm 1. The proposed algorithm consists of three main steps. First, weights are initialized and output is computed at the output layer (δ_k) using the sigmoid activation function. Second, the error is computed at hidden layers (δ_h) for all hidden units. Finally, in an exceedingly backward manner, all network weights ($\omega_{i,j}$) are updated to cut back the network error.

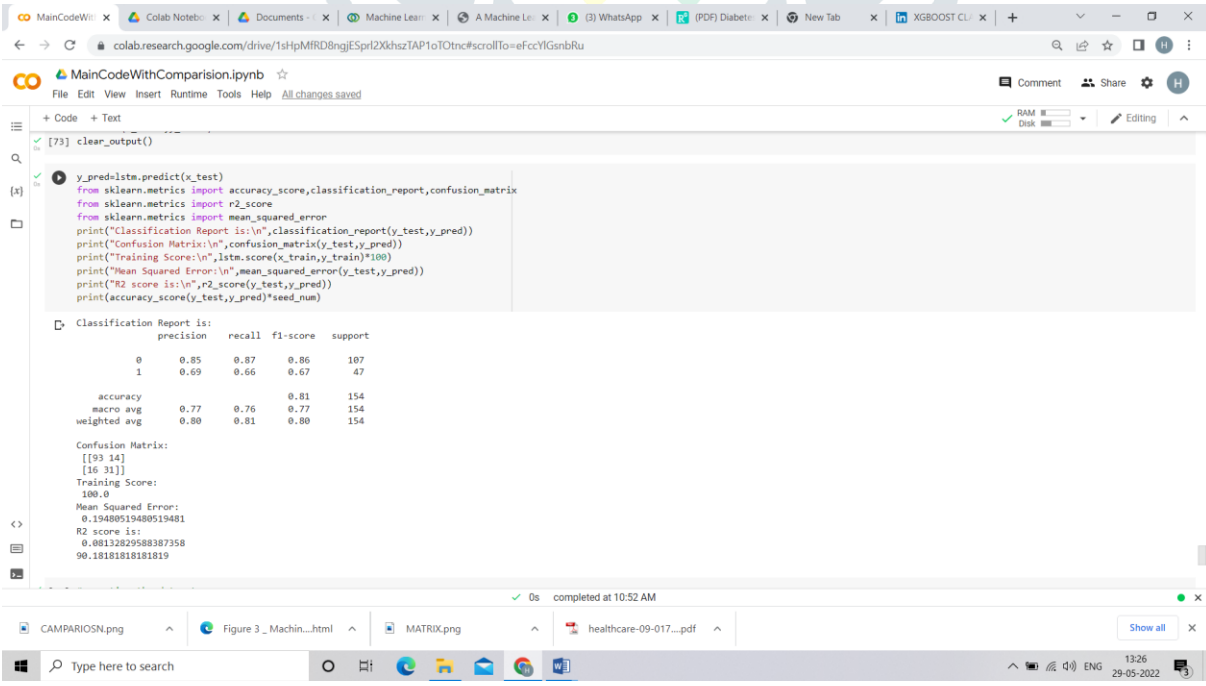
Figure 2 shows the multilayer perceptron classification model architecture where eight neurons are utilized in the input layer because we've eight different variables. The center layer is the hidden layer where weights and input are going to be a sigmoid unit. In the end, results will be computed at the output layer. Backpropagation is used for updating weights so that errors can be minimized for predicting class labels. For simplicity, only one hidden layer is shown in the architecture, which in reality is much denser.

Input data from the input layer are computed on the hidden layers with the input values and weights initialized. Every unit in the middle layer called the hidden layer takes the net input, applies the activation function "sigmoid" on it, and transforms the massive data into a smaller range between 0 and 1. The calculation is functional for every middle layer. The same procedure is applied to the output layer, which leads to the results towards the prediction of diabetes.**77.27%** accuracy is obtained by applying this algorithm.



4.1.4. Long Short-Term Memory

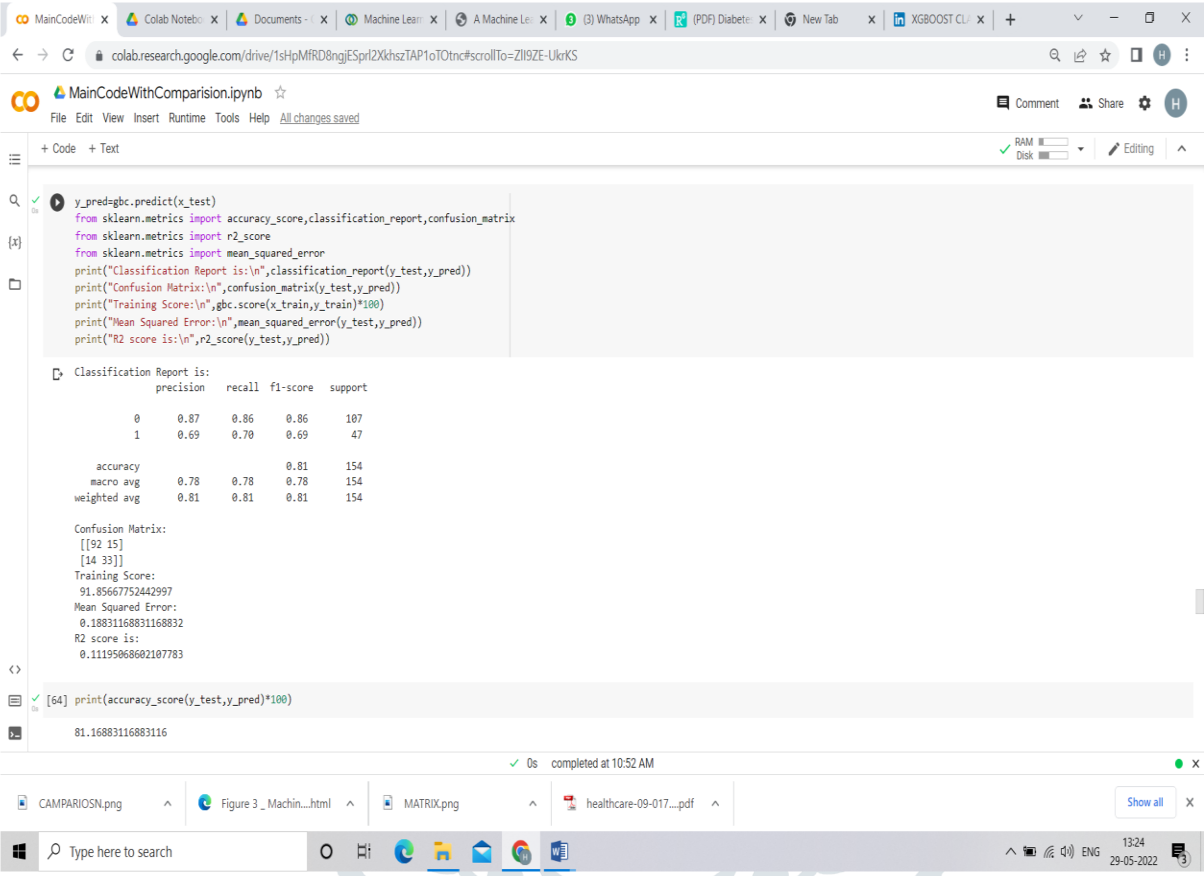
For diabetic forecasting, we've calibrated the long STM algorithm with our experimental setup. The proposed approach outperformed as compared to other state-of-the-art techniques implemented. LSTM relies on recurrent neural network (RNN) architecture and its feedback connections that make it suitable for diabetes forecasting. LSTM mainly consists of a cell, a keep gate, a write gate, and an output gate. The key behind using LSTM for this problem is that the cell remembers the patterns over an extended period, and three portals help regulate the data flow in and out of the system. **90.18%** accuracy is obtained by applying this algorithm.



4.1.5 Gradient Boosting

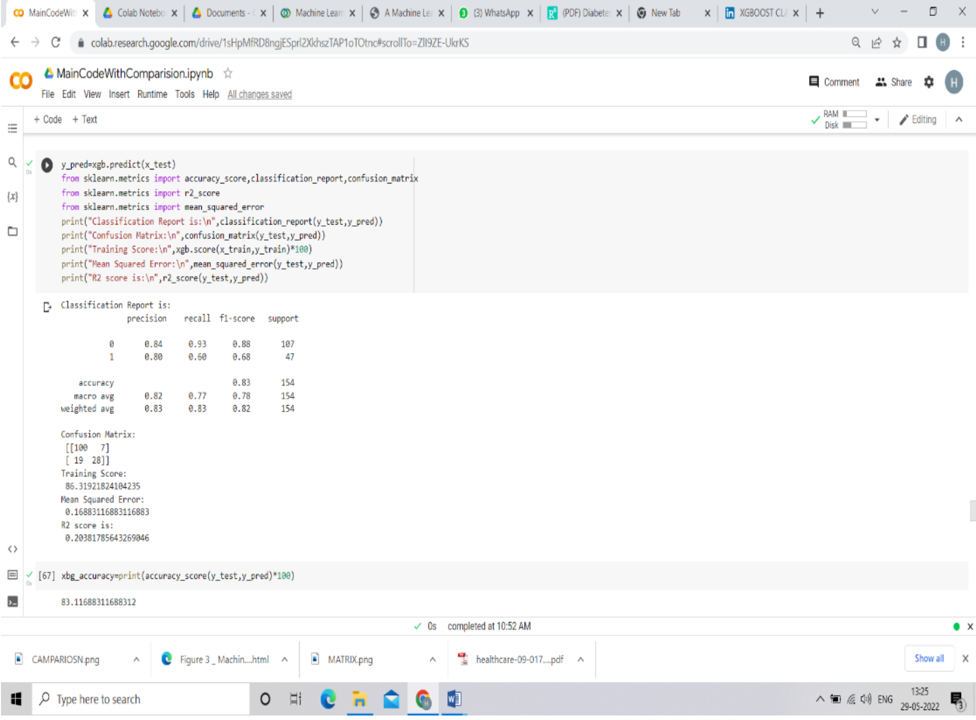
Gradient boosting could be a style of machine learning boosting. It relies on the intuition that the simplest possible next model, when combined with previous models, minimizes the prediction error. The key idea is to line the target outcomes for this next model to reduce the error. The target outcome for every case within the data depends on what proportion of changing that case's prediction impacts the prediction error:

- If a tiny low change within the prediction for a case causes a large drop in error, then the next target outcome of the case is a high value. Predictions from the new model that is close to its targets will reduce the error.
- If a small change in the prediction for a case causes no change in error, then the next target outcome of the case is zero. Changing this prediction doesn't decrease the error.

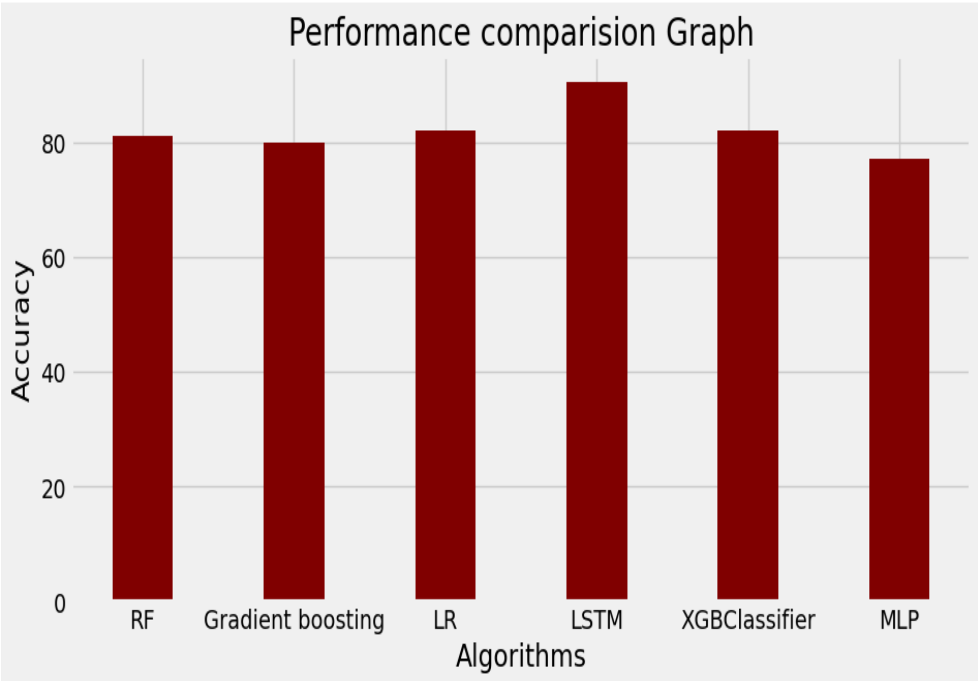


4.1.6 XG Boost

XG Boost classifier may be a Machine learning algorithm that's applied for structured and tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost is an extreme gradient boost algorithm. This means it's a giant Machine learning algorithm with plenty of parts. XGBoost works with large, complicated datasets. XGBoost is an ensemble modeling technique. **83.11%** accuracy is obtained by applying this algorithm. **81.16%** accuracy is obtained by applying this algorithm.



COMPARISON OF ALGORITHMS:



5. Algorithms for Implementation of Hypothetical System for Healthcare:

In the existing survey, they have detected and predicted diabetes by using the three different classifiers that have been employed, i.e., random forest (RF), multilayer perceptron (MLP), and logistic regression (LR). For predictive analysis, we have employed long short-term memory (LSTM), moving averages (MA), and linear regression (LR). For experimental evaluation, a benchmark PIMA Indian Diabetes dataset is used. During the analysis, it is observed that MLP outperforms other classifiers with 86.08% of accuracy and LSTM improves the significant prediction with 87.26% accuracy of diabetes.

By applying six various machine learning algorithms such as Random Forest (RF)-78.56, Logistic Regression (LR)-82.46, Multilayer Perceptron (MLP)-77.27, Long Short-Term Memory (LSTM)-90.18, gradient Boosting-81.16%, and XG-Boost-83.11%. Finally, by comparing all these six algorithms LSTM shows the highest accuracy of 90.18%.

6. Conclusion:

One of the important real-world medical problems is the detection of diabetes at its early stage. During this study, systematic efforts are made in designing a system that ends up within the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john's Diabetes

Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using the Decision Tree algorithm.

References:

1. World Health Organization, *Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World*, World Health Organization, Geneva, Switzerland, 2019.

2. R. Williams, S. Karuranga, B. Malanda, et al., "Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas," *Diabetes Research and Clinical Practice*, vol. 162, Article ID 108072, 2020.View at: Google Scholar

3. American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. Supplement 1, pp. S81–S90, 2014.View at: Publisher Site | Google Scholar

4. G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Calibration of minimally invasive continuous glucose monitoring sensors: state-of-the-art and current perspectives," *Biosensors*, vol. 8, no. 1, 2018.View at: Publisher Site | Google Scholar

5. N. N. Tun, G. Arunagirinatha, S. K. Munshi, and J. M. Pappachan, "Diabetes mellitus and stroke: a clinical update," *World Journal of Diabetes*, vol. 8, no. 6, 2017.View at: Publisher Site | Google Scholar

6. M. J. Davies, D. A. D'Alessio, J. Fradkin, et al., "Management of hyperglycemia in type 2 diabetes, 2018. a consensus report by the American diabetes association (ADA) and the European Association for the study of diabetes (easd)," *Diabetologia*, vol. 61, no. 12, pp. 2461–2498, 2018.View at: Publisher Site | Google Scholar

7. D. Bruen, C. Delaney, L. Florea, and D. Diamond, "Glucose sensing for diabetes monitoring: recent developments," *Sensors*, vol. 17, no. 8, 2017.View at: Publisher Site | Google Scholar

8. S. Wadhwa and K. Babber, "Artificial intelligence in health care: predictive analysis on diabetes using machine learning algorithms," in *Proceeding of the International Conference on Computational Science and Its Applications*, pp. 354–366, Springer, Cagliari, Italy, July 2020.View at: Publisher Site | Google Scholar

9. P. Tedeschi and S. Sciancalepore, "Edge and fog computing in critical infrastructures: analysis, security threats, and research challenges," in *Proceeding of the 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 1–10, IEEE, Stockholm, Sweden, June 2019.View at: Google Scholar

10. M. V. D. Schaar, A. M. Alaa, A. Floto, et al., "How artificial intelligence and machine learning can help healthcare systems respond to COVID-19," *Machine Learning*, vol. 110, no. 1, pp. 1–14, 2021.View at: Publisher Site | Google Scholar

11. M. H. Arnold, "Teasing out artificial intelligence in medicine: an ethical critique of artificial intelligence and machine learning in medicine," *Journal of bioethical inquiry*, vol. 18, no. 1, pp. 121–139, 2021.View at: Publisher Site | Google Scholar