Available online at http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 2, issue 6, pp. 668- 674, June -2022
10.53633/ijmasri.2022.2.6.008

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

EARLY-STAGE DIABETES PREDICTION USING DEEP LEARNING AND BOOSTING ALGORITHMS

Yatharth Kathuria[1], Vishal Modani [2], Sachin Garg [3],Varun Goel [4]
[1, 2, 3, 4] *Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi-110086,*
kathyatharth1999@gmail.com, vishalmodani647@gmail.com, sachingarg@mait.ac.in, varungoel.cs@gmail.com

Abstract

Diabetes has evolved as one of the most dangerous threats to the human world. Many are becoming its victims and are unable to come out of it even though they are working to avoid it from growing further. Cloud Computing and the Internet of Things (IoT) are two tools that play a very important role in today's life regarding many aspects and our poses including healthcare monitoring of patients and elderly society. Diabetes Healthcare Monitoring Services is very important nowadays because physically going to hospitals and standing in a queue is a very ineffective version of patient monitoring. If a patient has very chronic diabetes and it is not detected at an early stage, he/she might have to spend his/her time in long queues for a diagnosis which can be dangerous if left undetected in a long run. Diabetes can also act as a means for other diseases like heart attack, kidney damage, and somewhat blindness. This project makes use of various algorithms such as Perceptron, Artificial Neural Networks, Ada Boosting and Gradient Boosting with the help of which we can easily find out the accuracy of a model predicting that a person has diabetes or not.

Keywords: Deep Learning, Boosting, Diabetes, Analysis Metrics.

Introduction

Diabetes is a noxious disorder withinside the global. Diabetes is brought on due to weight problems or excessive blood glucose level, and so forth. It impacts the hormone insulin, ensuing in ordinary metabolism of crabs and enhancing ranges of sugar withinside the blood. Diabetes happens whilst the frame does now no longer make sufficient insulin. According to (WHO) World Health Organization approximately 422 million human beings be afflicted by diabetes mainly from low- or idle-earnings countries (Agrawal and Dewangan (2015). And this will be extended to 490 billion as much as the 12 months 2030 (Early stage diabetes risk prediction dataset. Data Set, Aug 2020)However, the superiority of diabetes is discovered amongst numerous Countries like Canada, China, and India, and so on. The populace of India is now greater than one hundred million so the real range of diabetics in India is forty million. Diabetes is a chief reason of dying withinside the global. Early prediction of sicknesses like diabetes may be managed and shop human lifestyles. To accomplish this, this painting explores the prediction of diabetes through taking numerous attributes associated with diabetes disorder (Mitushi Soni and Sunita Varma, 2020). For this purpose, we use the Dataset taken from the UCI device getting to know Repository (Islam, 2010). we observe numerous Deep Learning and Boosting Techniques to are expecting the opportunity of getting diabetes. Machine Learning is a way this is used to teach computer systems or machines explicitly. Deep Learning is the most exciting and powerful branch of Machine Learning that consists of several hidden layers made up of neurons. Various Machine Learning Techniques and other deep learning and boosting methods offer green effects to gather expertise through constructing numerous class and ensemble fashions from the gathered datasets. Such gathered records may be beneficial to are expecting diabetes. Various strategies can successful to do prediction, however, it’s difficult to pick the great method. Thus, for this purpose, we observe famous class and ensemble techniques to the dataset for prediction. Diabetes may be successfully controlled when caught early. However, whilst left untreated, it is able to result in capacity headaches that consist of heart damage. disorder, stroke, kidney damage, and nerve damage. Normally when you devour or drink, your frame will damage down sugars out of your meals and use them for electricity to your cells (Wang, 2019) . To accomplish this, your pancreas desires to supply a hormone referred to as insulin. The insulin is what allows the method of pulling sugar from the blood and setting it withinside the cells for use, or electricity. If you have diabetes, your pancreas both produces too little insulin or none at all. The insulin can’t be used successfully. This allows the blood organization disadvantaged of glucose ranges to upward push at the same time as the relaxation of your cells are wished electricity (Wang *et al.*, 2019).This can result in an extensive sort of troubles affecting almost each predominant body device. Diabetes also can have an effect on your pores and skin, the most important organ of your frame. Along with dehydration, your body’s loss of moisture because of excessive blood sugar can reason the pores and skin in your feet to dry and crack (Diabetes Daily, 2020). It’s vital to absolutely dry your ft after bathing or swimming. You can use petroleum jelly or mild creams, however keep away from letting those regions end up too1 moist. Moist, heat folds withinside the pores and skin are vulnerable to fungal, bacterial, or yeast infections (Diabetes Daily, 2020). These have a tendency to develop among arms and toes, the groin, armpits, or withinside the corners of your mouth. Symptoms consist of redness, blistering, and itchiness. High-pressure spots beneath Neath your foot can result in calluses. These can end up inflamed or expand ulcers. If you do get an ulcer, see your doctor right now to decrease the threat of dropping your foot (Diabetes Daily, May 2020)You will also be greater inclined to boils, inflamed nails.

Review of literature

Diabetes is one of the fastest-developing persistent lifestyles-threatening sicknesses which have already affected 422 million human beings international in step with the record of the World Health Organization (WHO), in 2018. Due to the presence of a highly lengthy asymptomatic segment, early detection of diabetes is constantly preferred for a clinically significant final result. Around 50% of everybody stricken by diabetes are undiagnosed due to its lengthy-time period asymptomatic segment. The early prognosis of diabetes is most effective feasible through right evaluation of each not unusual place and much less not unusual place signal signs, which may be discovered in exceptional levels from disorder initiation as much as prognosis. Data mining class strategies were nicely generic through researchers for the threat prediction version of the disorder. To are expecting the chance of having diabetes calls for a dataset, which includes the records of newly diabetic or would-be diabetic sufferers. In these paintings, we've got used this kind of dataset of 520 times, which has been gathered the usage of direct questionnaires from the sufferers of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and authorised through a doctor. We have analysed the dataset with Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm and after making use of tenfold Cross-Validation and Percentage Split assessment strategies, Random-forest has been discovered to have the great accuracy in this dataset. Finally, a generally accessible, person-pleasant device for the end-person to test the threat of getting diabetes from assessing the signs and beneficial hints to govern over the threat elements has been proposed (Ahmed, 2016). This mission pro- poses a powerful method for in advance detection of diabetes disorder. An assessment of the exceptional device getting to know strategies used on this examine exhibits which set of rules is great proper for the prediction of diabetes. Diabetes Prediction is turning into the place of hobby for researchers to teach this system to perceive the affected person is diabetic or now no longer through making use of right classifier at the dataset. Based on preceding studies paintings, it's been found that only some class algorithms aren't sufficient. Further, using cross-validation does now no longer enhance the very last accuracy of the bulk of the algorithms. Hence a device is needed for Diabetes Prediction is a vital place in computer systems, wherein all of the class algorithms are carried out and their accuracies are as compared to get pick the great-proper one.

Dataset description

S.No	Name of the attribute
1	Age
2	Sex
3	Polyuria
4	Polydipsia
5	Sudden Weight Loss
6	Weakness
7	Polyphagia
8	Genital Thrush
9	Visual Blurring
10	Itching
11	Irritability
12	Delayed healing
13	Partial paresis
14	Muscle stiffness
15	Alopecia
16	Obesity
17	Class

In this research, the early-stage diabetes risk prediction dataset collected from the UCI machine learning repository was used. The dataset was created through a direct questionnaire among diabetic and non-diabetic patients from the diabetes Hospital of Sylhet, Bangladesh. This dataset consists of 520 records and 17 attributes. Among them, 320 records are positive and 200 records are negative (Islam, 2010). A brief description of attributes is given in the below table.

This class variable shows the outcome of either 0 or 1 for diabetics which indicates positive or negative for diabetics

Research Methodology

This is the most important phase which includes model building for the prediction of diabetes. In this, we have implemented various algorithms for diabetes prediction. The procedure of Proposed Methodology is:

Step1: Import required libraries, Import diabetes dataset.

Step2: Encoding the dependent Variable, Splitting the dataset, and Feature Scaling.

Step3: Perform a percentage split of 75% to divide dataset as Training set and 25% to Test set.

Step4: Select the algorithm i.e., Perceptron, Artificial Neural Network, Ada Boosting (Decision Tree Classifier), Gradient Boosting.

Step5: Build the model for the mentioned algorithm based on the training set.

Step6: Test the model for the algorithm based on the test set.

Step7: Perform dimensionality reduction for visualization.

Step8: After analysing based on various measures conclude the best performing algorithm.

A. Data Pre-Processing

Data pre-processing is the most important process. Mostly healthcare-related data contains missing values, differences in the range of values, etc that can cause the effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data pre-processing is done. To use various techniques on the dataset effectively this process is essential for accurate results and successful prediction. In the data pre-processing, we have first imported sufficient libraries such as NumPy, pandas, matplotlib, TensorFlow and imported datasets as well. We have pre-checked for the null values and missing data in each of the attributes. We have Encoded the Categorical Data after which Splitting of the Data frame into Train and Test Sets is being done Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train the algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, the aim of normalization is to bring all the attributes under the same scale. Feature Scaling is further done to improve the training process and thus it will improve the final predictions.

B. Algorithms

B.1 Deep Learning Algorithms

- **Perceptron Learning Model:**

The Perceptron algorithm is **a two-class (binary) algorithm**. It is a type of neural network model, perhaps the simplest type of neural network model. It consists of a single node or neuron that takes a row of data as input and predicts a class label. It is the foundation of many modern neural networks such as ANN, CNN, etc. The module sklearn contains a Perceptron Class.

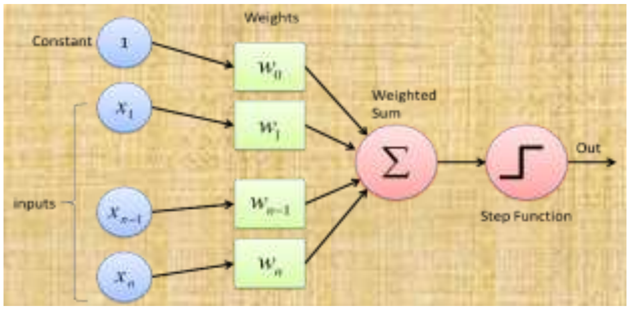
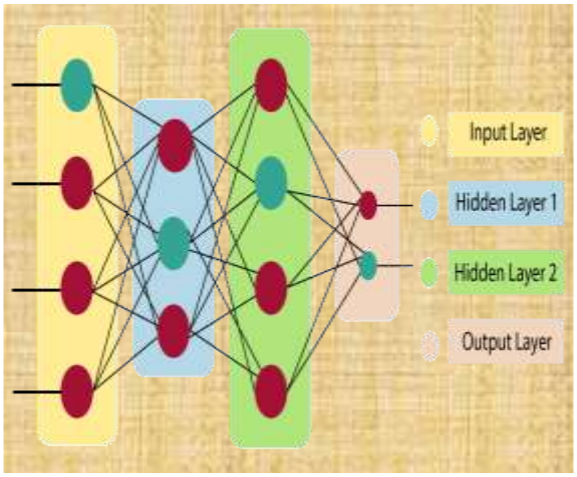


Fig. 1: Perceptron Learning Model

- **Artificial Neural Network (ANN):**

The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks (Maniruzzaman, 2018). ANN learns through the help of hidden layers that distills important pattern from the dataset analyze it and passes to the nest layer. Learning also involves updation of weights which is done by back propagation of error. The learning takes place till the error is minimum.



B.2 Boosting Algorithms

Ada-Boosting:

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method (Maniruzzaman, 2018) The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called **Decision Stump.** What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

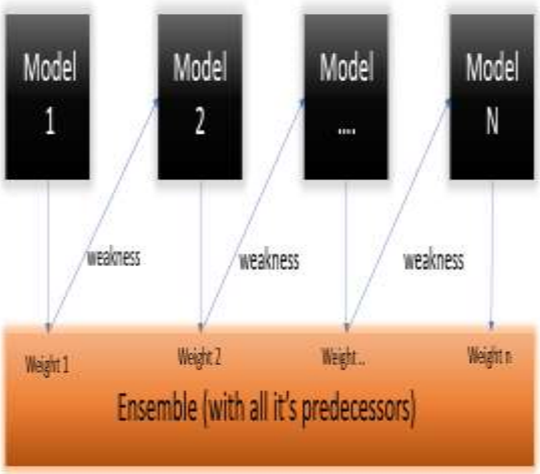


Fig. 3: Ada-Boosting

- **Gradient Boosting:**

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e., Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model. Unlike, Ada boosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us. The base estimator for the Gradient Boost algorithm is fixed and i.e., *Decision Stump*. Like, AdaBoost, we can tune the n_estimator of the gradient boosting algorithm. However, if we do not mention the value of n_estimator, the default value of n_estimator for this algorithm is 100. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

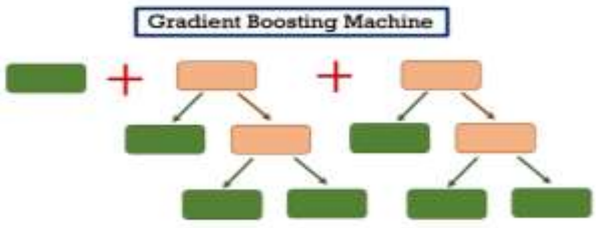


Fig 4: Gradient Boosting

C. Evaluation Criteria

It is the most common evaluation metric for classification and other problems. It is defined as the number of correct predictions against the number of total predictions. Also finding only the accuracy score sometimes is not enough. So, we also look at other performance metrics like Precision (measuring exactness), Recall (measuring completeness) , and the F1 Score (compromise between Precision and Recall).

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = 2 * Precision * Recall / (Precision + Recall)

Result

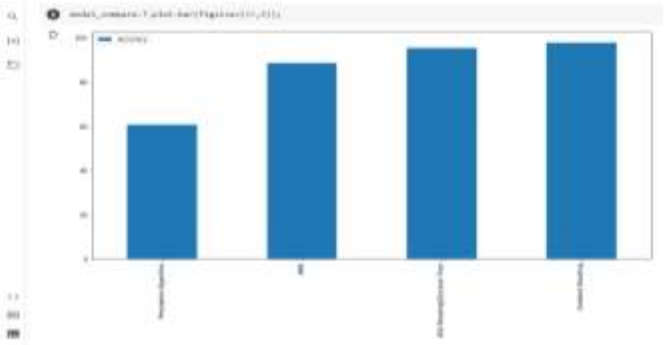
A. Final Results Analysis

- Finally, the accuracies of all the models are being compared using a bar graph with the help of pandas library and plot function.
- After a comparison of the accuracies, it is found that **Gradient Boosting Model** having **accuracy**

= **97.692** is the best one for "**EARLY-STAGE DIABETES PREDICTION**".

Name Of Algorithm	Accuracy	Precision	f1-score	Recall - score
Perceptron Model	60.769%	0.592	0.744	1.000
ANN	88.462%	0.864	0.903	0.946
Ada Boosting	95.385%	0.947	0.960	0.973
Gradient Boosting	97.692%	0.973	0.980	0.986

- Perceptron Model and ANN has a accuracy difference of about 28% which clearly states that the dataset taken is a non-linear dataset due to which there is such a low accuracy value for Perceptron model (as it uses a single neuron in comparison to the ANN model). So, ANN can be used in microwave modelling and optimization problems where perceptron is not much robust.
- Ada Boosting and Gradient Boosting have a difference of 2% for this particular dataset which is due to the loss function used in each of them i.e. exponential loss function in Ada Boost and differential loss function in gradient boosting.



B. Pros and Cons of Each Model

Algorithm	Pros	Cons
Perceptron Model	As recognized as Linear Binary Classifier, the perceptron model is extremely efficient and helpful in arranging the input data and classifying the same in different classes	The output values of a perceptron can take on only one of two values (0 or 1) due to the hard-limit transfer function, not applicable to non-linear separable dataset
ANN	Ability to work with incomplete knowledge, having fault tolerance, make use of multiple hidden layers to extract important pattern from input data.	Hardware dependence, determination of proper network structure.
Ada Boosting	Accuracy of weak classifiers can be improved	Need a quality dataset, sensitive to outliers
Gradient Boosting	No data pre-processing required, handles missing data.	Expensive Computation

Future Scope

A particular method to identify diabetes is not a very sophisticated way for initial diabetes detection and it is not fully accurate for predicting diseases. That's why we need a smart hybrid predictive analytics diabetes diagnostic system that can effectively work with accuracy and efficiency. So, after using data mining and machine learning classification algorithms for exploring and utilizing to support the medical decisions, which improves in diagnosing diabetic patients. Due to the dataset, we have till date is not up to the mark, we cannot predict the type of diabetes, so in the future, we aim to predict the type of diabetes and explore it, which may improve the accuracy of predicting diabetes. Along with that, various deep learning and boosting classification methods help to improve the efficiency of the diagnosis.

Reference

1. Diabetes, May 2020, [online] Available: https://www.who.int/newsroom/factsheets/detail/diabetes.
2. Maniruzzaman, M., Rahman, M. J. Al-Mehedi Hasan, M. (2018). "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers", *J Med Syst*, vol. 42, pp. 92.
3. Diabetes Daily, May 2020, [online] Available:https://www.diabetesdaily.com/learn-about-diabetes/what-isdiabetes/how-manypeople-have-diabetes/
4. Safial Ayon and Md. Islam, (2019). "Diabetes Prediction: A Deep Learning Approach", *International Journal of Information Engineering and Electronic Business*, vol. 11, pp. 21-27.
5. Wang, Q., Cao, W. Guo, J. Ren, J. Cheng, Y and Davis, D. N. (2019). "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values", *IEEE Access*, vol. 7, pp. 102232-102238.
6. Agrawal, P and Dewangan, A. (2015). "A brief survey on the techniques used for the diagnosis of diabetesmellitus", *Int. Res. J. Eng. Technol. (IRJET)*, vol. 02, no. 03, 2015.
7. Ahmed. Developing a predicted model for diabetes type 2 treatment plans by using data mining, (2016).
8. Islam, M. M. F., Ferdousi, R. Rahman, S and Bushra, H. Y. (2010). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques", *Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing*, vol. 992, 2020, [online] Available: https://doi.org/10.1007/978-981-13-8798-2_12.
9. Early stage diabetes risk prediction dataset. Data Set, Aug 2020, [online] Available:https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
10. Google Collaboratory, Aug 2020, [online] Available:https://colab.research.google.com/notebooks/intro.ipynb.
11. Mitushi Soni and Sunita Varma. (2020). Diabetes Prediction using Machine Learning Techniques, *International Journal Of Engineering Research & Technology* (Ijert) Volume 09, Issue 09.

*****