International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

Diabetes Disease Prediction using Regularized Gradient Boosting Framework

Nidhi Hardaha, Prof. Prateek Gupta
[1]Student, Shri Ram Institute of Science & Technology, Jabalpur, M.P.
[2] Prof, Shri Ram Institute of Science & Technology, Jabalpur, M.P.

Submitted: 20-03-2022 Revised: 27-03-2022 Accepted: 30-03-2022

ABSTRACT:
Data Mining performs a major role in healthcare services because disease recognition and investigation contains a vast amount of data. These conditions generate several data managing problems, and to operate efficiently. The healthcare datasets are undefined and influential and it is extremely monotonous to manage and to operate. To get better of the exceeding problems, numerous analyses present various ML algorithms for different disease examination and prediction. The undertaking of disease identification and prediction is an element of classification and forecasting.
Exploring important features of diabetes through analytical methods of data mining is able to predict and prevent diabetes. This paper presents a diabetes prediction algorithm based on XGBoost algorithm with the numerical features being separated while some important features are extracted from the text features of experiment data. Experiment results show that accuracy of diabetes prediction based the improved XGBoost algorithm with features combination is 80%, which is feasible and effective method for diabetes prediction.

Keywords: Diabetes prediction, Grid Search, XGBoost, PIMA, Machine Learning, BMI.

I. INTRODUCTION

Health care is one of those fields in the modern times which generates large amount of data. These data's if used technically in a better way can produce significant outcomes in research and engineering [1]. The records which are just patient information in the hospitals can be transformed into more useful data in data mining [2]. Diabetes is one of the most hazardous chronic diseases. The presence of glucose in the blood of a person regulates the body metabolism. If the glucose is too low it is termed as low glucose level in blood, likewise if it is high, blood is said to have high glucose levels, but having high glucose levels to a prolonged period stimulates the liver not to produce enough insulin that regulates glucose level in blood this is called as diabetes mellitus, simple diabetes. Diabetes mellitus is considered one of the deadliest yet common diseases in the world. It affects millions of people in the world and is becoming more and more common in India. A high sugar diet and other unhealthy eating habits and lifestyle choices, such as lack of regular physical activity is the cause of it. Genetics also play a major role in the onset of the disease. It is predicted that by 2034, the number of diabetes patients would reach around 592 million based on statistics calculated by the World Diabetes Federation [3].

Diabetes is characterised by insulin resistance, a major problem for the body. Due to insufficient levels of insulin in the body, which is triggered by hormonal imbalance and impact of pancreas, the body sugar levels do not stabilise. There are two types of diabetes: type 1 and type 2 diabetes. Type 1 diabetes is a far more serious condition than type 2 diabetes. In type 1 diabetes, the pancreas does not produce insulin at all, and the patient needs to take insulin through injections. In type 2 diabetes, the pancreas produces insulin, but it is not sufficient and medications to make the pancreas produce more insulin or injections need to be used. Type 2 is more common, almost 90% of the diabetes cases are type 2 [4] health problems.

Since this disease has been first detected, scientist of the different era was researching to cure of diabetic. But after researching by the different eras researchers, they all have to fail to find out the permanent solution for this disease. As a result, people can't cure their disease. In order get solution of this problem the data scientist came up with an approach. Their method was that as there are no processes to cure the diabetic. They can predict it by using different parameter. As a result, people can be caution by predicted. To do that researcher

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

collected data from the various people who have diabetic. Then apply different machine learning algorithm so that it can give the best accuracy. In our research work we have analyzed and then prediction the diabetic. Here, in the dataset, there are many attributes to classify the disease. Nowadays, a critical challenge for real-world medical problems is diabetes patients' diagnosis at an early stage [5-6].

Diabetes can cause many complications such as diabetic ketoacidosis, hyperosmolar hyperglycemic state, and death. Therefore, many research works have been conducted using real-world medical data to model and predict diabetes at an early stage [7-9]. In [10], the authors presented the importance of data quality in the therapeutic area. Authors proposed a 4- stage data pre-processing approach for missing values. Moreover, the SVM algorithm was used to classify diabetic patients' data and records. The authors implemented an optimized algorithm which was applied to the diabetes dataset to improve the accuracy, sensitivity, and specificity. In [11], researchers proposed a novel prediction model via improved K-means and logistic regression algorithms. The proposed model has proven to be appropriate for predicting type 2 diabetes mellitus. In [11], the authors improved accuracy and furthermore, the proposed model can adapted to any dataset. Authors showed that the proposed model has higher accuracy than other traditional schemes. . Therefore, it is useful for the realistic health management of diabetes. In [12], the authors proposed a new hybrid imputation method, Fuzzy principal component analysis (FPCA)– Support Vector Machine (SVM)– fuzzy c-means (FCM), to solve the problem of missing data in medical datasets.

In this research, our goal is to improve the accuracy of diabetic detection using one of the popular boosting model classifier named XGBoost. We also compare the accuracy of some machine learning algorithms like Logistic Regression (LR), naive Bayes, KNN and Random Forest (RF) by analyzing the dataset and compare their performance and to integrate these techniques in a system.

Modern medicine generates a great deal of information stored in the medical database. For example, medical data may contain MRIs, signals like ECG, clinical information like blood sugar, blood pressure, cholesterol levels, etc., as well as the physician's interpretation. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database increasingly becomes necessary. Data mining in medicine can deal with this problem. It can also improve the management level of hospital information and promote the development of telemedicine and community medicine. The goal of data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making.

Data mining methods may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which once evaluated and verified may be embedded within clinical information systems. The aim of preprocessing data is to remove the noise, outliers and to discover the important features existing in the raw data. Pre-processing stage includes cleaning, normalization, transformation, and feature selection. Learning becomes simpler if features are identified at pre-processing stage. The product of data pre-processing is the training set. With the training set, the learning model has to learn from it.

Diabetes is a non-communicable disease and increasing at an alarming rate all over the world. Having a high sugar level in blood or lack of insulin are the primary reasons. So, it is important to find an effective way to predict diabetes before it turns into a major problem for human health. It is possible to take control of diabetes on an early stage if we take precautions.

II. LITERATURE REVIEW

Classification can be described as a supervised learning algorithm in the Machine learning process. It assigns class labels to data objects based on prior knowledge of class which the data records belong. It is a Data mining technique, has made it possible to co-design and co-develop software and hardware, and hence, such components. However, integration deals with knowledge extraction from database records and prediction of class label from unknown data set of records. In classification a given set of data records is divided into training and test data sets. The training data set is used in building the classification model, while the test data record is used in validating the model. The model is used to classify and predict new set of data records that is different from both the training and test dataset. Supervised learning algorithm (like classification) is preferred to unsupervised learning algorithm (like clustering) because its prior knowledge of the class labels of data records makes feature/attribute selection easy and this leads to good prediction/classification accuracy.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

Some of the common classification algorithms used in Data mining and decision support systems are: Neural networks, Logistic regression, Decision tree etc. Among these classification algorithms Decision tree algorithms is the most commonly used because it is easy to understand and cheap to implement. It provides a modeling technique that is easy for human to comprehend and simplifies the classification process. Most Decision tree algorithms can be implemented in both serial and parallel form while others can only be implemented in either serial or parallel form. Parallel implementation of decision tree algorithms is desirable inorder to ensure fast generation of results especially with the classification/prediction of large data sets; it also exploits the underlying computer architecture. But serial implementation of decision algorithm is easy to implement and desirable when small-medium data sets are involved.

A. Logistic Regression

Logistic Regression [12] is a classification algorithm for the probability of occurrence of an event, whether that event will occur or not. It is used to portray a binary or a categorical outcome with only 2 classes. It is similar to linear regression with the only difference being that the outcome of the variable is categorical instead of a continuous variable. It uses Logit Link function, in which the data values are fitted, for prediction. The mathematical interpretation defines Logit function as the natural log of the odds that Y equals one of the categories [12]. If p is the probability then, the logit function for p is defined as:

Logit (p) = ln (p/1-p)

B. Decision Tree

Similar to the tree analogy in real life, the Decision tree is a machine learning algorithm, used for both classification and regression analysis. It is a tree-like graph beginning with a single node, and branching into its possible outcomes. Unlike the linear models, a decision tree is a supervised learning that maps non-linear relationships as well. The data sample is divided into homogeneous subsets based on the most notable splitter in input attributes. The splitter is identified using various algorithms such as Gini Index, Chi-Square, Information Gain and Reduction in Variance.

C. Random Forest

Before we start let us give brief information about the random forest classifier. It is a flexible, easy to use machine learning algorithm that produces, even without hyper- parameter tuning, a great result most of the time [13]. Moreover, it is also known to be one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks. Here down below we are going to show how the random forest algorithm works and several other important things about it.

Random forest is a supervised learning algorithm. Like you will already see from its name, it creates a forest and makes it somehow random. The 'forest' it builds, is an ensemble of decision trees, most of the time trained with the 'bagging' method. The general plan of the textile technique is that a mixture of learning models will increase the result. If you want to say it in simple words: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction [13]. One massive advantage of random forest is that it is used for each classification and regression issues, which form the majority of current machine learning systems. Random forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, you do not have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. We already have told that with random forest we can also deal with regression tasks by using the random forest repressor. Random Forest adds further randomness to the model whereas growing the trees. Instead of sorting out the foremost necessary feature whereas rending a node, it searches for the best feature among a random subset of features. This ends up in a good diversity that typically ends up in a more robust model. So, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random. In addition to that using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

D. Support Vector Machine

Before we start discussing about SVM (support vector machine) we need to be accustomed with linear regression and Logistic regression algorithms. If not it is suggested to look at them before moving on to support vector machine. Support vector machine is another simple algorithm that every machine learning expert should have in his or her arsenal. In this scenario Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Moreover support vector machine, abbreviated as SVM can be used for both

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

regression and classification tasks. But it is widely used in classification objectives [14]. Now let us a bit more about what is support vector machine. The objective of the support vector machine algorithm is to find a hyper plane in an N-dimension space (N – the number of features) that distinctly classifies the data points [14]. Now in order to separate the two classes of data points there are many possible hyper planes that could be chosen. Here hyper planes are decision boundaries that help classify the data points. Data points falling on either aspect of the hyper plane will be attributed to completely different categories. Here one more thing we would like to add is that the dimension of the hyper plane depends upon the number of features. If we can find the number of input feature is 2 then the hyper plane is just a line. If the number of input feature is 3 then the hyper plane becomes a two dimensional plane. From here what we can understand is that it becomes very difficult to imagine when the number of feature exceeds 3. We should always remember that support vectors are data points that are close to the hyper plane and influence the position and orientation of the hyper plane. Using these support vectors, we maximize the margin of the classifier [15]. Now if we delete the support vectors it will change the position of the hyper plane. These are the points that will eventually help us build our SVM. Now we are going to talk a bit about the large margin intuition. To start we can say that in logistic regression we take the output of the linear function and squash the value within the range of [0, 1] using the sigmoid function. If the squashed value is greater than a threshold value (.5) we assign it as label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold value are changed to 1 and -1 in SVM, by doing this we can obtain this reinforcement range of values ([-1, 1]) which acts as margin.

Support Vector Machines, also called Support Vector networks are supervised learning algorithms used for both classification and regression analysis. It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of margin between the hyperplane. There are different kernels available for mapping of linear or no linear data points in a multidimensional space for separation. For our analysis, we have used only the Linear and Radial basis function as kernel.

E. Adaptive Boosting

Adaptive Boosting (AdaBoost) formulated by Yoav Freund and Robert Schapire [16] is a machine learning algorithm used for classification as well as for regression analysis. It involves the conversion of a weak classifier into a strong one using the ensemble technique. For this purpose, the predictions of each weak classifier are merged using weighted average or by taking into account their prediction accuracy as a metrics. Initially, all the attributes are given equal weights, and then the algorithm assigns a higher weightage to the inaccurate observation [17]. The error is then propagated with every prediction and multiple iterations are done to reduce it until the prediction become accurate.

Our motive behind surveying papers is to get an idea how warn a human being beforehand if he/she is going to have any diabetic disease. All the possible algorithms regarding data mining have been analysed and studied deeply so that the diabetic disease prediction provides maximum accuracy with minimum attributes used. Moreover, focus is made on choosing the correct algorithm that provides accurate results. But the main problem of Data mining is using different algorithms for detection of diabetic disease. Some algorithms are diagnose is less accurate and time consuming. So we propose a method that will detect the diabetic disease with maximum accuracy.

III. PROPOSED WORK

XGBoost (Extreme Gradient Boosting) is also referred by gradient boosting, stochastic gradient boosting, multiple additive regression trees or simple gradient boosting machines. It is a kind of supervised machine learning algorithm which can be regarded as an improved version of gradient boosting machine. Boosting as also is an ensemble technique which leverage the errors made by existing models by correcting them until no errors can be corrected by adding models sequentially. XGBoost models are based on the technique wherein we predict the errors of models are predicted by newer models which are then added together to make a final assessment of the prediction. XGBoost algorithm is called gradient boosting because it particularly minimizes the loss when adding new models using the gradient decent algorithm.

It takes classification and regression tree (CART) as the base learner. Basic structure is shown in figure below in which X is the spectral absorbance matrix in this model and y is the concentration of a certain metal ion. According to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

the additive training strategy of boosting, each tree is constructed based on learning from the residual δ of the previous tree.

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i)$$

is the prediction of the k-th iteration. At every iteration, XGBoost optimizes the model and decreases the prediction error. The final prediction output is generated by the weighted summation of trees as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F},$$

where F is the space of functions containing all regression trees; K denotes the number of trees. To learn function fk of each tree, XGBoost establishes an objective function with regularization:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

Where f is all learnable parameters in XGBoost; $l(y_i, \hat{y}_i)$ is the loss function representing the error between the predicted concentration $\hat{y}_i$ and the actual concentration y_i, the smaller the l is, the better the performance of the algorithm; $\Omega(f_k)$ is the regularization term to penalize the model complexity and prevent over-fitting. When XGBoost uses the square loss function to measure error, the second derivative Taylor expansion of the loss function can assist the model to optimize the objective quickly. The second derivative Taylor expansion of the loss function after k-th iteration is given as follows:

$$\mathcal{L}(\phi)^{(t)} \simeq \sum [l(y_i^{(t)}, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t),$$

Where gi and hi are the first and second derivative of the loss function. It can be learned that the loss function only depends on the first and second derivatives of each data point. To predict the ion concentrations, the essential step in the XGBoost learning algorithm is to optimize the XGBoost algorithm parameters, booster parameters, and learning parameters.

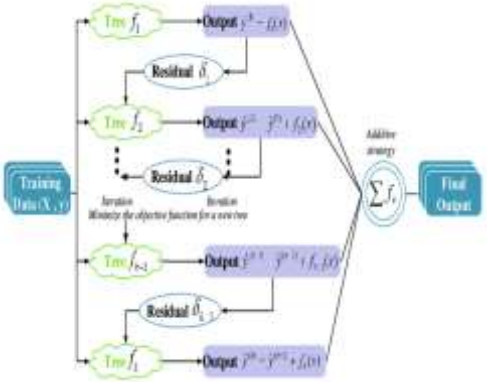

Figure 4.2: Proposed Working Model.

Proposed Algorithm

Input: Dataset Containing values of Symptoms.
Output: 1 (Diabetics) or 0 (Not Diabetics).

Proposed Algorithm ( )
Step 1: Read the database.
Step 2: Apply preprocessing.

- 2.1: Count all the entries.
- 2.2: Find mean of all the columns.
- 2.3: Find min and max values in the dataset for individual features.
- 2.4: Apply Scaling Transformation.
- 2.5: Remove abnormal features.
- 2.6: Extract the keywords.

Step 3: Select the parameters for tuning.
Step 4: Get the best hyperparameters.

- 4.1: Apply Grid Search.
- 4.2: Set the value of gamma (between 0 to 0.1).
- 4.3: Set the values of eta. (In between 0.1 to 0.3).

Step 5: Start creating trees with above values.
Step 6: Final tree is created using eta.
Step 7: Calculate the Accuracy.
Step 8: Print the result.
Step 9: End of algorithm.

IV. RESULTS AND EVALUATION

In Proposed system has been evaluated on various parameters with various existing classification methods like logistic regression, support vector machine, random forest & k nearest neighbour. It is found that extreme gradient boosting (XGBoost) method achieves good results as compare to others.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1211-1217 www.ijaem.net ISSN: 2395-5252

Classifier	Accuracy (in %)
LR	78.13
SVM	77.60
KNN	75.52
Proposed	**80.00**

Table 5.1: Performance evaluation.

V. CONCLUSION

The main motivation of this thesis is to provide an insight about detecting and curing diabetic disease using data mining technique. For this thesis, data were collected from PIMA Data Sets. These datasets are fed in to Naive Bayes, SVM, KNN, Decision Tree and Random forest and proposed method for diabetic prediction, in which proposed method gave the best result with the highest accuracy. Valid performance is achieved using XGBoost algorithm in diagnosing diabetic diseases and can be further improved by increasing the number of attributes.

Thus, in an environment similar to that of the used dataset, if all the features are preprocessed such that they acquire normal distribution, XGBoost is a good selection to obtain a robust prediction model. And, such models provide a valuable assistant to the society for health care management domain.

REFERENCES

[1] Jothi, Neesha, and Wahidah Husain. "Data mining in healthcare–a review." Procedia Computer Science 72 (2015): 306-313.

[2] Raghupathi, Wullianallur. "Data mining in healthcare." Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management (2016): 353-372.

[3] Nongyao Nai-arun, Rungruttikarn Moungmai, “Comparison of Classifiers for the Risk of Diabetes Prediction”, Procedia Computer Science 69, pg 132 –142, 2015.

[4] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15, pg 104–116, 2017.

[5] Y. Xiaodong, D. Fan, A. Ren, N. Zhao, S. A. Shah, A. Alomainy, M. Ur-Rehman, and Q. H. Abbasi, "Diagnosis of the Hypopnea syndrome in the early stage," Neural Computing and Applications, vol. 32, no. 3, pp. 855-866, 2020.

[6] M. B. Khan, Y. Xiaodong, A. Ren, M. A. M. Al-Hababi, N. Zhao, L. Guan, D. Fan, and S. A. Shah, "Design of software defined radios based platform for activity recognition," IEEE Access, vol. 7, pp.31083-31088, 2019.

[7] S. K. Somasundaram and P. Alli, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," Journal of Medical Systems, vol. 41, no. 12, pp.201, 2017.

[8] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images,” Computer methods and programs in biomedicine, vol. 157, pp.121-128, 2018.

[9] A. Choudhury and D. Gupta, "A survey on the medical diagnosis of diabetes using machine learning techniques, "Recent Developments in Machine Learning and Data Analytics, Springer Singapore, vol. 740, pp. 67-78, 2019.

[10] M. Alirezaei, S. T. A. Niaki, and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines," Expert Systems with Applications, vol. 127, pp. 47-57, 2019.

[11] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp.100-107, 2018.

[12] S. Sperandei, “Lessons in biostatistics Understanding logistic regression analysis,” Biochem. Medical, vol. 24, no. 1, pp. 12–18, 2014.

[13] L. Verma, S. Srivastava, and P. Negi, “A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,” Journal of medical systems, vol. 40, no. 7, p. 178, 2016.

[14] V. Jakkula, “Tutorial on Support Vector Machine (SVM),” Sch. EECS, Washingt. State Univ., pp. 1–13, 2006.

[15] N. Cristianini and J. Shawe-Taylor, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,” vol. 22, no. 2, pp. 103–104, 2000.

[16] Y. Freund and R. Schapire, “A Tutorial on Boosting,” pp. 1–35, 2013. [11] R. Rojas, “AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting,” Writing, pp. 1–6, 2009.

[17] Liaqat Ali, Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilarz, And Xiong Xingzhong," An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease', 2169-3536 (c) 2018 IEEE.