Research Article

Prediction and diagnosis of future diabetes risk: a machine learning approach

Roshan Birjais[1] · Ashish Kumar Mourya[1] · Ritu Chauhan[2] · Harleen Kaur[1]

Abstract

Machine learning is a subset of Artificial Intelligence when combined with Data Mining techniques plays a promising role in the field of prediction. We live in an era where data generation is exponential with time but if the generated data is not put to work or not converted to knowledge data, its generation is of no use. Similarly, in Healthcare also, data availability is high, so is the need to extract the information from it for better prognosis, diagnosis, treatment, drug development, and overall healthcare. In this research, we have tried to focus more on diagnosis of Diabetes disease, which is one of the fastest growing chronic diseases all over the world as declared by World Health Organization in the year 2014. We have also tried to show the different techniques like Gradient Boosting, Logistic Regression and Naive Bayes, which can be used for the diagnosis of diabetes disease with attained accuracy as 86% for the Gradient Boosting, 79% for Logistic Regression and 77% for Naive Bayes.

Keywords Machine learning · Gradient boosting · Logistic regression · Naive Bayes · Diabetes

1 Introduction

Machine learning is a discipline where machines are instructed without human intervention by mean of algorithms. We can train them to perform a particular task and based on the training they can be used to handle the similar job without being explicitly programmed. Training the machine with some algorithm and feeding them with the dataset results into the formation of a classifier and to find the accuracy of the classifier we test it in the testing phase. Accuracy is always major issue in medical science and with different algorithms we can obtain different accuracies on same data set and it is very important to see which algorithm provides best result in order to obtain better classifier to do the better classification. Machine learning can be used in almost every field nowadays. Using it in the field of medical science can prove to be very beneficial in the improvement of healthcare [1].

Health care is always a big issue for any nation and is always challenging thing to provide. Better, the health care of a nation better is the condition of the inhabitants living there. Improvement in the health care can directly result in economic growth because a healthy person can prove to be a big asset to the nation and can conduct activities effectively in the workforce than any unhealthy person. Health care is aggregation and integration of all the measures, which can be taken to improve the health system. Healthcare constitutes prevention, diagnosis and treatment. Improving healthcare should be the first priority and major task to ponder on. Usage of technologies in the improvement of the health care is proved very beneficial [2, 3]. Machine learning can help to prevent, detect and treat various health conditions. Machine learning and data mining techniques are the best sources to improve healthcare system. Manual detection or diagnosis done by doctors is time consuming and lacks accuracy. Machines which are made to learn how to detect diseases using machine

✉ Harleen Kaur, harleen.unu@gmail.com | [1]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India. [2]Department of Biotechnology, Amity University, Noida, U.P., India.

SN Applied Sciences (2019) 1:1112 | https://doi.org/10.1007/s42452-019-1117-9

Received: 10 May 2019 / Accepted: 17 August 2019 / Published online: 28 August 2019

learning and data mining can better diagnose the problem and that too with high accuracy. Machine Learning (ML) not only can prove beneficial in diagnosis and prognosis of a disease but can also prove to be helpful in personalised treatment and behavioural modification, manufacturing of drugs and discovering new patterns resulting in new medications and treatments, clinical trial research and smart electronic health records. With machine learning we can do all this and in a better way [4].

2 Literature survey

Patients with various health related problems lead to severe discomforts, costly treatments, disabilities and much more. Prognosis of which helps healthcare providers to provide better and preventive measures to improve patient safety, and quality of care while reducing the medical costs. A multilayer neural network trained by Levenberg–Marquardt algorithm and a probabilistic neural network is used for diabetes disease diagnosis [5]. The authors compared the result with previous studies of the same data sets provided in UCI repository and found ANN could be successfully used in diagnosis of pima-diabetes. in their research multi-layer neural network (MLNN) with linear model (LM) (10× FC) provides an accuracy of 79.62% and probabilistic neural network (PNN) (10× FC) with 78.05%. Using conventional valid MLNN with LM showed 78.13% accuracy that is huge success compared to the previous studies. Another paper on same dataset using different methods showed different percentage of accuracy [6]. Algorithms like expectation maximization (EM), k-nearest neighbor (kNN), K means, Amalgam kNN, adaptive neuro fuzzy inference system (ANFIS) when fed with PIMA diabetes data set gave good results as shown in the below table. Pima Indian Diabetes data set results in a huge number of association rules which can then be deduced on the basis of the coverage and percentage of accuracy [7].

Machine Learning not only can be used for the prognosis and diagnosis but also has many other applications in medical field. Computer based expert system for recording and interpreting patient's context, is not only helpful for patients but also for doctors. (1) METABO is such a system for diabetic monitoring and management system [8]. This system consists of a patient's mobile device (PMD), (2) different types of unobtrusive biosensors, (3) a Central Subsystem (CS) located remotely at the hospital and (4) the Control Panel that helps in comprehensive monitoring of the actions of the patient and disease related data. Breault et al. [9] shows rising growth of diabetes in US which makes it very important issue to ponder on. Different classification technique [10, 11] like decision tree (DT), support vector machine (SVM), Naive Bayes (NB), decision stump (DS), when evaluated for performance check without boosting gave accuracy as 76%, 79.68%, 78.1%, 74.47% respectively and the performance evaluation of Adaboost with DT, SVM, NB, DS as base classifier resulted in the accuracy improvement except SVM which does not show any accuracy development as mentioned in the Table 1 below.

We live in an era where data is generated exponentially which results into summation of huge data. Especially in health care data availability is high but need for extracting knowledge out of it is also high otherwise collection of abundant health care data is waste if not put to use. Data mining and machine learning helps in finding the useful information, which can be further used extensively. In health care, data mining and machine learning can be used to make patient care better, best practices, effective treatments can be provided to the patient, fraud detection and more affordable healthcare services [12]. Data mining can also be used to detect outburst of plagues before time (prediction) by observing the trends in the symptoms/complaints of patients.

3 Material and methodology

In this paper we have used three techniques of machine learning –Gradient boosting, logistic regression and Naive Bayes to do the better diagnosis of diabetes disease. Using these three algorithms on Pima Indian diabetes data set, we can do the diagnosis whether the person is diabetic (1) or non-diabetic (0). With minor changes in the life style

Table 1 Analysis of various methods on diabetes dataset (Pima Indian diabetic set)

References	Objective	Method/technique	Analysis
[1]	Diagnosis of diabetes on pima-diabetes	MLNN, PNN	Using 10× FC (MLNN = 79.62, PNN = 78.05) Otherwise PNN = 78.13)
[6]	Prediction and diagnosis of diabetes	EM, KNN, K means, Amalgam KNN, ANFIS with adaptive KNN	< 70%, 73.17%, 66–77%, > 80%, respectively
[7]	Prediction and diagnosis of diabetes mellitus	DT, SVM, NB, DS	76%, 79.68%, 78.1%, 74.47% respectively.
[8]	Classification of diabetes disease	LS-SVM, GD-LS-SVM	78.21%, 82.05%

and in the eating habits, pre diabetic patients can be prevented from being diabetic, if not forever but at least for some duration of their lifetime.

3.1 Diabetes

One of the fast growing chronic diseases in the world is Diabetes. According to WHO, the number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014 [13] which means 8.5% of world's population has diabetes which was just 4.7% in 1980. Considering its fast rising growth, diagnosis and prognosis is necessary.

Diabetes is considered a disease in which the glucose level in the blood gets increased because of its non outreach to the cells present in the body. Blood glucose is formed because of the food you eat and is considered the main source of energy. However, increased level of blood glucose (also called as blood sugar) can lead to many diabetic related health problems. Similarly, the decreased level of blood sugar can also lead to hypoglycaemia or other sugar problems. To maintain the optimal glucose level in the body, two pancreatic endocrine hormones—Insulin and Glycogen play an important role. Both of these hormones are secreted by islet cells [14] present in the pancreas in response to the blood sugar levels.

Islet cell has two types—Beta cells and Alpha cells. Insulin is secreted by the beta cells [15]. The prompt of insulin secretion is regulated by the presence of glucose level in the blood. It increases with the increase in the blood sugar level and decreases with the decrease in the blood sugar level. Insulin enables muscle, fat cells and red blood cells to absorb glucose from the blood to use it for energy, hence reducing the high blood glucose levels into the normal range. Diabetes is of two types—Type 1 and Type 2.

In type 1 diabetes due to insufficiency in the production of insulin or there may be no production at all because of which cells are not able to absorb glucose which result into excess glucose in the blood and can lead to many problems like heart attack, renal failure etc.

In type 2 diabetes, insulin is produced in small quantity, which is not usually sufficient, or there exists insulin resistance thus glucose level remains high in blood which results into many fatal complications.

3.2 Data set description

Pima Indians diabetes data set is available in UCI machine learning repository and this set has 768 instances and 8 number of attributes [16]. Objective of this data set is diagnosis of diabetes of Pima Indians living in America. Based on the various attributes provided in this data set, this paper shows whether a person is diabetes positive or not using Gradient Boosting, Logistic Regression and Naive Bayes. In this database, all patients are females and are of age at least 21 years. The total numbers of instances in the database are 768 out of which 268 are diabetic and 500 are non-diabetic and are described as 1 and 0 respectively in the class attribute. This database is actually owned by the National Institute of Diabetes, Digestive, and Kidney Diseases. Till 2011 the diabetes data set present in UCI machine learning repository was considered to have no missing values but later it is found that in place of missing values there are zeros and having values as zeros at these places is biologically impossible such as in age or blood pressure attributes (Table 2). Also zero body mass index, 0-plasma glucose and 2-h serum insulin contains almost 50% impossible values. Attributes in this database is either integer or real [17].

A good data set is one in which the features are highly correlated to the target class and are highly uncorrelated to each other. To find the uncorrelated attributes feature selection is done through a correlation-based approach. Figure 1 shows age, BMI, and glucose are important features for the diagnosis of diabetes disease and are highly uncorrelated to each other.

Table 2 Characteristics of Pima Indians diabetes

Attribute number	Attribute name	Attribute description	Attribute type and measurement
1	Pregnancy	Number of times the female is pregnant	Numeric
2	Plasma glucose	Measured using a 2 h oral glucose tolerance test	Numeric
3	Blood pressure	Diastolic blood pressure	Numeric (mm Hg)
4	Triceps	Triceps skin fold thickness	Numeric (mm)
5	Serum insulin	2 h serum insulin	Numeric (μ U/ml)
6	BMI	Body mass index	Numeric [weight in kg/(height in m)^2]
7	Pedigree	Diabetes pedigree function	Numeric
8	Age	Age of patients	Numeric (years)

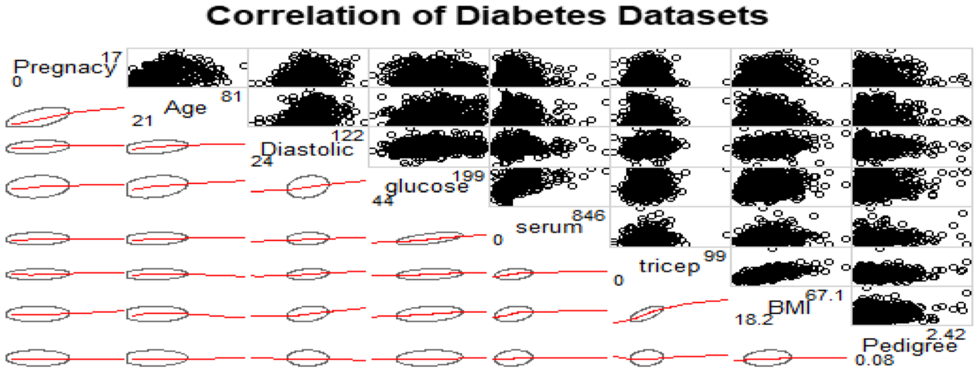

Fig. 1 Correlation between the independent attributes of the diabetes data set

3.3 Data pre-processing

Chances are very high that the secondary data obtained from any repository has missing values or may contain outliers as well. When the data set is medical dataset then chances of having missing values in the data increases. Also proper selection of attributes is very important aspect for building model free from correlated variables, biases and unwanted noises. Many pre-processing techniques are given in [18, 19]. We have used Boruta package in R for variable selection in order to get the better performance. Boruta works as a wrapper algorithm around Random Forest.

To handle missing values effectively in the data set there are many ways of doing it. Impact of pre-processing on data increases the performance level of a model to a great extent [20]. We have used KNN imputation for predicting values and to reduce bias. KNN imputation uses K-Nearest Neighbour approach to impute missing values in a data set. It uses Euclidean distance to identify 'K' closest observations and computes the weighted average of these 'K' observations. KNN imputation is used for data that are discrete, continuous, categorical, and ordinal, which makes it better technique in dealing with all kinds of missing data.

3.4 Machine learning models

Diagnosis of diabetes is a binary classification problem, which means we need to analyse whether a patient is suffering from diabetes or non-diabetic (1 or 0) based on attributes available in the Pima Indians diabetes dataset. Using all the above-mentioned attributes three techniques are used for the process of diagnosis of diabetes. The techniques are Gradient Boosting, Logistic Regression, and Naive Bayes. These techniques are categorised into the supervised learning of machine learning. Supervised learning [21] is labelled learning which means the output belongs to some labelled class—diabetic or non diabetic.

3.4.1 Gradient boosting

Gradient Boosting algorithm is a supervised learning technique (machine learning) based on gradient descent method [22]. It can be used for classification as well as regression. It has ability to reduce variance and bias, and helps weak learner to become stronger. This model improves prediction accuracy by adding additional trees to correct mistakes made by previous base models [23, 24]. Its principle is based on the Optimization technique that finds the values of coefficients of (*f*) function that minimize a (*c*) cost function with differential loss function. Like other boosting algorithms, a set of weak learners become ensemble to a single strong learner by sequentially. In gradient boosting, each tree is build one by one by making correction in error rate of previous ones where as in random forest trees are build randomly. Gradient boosting is time consuming in comparison to random forest.

Suppose we have the training sample $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$ and loss function is $\mathfrak{L}(y, \mathcal{F}(x))$ with *l*th number of iterations. Now applying empirical risk minimization principle (Eq. 1) with the model we get

$$\mathcal{F}_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right) \tag{1}$$

3.4.2 Logistic regression

Logistic regression also called as logit regression or even logit model is another supervised learning technique [18, 25, 26, 27] from the field of statistics borrowed by machine learning which a predictive analysis. It is a classification algorithm which means the output it provides is discrete (0/1, true/false, yes/no). Logistic regression (LR) is used to describe the relationship between one dichotomous dependent attribute (Y) and one or more nominal independent variables (X) (Fig. 2). The dependent variable constitutes the target class or the labelled class we want to predict and the independent variable are the parameters or the variables, which

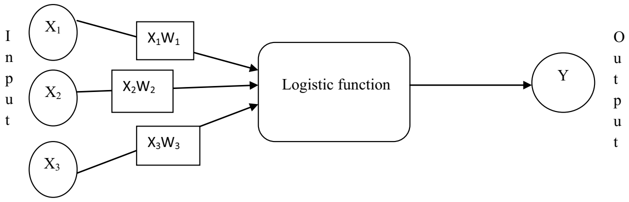

Fig. 2 Representation of logistic regression model

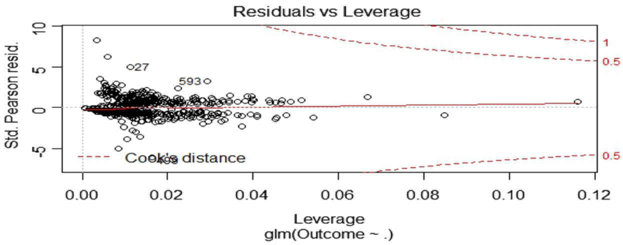

Fig. 3 Correlation between various attributes in Logistic Regression model

we use to do prediction. Mathematically LR classifier will use the logistic function to do the prediction.

In order to do the prediction LR needs activities score, weight, and the expected target. The activities constitute the independent attribute and the activity score are the numerical value of the independent attribute. The weights corresponds to the particular target which means if the weight is 0.5 we can say the model is 50% confident about the target class whether be diabetic or non diabetic. Since this being binary classification, if the weights for the target class diabetic are positive then for the non-diabetic they are negative. To do the prediction we need to know the activity score for each activity and the corresponding weight (Eq. 2), which provide us the logit.

$$Logit\ =\ Activity\ score\ *\ Weight \tag{2}$$

Logit is also called as score. It is the multiplication of activity score and the corresponding weight. The logit is passed into the logistic function in order to get the probability for each target class and the class with high probability is considered to be predicted as the target class for the given attribute. Thus we predict whether the person is diabetic or not with the calculated logits. The simple logistic regression has the form

$$\text{logit(pr[Y])} = \ln\left(\frac{p(\pi)}{1-p(\pi)}\right) = \alpha + \beta X \tag{3}$$

The logit is natural logarithm (ln) of the ratio of probabilities (π) of (Y) happening (i.e. patient is diabetic) to the probabilities ($1-\pi$) of (Y) not happening (i.e. patient is non-diabetic) (Eq. 3).

To check the credibility of the logistic regression residuals are important and are used to find difference between the observed value and the predicted value (Eq. 4).

$$\text{Residual}\ =\ \text{Observed} - \text{Predicted} \tag{4}$$

Figure 3 shows the correlation between the observed and predicted values in the diabetic dataset for training Logistic Regression model better. Least the residual better is the model for prediction.

3.4.3 Naive Bayes

Naive Bayes classifier is a simple approach used for the classification task [28]. It is considered as one of the powerful algorithm, which uses clear semantics and representation. It is called naive because it takes into consideration the two assumptions. The first is that the predictive attributes are conditionally independent and the second one is that there are no hidden or latent attributes. Naive Bayes classifier is best to use for textual data analysis and its working is based on Bayes theorem. Bayes theorem works on conditional probability, which means predicting something, based on certain criteria, which has already taken place. In clear words, conditional probability shows something will happen given that something else has already happened [29, 30]. If A represents the dependent attribute and B represents the prior event or the prior happened independent attribute then Bayes theorem is given by Eq. (5)

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{5}$$

Given that 'B' has already happened, what is the probability that 'A' will happen is equal to the product of the probability of two probabilities divided by probability of the already happened event. Naive Bayes classifier predicts membership probabilities for each class. The target class with highest probability is depicted as the most likely class. In the real datasets, we come across multiple attributes associated with the problem like BMI, age, plasma glucose concentration, blood pressure etc. Since considering all attributes makes the calculations complicated, thus feature independence approach is used to treat each attribute as an independent.

$$P(A|Multiple\ attributes) = \frac{P(B1|A) * P(B2|A) * \ldots P(Bn|A) * P(A)}{P(Multiple\ attributes)} \tag{6}$$

Using Eq. (6), we take pregnancy, BMI, triceps skin fold thickness, pedigree, plasma glucose, age, blood pressure and serum insulin as multiple necessary independent attributes for predicting a person is as diabetic is shown as below

$$\mathrm{P(D|Pregnancy,\ BMI,\ Triceps,\ pedigree,\ plasma\ glucose,\ age,\ BP,\ Insulin)} = \frac{P(pregnancy|D) * P(BMI|D) * P(Triceps|D) * P(pedigree|D) * P(plasma\ glucose|D) * P(age|D) * P(BP|D) * P(Insulin|D) * P(D)}{P(pregnancy,\ BMI,\ triceps,\ pedigree,\ plasma\ glucose,\ age,\ BP,\ Insulin)}$$

where D represents diabetic—patient who is diagnosed as positive and to know the probability of a person is non diabetic is shown as follows

$$\mathrm{P\left(ND|Pregnancy,\ BMI,\ Triceps,\ pedigree,\ plasma\ glucose,\ age,\ BP,\ Insulin\right)} = \frac{P(pregnancy|ND) * P(BMI|ND) * P(Triceps|ND) * P(pedigree|ND) * P(plasma\ glucose|ND) * P(age|ND) * P(BP|ND) * P(Insulin|ND) * P(D)}{P(pregnancy,\ BMI,\ triceps,\ pedigree,\ plasma\ glucose,\ age,\ BP,\ Insulin)}$$

where ND represents non-diabetic—patience who is diagnosed as negative.

If P(ND|pregnancy, BMI, triceps, pedigree, plasma glucose, age, BP, insulin) > P(ND|pregnancy, BMI, triceps, pedigree, plasma glucose, age, BP, insulin) then we say patient is diagnosed as diabetic and if P(ND| pregnancy, BMI, triceps, pedigree, plasma glucose, age, BP, insulin) > P(D|pregnancy, BMI, triceps, pedigree, plasma glucose, age, BP, insulin) then we say patient is non diabetic.

4 Results and discussions

In order to measure and validate the performance of the classifiers some metrics are taken into consideration in terms of accuracy, specificity, sensitivity, error rate. These metrics usually provide quantitative results. In addition, ROC (receiver operating characteristics) used shows the trade-off between sensitivity and specificity. The graph depicts that how much accurate the test is if curve of ROC follows the left hand border and then the top border of the ROC space. If it comes to 45 degree of the ROC space, it means that the test performed is less accurate. ROC is important to visualize the over prediction and under prediction. Area under curve (AUC) is an additional means to assess the classification performance. AUC is ROC related metric. The four possible states i.e. true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are required to assess the performance. The three metrics are calculated as follows:

(1) Accuracy (Acc)

It gives the overall performance of the classifier. It shows the ability of a classifier to correctly predict given samples and given by

$$\mathrm{Acc} = \frac{Tp + TN}{TP + TN + FP + FN} \tag{7}$$

(2) Sensitivity (Sn)

This metrics tells us about how well the classifier is doing to identify the positive results and is given by

$$\mathrm{Sn} = \frac{Tp}{TP + FN} \tag{8}$$

(3) Specificity (Sp)

On contrary to sensitivity specificity tells us how well the classifier is doing to identify the negative results and is given by

$$\mathrm{Sp} = \frac{TN}{TN + FP} \tag{9}$$

The data set is separated into two parts for the purpose of testing and training. The 70% portion of data is used for training purpose and 30% for testing the accuracy of trained model.

4.1 Experimental results

The records present in the Pima Indians diabetes data set is augmented with a diabetes status positive/negative based on the medical test results and this status is the class label for each instance provided in the data set. We have used holdout method to evaluate the performance of the classifiers. Gradient Boosting experiments are carried out using threefold cross validation and number of iterations = 3 in order to obtain sustainable and reliable results. To assess the overall or discriminative capability of

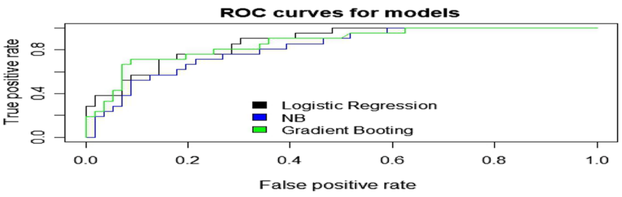

Fig. 4 ROC curve for the prediction classifiers for diagnosis of diabetes disease

Table 3 Analysis of various methods

S. no.	Methods	Testing accuracy (%)	Sensitivity (%)	Specificity (%)
1	Gradient boosting	86	89.5	71.4
2	Logistic regression	79.2	77.8	100
3	Naive Bayes	77	80.4	66.6

binary classifiers in for the better diagnosis or prognosis of diabetes, receiver-operating curve is used as a tool.

ROC is a plot between true positive rate and false positive rate (1—specificity) which shows quality or performance of the diagnosis of diabetic disease. ROC curve theoretically takes the values between 0 and 1 and it is found the classifier which takes value of 1 is an ideal classifier. However, the practical lower bound for random classifier is 0.5 which means classifier has no discriminative capability whereas classifier with roc significantly more than 0.5 has at least some discriminative capability. Figure 4 depicts the experimental results of the study using gradient boosting, logistic regression and Naive Bayes. Among all the three classifiers, the gradient boosting shows high reliability of discriminative capability among all the methods.

Gradient Boosting has accuracy of the testing data as 86% and Naive Bayes as 77% and Logistic Regression as 79%. Thus from Table 3 we can conclude Gradient Boosting is better than the Naive Bayes and Logistic Regression in the process of predicting whether a person is diabetic or non diabetic.

5 Conclusion

Traditionally, Doctors have evaluated whether the person is diabetic with the help of some diagnostic test. First, they have checked the serum and plasma glucose rate per hour. Diagnosis of diabetic person has historically included fasting blood glucose higher than prescribed rate. Another factor like Body mass index has also played very important role during diagnosis of a diabetic pregnant woman, compared with women with a pre pregnancy BMI < 29 kg/m^2, women with a BMI > 29 kg/m^2 has an 10-fold increased risk of developing type 2 diabetes. Consequently, both of these factors such as BMI and Plasma glucose have also significantly co-related attributes during our study. Boruta algorithm has been used for this purpose. Gradient Boosting is one of the best machine learning techniques for the regression and classification problems. The results of the implementation show that gradient boosting has prediction accuracy of 86%, which is greater than the other two techniques used. The data set used is Pima Indians diabetes dataset. In any research, data pre-processing is an important step in order to build the better and reliable model for the process of prediction. In future, similar approaches can be applied on other disease datasets like cardiovascular disease or oncology based diseases for the purpose of prediction. Moreover, same techniques can be used for pathological and rare disease prediction in order to enhance the overall healthcare.

Acknowledgements This research work was catalyzed and supported by the National Council for Science and Technology Communications (NCSTC), Department of Science and Technology (DST), Ministry of Science and Technology (Govt. of India), New Delhi, India [Grant Recipient: Dr. Harleen Kaur and Grant No. 5753/IFD/2015-16]. The authors gratefully acknowledge financial support from the Ministry of Science and Technology (Govt. of India), New Delhi India.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Manos A, Sattler M, Alukal G (2006) Make healthcare lean. Qual Prog 39(7):24
2. Yancy P, Handley A (2004) Can everyone be refreshed? J Contin Educ Nurs 35(2):80–83
3. Kaur H, Alam MA, Jameel R, Mourya AK, Chang V (2018) A proposed solution and future direction for blockchain-based heterogeneous medicare data in cloud environment. J Med Syst 42(8)
4. Yue X, Wang H, Jin D, Li M, Jiang W (2016) Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. J Med Syst 40(10):218
5. Temurtas H, Yumusak N, Temurtas F (2009) A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl 36(4):8610–8615
6. Veena V, Ravikumar A (2014) Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. Int J Comput Appl 95(17):12–16
7. Patil BM, Joshi RC, Toshniwal D (2010) Association rule for classification of type-2 diabetic patients. In: 2010 2nd International conference on machine learning and computing association. IEEE Computer Society, pp 330–334

8. Georga E et al (2009) Data mining for blood glucose prediction and knowledge discovery in diabetic patients: the METABO diabetes modeling and management system. In: 31st Annual international conference. IEEE EMBS Minneapolis, Minnesota, USA, vol 43100, pp 5633–5636
9. Breault JL, Goodall CR, Fos PJ (2002) Data mining a diabetic data warehouse. Artif Intell Med 26:37–54
10. Vijayan V, Ravi K (2015) Prediction and diagnosis of diabetes mellitus—a machine learning approach, December, pp 122–127
11. Polat K, Güneş S, Arslan A (2008) A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Syst Appl 34(1):482–487
12. Koh HC, Tan G (2005) Data mining applications in healthcare. J Healthc Inf Manag 19(2):64–72
13. WHO (2017) Diabetes. WHO, Geneva
14. Endocrineweb (2016) Normal regulation of blood glucose. https://www.endocrineweb.com/conditions/diabetes/normal-regulation-blood-glucose. Accessed 03 Feb 2016
15. Davidson M, Peters M, Ruchi A, Harmel AP (2004) Diabetes mellitus, 5th edn. Saunders, Philadelphia
16. UCI (2017) Pima indians diabetes data set. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. Accessed 20 Dec 2017
17. Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 03 Jan 2010
18. Kotsiantis SB, Kanellopoulos D (2006) Data preprocessing for supervised leaning. Int J Comput Appl 1(2):1–7
19. Devore J (2007) Making sense of data: a practical guide to exploratory data analysis and data mining. Glenn J Myatt 67:370–371
20. Jayalakshmi T, Santhakumaran A (2010) A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. In: 2010 International conference on data storage and data engineering. IEEE Computer Society, pp 159–163
21. Pal CJ, Witten IH, Frank E, Hall MA (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
22. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21
23. Kaur H, Wasan SK (2006) Empirical study on applications of data mining techniques in healthcare. J Computer sci 2(2):194-200
24. Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Trans Res Part C: Emerg Technol 58:308–324
25. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Procedia Procedia Comput Sci 82:115–121
26. Hayes AF, Matthes J (2009) Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. Behav Res Methods 41(3):924–936
27. Baboota R, Kaur H (2019) Predictive analysis and modelling football results using machine learning approach for English Premier League. Int J Forecast 35(2):741-755
28. McCallum A, Nigam K (1998) A comparison of event models for Naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol 752, no. 1, pp 41–48
29. Zhang ML, Peña JM, Robles V (2009) Feature selection for multi-label Naive Bayes classification. Inf Sci 179(19):3218–3229
30. Frank E, Trigg L, Holmes G, Witten IH (2000) Naive Bayes for regression. Mach Learn 41(1):5–25