Early Stage Prediction of Type Two Diabetes in Females

Amal S Menon
MTech. Student, Dept. of Computer Science and Engineering, BMS College of Engineering,
Bengaluru, India

Gowrishankar S
Professor, Dept. of Computer Science and Engineering,
BMS College of Engineering,
Bengaluru, India

***Abstract*— The advanced improvement in health science and technology has generated huge amount of data in health records and clinical format**

ion. Application of machine learning and data mining techniques are used to transform the medical field into more intelligent and accurate.

Diabetes mellitus (DM) is a group of metabolic disorder exerting pressure on human health around the world. Diabetic leads to heart problems, kidney diseases, nerves damages, eye sight problems and also artery and vein damages. Existing set of machine learning algorithms are introduced to develop different types of models for prediction of diabetic conditions in patients and civilians. For better model of greater accuracy, this study introduces an enhancement of KNN algorithm, to predict the diabetic conditions.

Keywords*— *Clinical information; diabetes mellitus; metabolic disorder; machine learning algorithms; KNN algorithm.

I.INTRODUCTION

Diabetes mellitus is a chronic disease which occurs due to the lack of insulin in the human body. Pancreas is the insulin making organ in the human anatomy. When it is unable to continue its production of insulin, it leads to diabetes mellitus in human body. Insulin is a hormone which helps to convert glucose from the food we eat into energy. All carbohydrates are broken down into glucose, which in turn aids the glucose to get into cells. Several health problems are created along with the development of diabetes.

Mainly there are three kinds of diabetes- type 1, type 2 and gestational [1].

A. Type 1 Diabetes

These are mostly found in children and teenagers. In this case pancreas produces very little or no insulin, and the patient has to intake insulin as injection to maintain the blood glucose level. Although this is mostly found in children this can be developed at any age at any time.

B. Type 2 Diabetes

These are common in adults. Normally in human body after a certain age pancreas shows laziness to create insulin [11]. As a result, the patient requires oral drugs and insulin, to keep the glucose level under control.

C. Gestational Diabetes

This is a special type of diabetes which leads to high blood glucose during the pregnancy period in females. This will result in complications for both mother and the child. It will mostly disappear after the delivery but high probability exists for mother or child to be prone to develop type 2 diabetes in the future.

The main symptoms of diabetics are weight loss, increased thirst and hunger, increase in urination, wounds which do not heal, blurred vision and tiredness [2]. Type 1 diabetes symptoms may start early but type 2 diabetes symptoms develop slowly over years. Many people are unable to recognize the matter until they fall in the hands of major health traps such as blindness, heart failure etc.

Treatment of Type 1 diabetes is insulin injection or pumping insulin to the body. Thus, by taking insulin externally helps to stabilize the glucose in body cells. But the count of insulin must be precise and accurate according to the need of the body. This is based on food, general health, stress, emotions, sleeping habits. These factors relate to burn of extra glucose. If the patient consumes insulin above the need, it leads to dropping of blood sugar to dangerously low level and it can be life threatening. This is hypoglycemia. And if the intake of insulin is too little, the blood sugar can raise to dangerously high level. This is hyperglycemia and can lead to long term complications which can be life threatening.

Type 2 diabetes is mostly found in people, that is 90% of diabetic patients are in the category of type 2. Since it is developed after the age of 35, it is also called adult onset diabetes. People with type 2 can produce certain amount of insulin in their body, but that may not be sufficient to open the cells and to allow the glucose to enter. This is called insulin resistance. So additional insulin must be given from outside to solve this issue. Most of the symptoms of type 2 are similar as type 1. But in type 2 case the symptoms are usually slower, so the people are not aware about the diabetic disease until it progresses into a late stage.

Treatment of diabetes focuses on diet, exercise and intake of insulin as oral or injection. Diagnosis and prediction in diabetes mellitus is mainly detected by the following steps:

- A1C test: alpha-glycosylated hemoglobin test
- Random blood sugar test
- Fasting blood sugar test

Metabolomics, proteomics and genomics are the common biomarkers. Researches in the study of used membrane fluidity of type 1 diabetes and the decision system have directed to monitor the pathology of diabetes mellitus [3]. Applying machine learning techniques and data analysis to diabetic medical record have improved the traditional recognition of diabetes. The researchers have proposed a solution of hybrid machine learning to solve the problem of uneven of medical data distribution.

As of the survey of United States in 2015, 30.3 million people have diabetes, i.e., 9.4% of total population. And half of them were unaware that they had this disease [4]. To avoid this incident of unawareness, a program to confirm and predict whether one has the disease or the chance of being a patient in the coming years, a software to check this by themselves has to be introduced. The feature of these software should consider heredity of the tester. The scope of this software will be very high in the future because the living style and food habit of people is leading towards diabetes.

The study focuses on diabetes in female, and the software is developed for mainly checking for women who had undergone pregnancy, since they have high chance of falling under the grasp of diabetes due to gestational diabetes during pregnancies. This software gives them a future indication for the DM, so that they can take necessary prevention. This software is also applicable for ladies who had never undergone pregnancies

II.MACHINE LEARNING ALGORITHM

Machine learning algorithm connects the problems from the data samples which collected from general concepts of reasoning. Machine learning can be divided into 2 process [5]:

- Discovering data dependencies from a given dataset.
- From established relations to create outputs for new inputs.

Machine learning can be classified into: -

A. Supervised Learning

In which the system creates a function from a labelled training data. The target function is used to predict the value of a variable (dependent variable or output variable) from a set of features (independent variable or input variable) [6]. Set of input values of the function are called instants is its domain. Each case is explained by a set of characteristics or attributes or features. Training data is a subset of all features for which the target variable is known. To generate the best target function, the learning system, a training set, an alternative function called hypothesis is taken into consideration. Classification and regression are 2 kinds of learning tasks for the supervised learning [8]. Models which predict distinct classes like blood group is classification and the regression predicts the numerical value. Most common technique in supervised learning are decision tree (DT), K-Nearest neighbors, rule learning, genetic algorithms, support of vector machine (SVM), artificial neural network (ANN).

B. Unsupervised Learning

In which the learning system try to generate a structure from unlabeled data. i.e., the system tries to discover the hidden structure of data or relationship between variables.

C. Reinforcement Learning

Is to maximize the cumulative reward in which the system tries to learn through direct interaction with the environment. In this, the system has no prior knowledge about the behavior of the environment and the only way to solve this is through trial and error method. The main application of this is autonomous systems.

The ML algorithms that are considered for creating models to predict diabetes are usually, k-NN algorithm [10], Naïve Bayes algorithm, Support vector machine algorithm and random forest algorithm.

K- Nearest Neighbor (k-NN) yields very good results, but it is very simple compared to other. Figure 1, shows K-NN algorithm. It is a non-parametric, lazy and instance-based learning algorithm. In classification and regression problem this can be used. K-NN is applied to find out the class for classification of new unlabeled objects. K stands for number of neighbors which is generally odd. The nearest object to the data point is calculated using Euclidian's distance, Manhattan distance, Makowski distance and Hamming distance. According to distance, nearest neighbors k is selected which is used to determine the nearest class of the new object. Generally, this algorithm has high accuracy.

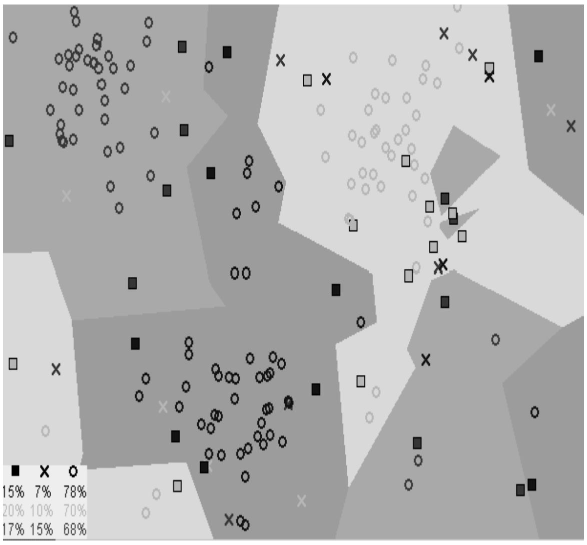


Fig. 1. K-NN Algorithm

Logistic Regression is another technique by machine learning from the field of statistics. It works out through binary classification. Logistic function is an S-shaped curve with a real valued number and mapped into a value between 0 and 1, but excluding 0 and 1. The equation is: **1/(1+e^-value)** where **e** is the base of natural logarithm. And **value** is the value actual numerical value. Logistic regression is a linear method, predictions are made using logistic function. Figure 2, shows Logistic Regression Example.

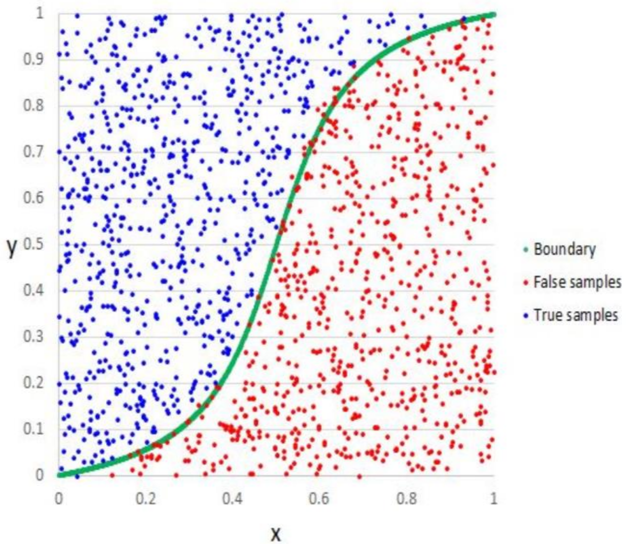


Fig. 2. Logistic Regression Example

Naïve Bayes algorithm reduces the complexity of conditional independence over training dataset [14]. The idea of conditional independence defines that the variables X, Y and Z, while X is conditionally independent of Y given Z. If the probability distribution governing X is independent of y given Z. X and Y are conditionally independent only if given Z occurs but X provides no information of Y occurring and Y provide no information regarding the occurrence of X. This makes the Bayes algorithm, naïve.

Given n different attributes values,

$$P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \quad \textbf{(1)}$$

Support vector machine (SVM) is a supervised learning algorithm and can be used for classification and regression challenges, but mostly used in classification. In SVM each data item will be plotted as a point in n-dimensional space (n is number of features) with the value of the feature belonging to the value of particular co-ordinate [13]. Then the classification is performed by finding the hyper-plane which differentiates into classes. Support vectors are the co-ordinates of individual observations. SVM classifier segregates the 2 classes in the best way. Figure 3, shows SVM algorithm.

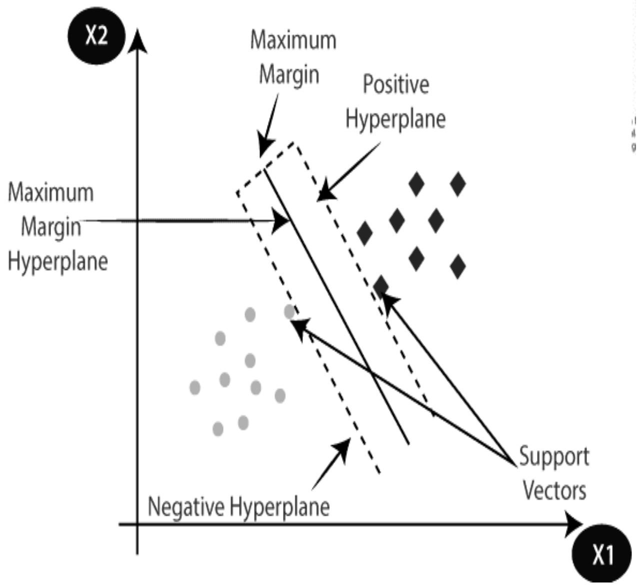


Fig. 3. Algorithm of SVM

Decision tree algorithm led to supervised learning. Here the leaf nodes of the tree indicate to respective class label and the internal nodes in this tree are the representations of the attributes. Regression problems can also be solved using this algorithm. The problem in decision tree algorithm is the root node identification. This process is called attribute selection. Information gain and Gini index are the popular attribute selection methods. The entropy changes if we use nodes to segregate the tree. Figure 4, shows decision tree algorithm.

Here the term entropy indicates the amount of random variable uncertainty.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v) \quad \textbf{(2)}$$

Here Gini index is a term to calculate the probability of incorrect identification of any given variables that is taken very randomly.

$$GiniIndex = 1 - \sum_j p_j^2 \quad \textbf{(3)}$$

where **P** is the proportion.

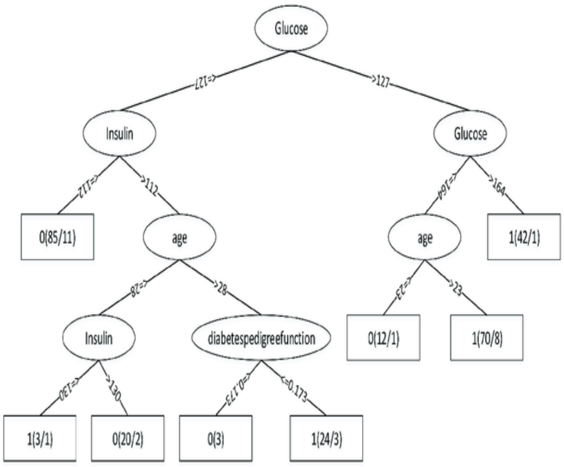


Fig. 4. Decision tree Algorithm

Random decision forest is a cluster of methods which consist of similarity [12]. Random forest algorithm makes decision tree on data samples. Then finally select the best solution by means of voting. Because of it reduces the over fitting by averaging the result, it is better than a single decision tree. Figure 5, shows Random decision forest.

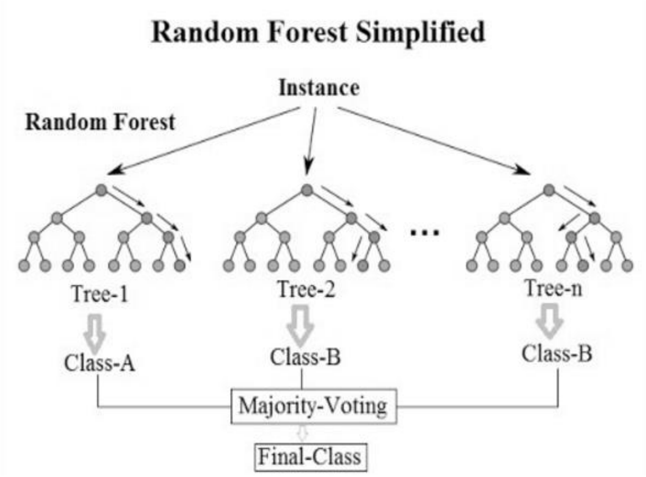


Fig. 5. Random decision forest

Working of random forest algorithm, first start with the selection of random samples. Then construct a decision tree for every sample, it will predict the result from every decision tree. Voting will be performed for every predicted result. The final result is the prediction that are most voted. Figure 6, shows Working of Random Decision Forest.

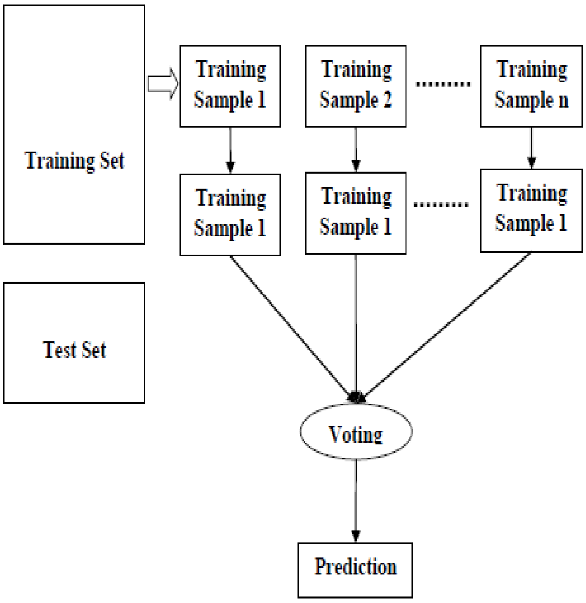


Fig. 6. Working of Random Decision Forest

III.FEATURE SELECTION

It is the process that is very imperative for data transformation in KDD. The process involves selection of dataset for future space and is very important for the model creation. It relates to different aspect of data analysis, better visualization and understanding of data, reducing computation time and duration of analysis, and better prediction accuracy are the advantages of feature selection. Feature selection process has two main approaches.

- To make an independent assessment based on general characteristics of data. This is called filter method because the feature method is filtered out before model construction.

- To use a machine learning algorithm to examine different subset of features and select one with best performance or accuracy. A model would be build using this algorithm. Algorithm wraps the whole feature selection process, so this method is called wrapper method.

IV.METHODS

A. *Center profiling* [9]

Targeting the hospital attributes, which include type of pattern of care and also population. This study was helpful for obtaining bias in the selection and more responsively defining the prediction difficulty to ML variables is to be used as feature for this initial analysis.

B. *Predictive model*

1) targeting: it is viable and feasible to aim different techniques of modelling for particular subset on the basis of literature review.

2) construction: to define a strategic for preprocessing problem if the target model has been selected.

3) validation: this model is specifically used to approach the working of proposed method.

Predictive model training is targeted on center profiling. This method also focuses on literature review. The study focuses at predictive models.

A diabetic study was conducted through survey in UK [7]. The results show there is no possibility to apply all the attributes to ICSM dataset. Instead of collecting data from diabetic diagnostic, ICSM data gives details following the patient's call on to the hospital. Also, in terms of predication accuracy these analysis does not provide a valid model.

Dataset of female patient above age of 20 have been collected from National Institute of Diabetes and Digestive and Kidney disease (NIDDK) via UCI machine learning repository. Table 1, shows the statistical report of dataset There are total 2780 instances classified into 2 classes – diabetic and non-diabetic.

Table 1. Statistical report of dataset

Attribute No.	Attribute	Variable Type Range	age
A1	Pregnancy (No of times pregnant)	Integer 0-17	0–17
A2	Plasma Glucose (mg/dL)	Real 0-199	0–199
A3	Diastolic Blood Pressure (mm Hg)	Real 0-122	0–122
A4	Triceps skin fold (mm)	Real 0-99	0–99
A5	Serum Insulin (mu U/ml)	Real 0-846	0–846
A6	Body mass index (kg/m^2)	Real 0-67.1	0–67.1
A7	Diabetes Pedigree	Real 0.078-2.42	0.078–2.42
A8	Age (years)	Integer 21-81	21–81

V.IMPLEMENTATION OF ML ALGORITHM

In order to create more accuracy some data pre-processing techniques such as co-relation, cleaning of empty or null values by replacing mean. As part of the challenge, the dataset given contains a lot of missing values. Ignoring the missing values from the samples would leave only with a very few training samples. Also, the research samples contain missing values and so we need to find ways to impute the missing values. Knowledge acquisition and reasoning architecture for predicting diabetes.

We must also predict the missing values correctly, otherwise the algorithm can take the wrong one. From the above discussed classification algorithms, we have created models using learn and attained accuracy as table below.

Table 2, shows. Comparison of different Machine Learning Algorithms. By comparing the above-mentioned algorithms, arrived at the conclusion that random forest algorithm is the most effective [16]. RF is mixture of decision trees and does not have the adverse effect of over-fitting, that is the reason behind the best performance of RF compared to other algorithms. Thus, it is decided to start with RF algorithm with modifications for getting better results. Before comparing the result provided by RF, the factors that decides it's accuracy, and the factors helps to avoid over-fitting, have to be examined. After the realization of these factors, we have reached the point RF is a group of tree structured classifier,

$$h(x,\Theta k)n\ k=1 \qquad (4)$$

where the {Θk} are identical, independent distributed random vectors.

Table 2. Comparison of different Machine Learning Algorithms

S. No.	Algorithms	Accuracy
1	KNN	77.38%
2	Naïve Bayes	73.77%
3	Logistic Regression	76.65%
4	SVM	77.13%
5	Decision Tree	97.23%
6	Random Forest	98.19%

Overfitting happens when we don't build a dataset based generic model. This increases the probability of getting errors. The generalization error can be used to measure the over-fitting. In RF the generalization error has a limiting value, because RF has less chance of fitting the data. Generalization error can be defined as:

$$PE = PX,Y\ (mg(X,Y) < 0) \qquad (5)$$

Here the value of marginal function is less than zero. In case of a random forest the margin function is:

$$mr(X,Y)=P\Theta(h(X,\Theta)=Y)-max\ j6=Y[P\Theta\ (h(X,\Theta)=j)] \qquad (6)$$

Reduction of correlation: Correlation can be considered as the measure of 2 trees in RF, this is defined as:

$$p(\Theta,\ '\Theta) = covX,Y(rmg(\Theta,X,Y),rmg('\Theta,X,Y))\ /\ sd(\Theta)sd('\Theta) \qquad (7)$$

p(Θ, ´Θ) : *rmg(´Θ,X,Y*

where co-relation, raw margin function, and ***sd(Θ)*** denotes the standard deviation function. The raw margin function is:

$$rmg(\Theta,X,Y) = I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{}j(X,Y)) \qquad \textbf{(8)}$$

The equation of correlation shows that we to reduce the correlation, the variance between the trees must be reduced.

VI.SELECTION OF ALGORITHM

Gradient Boosting is one of the important learning algorithms which makes a strong classifier combined with weak learners. Freidman constructed the building addictive models in a step wise manner and at each stage a weak learner is built by gradient. E.g.: A weighted sum of neural network using different convex loss functions. In GB algorithm is introduced by a statistical interruption of boosting. For example, minimizing the criterion,

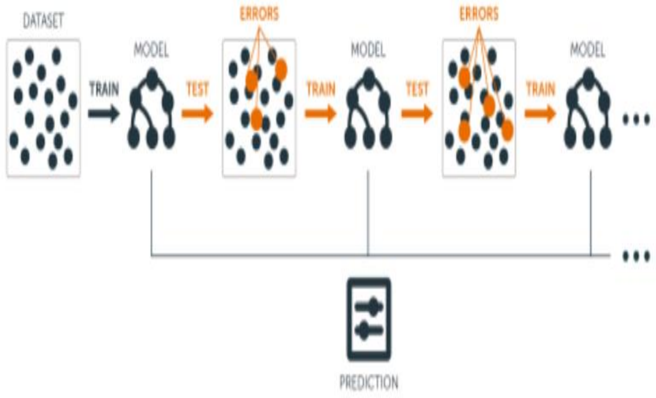
E[exp(-y F(x))] for decision function F(x). It is sufficient to minimize the conditional criterion on x, if E involves over the joint distribution of y and x.

$$E[\exp(-yF(\mathbf{x})) \mid \mathbf{x}] = p(y=1 \mid \mathbf{x})\exp(-F(\mathbf{x})) + p(y=-1 \mid \mathbf{x})\exp(F(\mathbf{x}))$$

$$\frac{\partial E[\exp(-yF(\mathbf{x})) \mid \mathbf{x})]}{\partial F(\mathbf{x})} = -p(y=1 \mid \mathbf{x})\exp(-F(\mathbf{x})) + p(y=-1 \mid \mathbf{x})\exp(F(\mathbf{x})) \qquad \textbf{(9)}$$

F(x) that reducing the condition E[exp(-y F(x))] is taken by activating the derivation to zero and is the symmetric logistic transform of p(y=1|x)

$$F(\mathbf{x}) = \frac{1}{2}\log\frac{p(y=1 \mid \mathbf{x})}{p(y=-1 \mid \mathbf{x})} \qquad \textbf{(10)}$$



Boosting can be explained as a step wise procedure for fitting addictive logistic regression models. Loss function that is

L=log(1+ exp(-2y F(x)))

is same as the Logit boost. Gradient nr is the negative derivation of the loss function by the decision function. From the algorithm shown below, it is concluded that GB is a greedy function.

Given: D: Training set, L: the learner, F: the decision function, I: iterations, S: total samples, v: regularization term.

```
1.    START
2.    Initialize F=0
3.    Initialize i =0
4.    Loop until i=I
5.    Initialize s=0
6.    Loop until s=S
7.    Compute:
          $r = \frac{2y_s}{1+\exp(2y_s F_{i-1}(x_s))}$
8.    Increment s by 1
9.    End loop
10.   Fit a weak learner, Li for train dataset
          $F_s(L_1(x),\cdots,L_s(x)) = F_{s-1} + \nu \cdot L_s(x)$
11.   Increment i by 1
12.   End loop
```

Measures Of Performance: Generally for evaluation of methods four measures are used. Prediction accuracy shows the number of correct samples. In imbalanced data, a simple model predicts all the samples as the negative class can increase the accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad \textbf{(11)}$$

F-measure makes Precision TP/TP+FP and Recall the TP/TP+FN for the positive class. If FN and FP are positive the it shows predictive model works well.

$$F-measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad \textbf{(12)}$$

Here G-Mean is calculated as the product of the prediction of classes. If the predictive model identifies negative samples, poor performance in predicting positive samples will make low G-mean value.

$$G-mean=\sqrt{\frac{TN}{TN+FP}\cdot\frac{TP}{TP+FN}} \tag{13}$$

ROC curve is a very important measure to evaluate the algorithm performance. It represents trade-off between the true positive and false positive. Upper curve shows the better performance.

Area below ROC curve is called AUC.

VII. ENHANCEMENT

We propose a new algorithm by enhancing GBM called Levelboost, by assigning costs based on the empirical distribution of class. We denote the minority class distributic Pˆ (y = 1) and the empirical majority Pˆ (y = -1) class as _ . For estimation of decision function F(x) the minimizing criterion is E[exp(−y F(x))]

But E need expectation of joint distribution of x and y, it is sufficient to minimize the conditional criterion on x. Bayes' theorem is then utilized to assign the weights:

$$\frac{p(y=1)}{\hat{p}(y=1)} \quad \text{and} \quad \frac{p(y=-1)}{\hat{p}(y=-1)}$$

by substituting the empirical class distributions with the equal class distributions on E[exp(−yF(x))|x]

$$\begin{aligned}E[\exp(-yF(\mathbf{x}))\,|\,\mathbf{x}] &= \hat{p}(y=1\,|\,\mathbf{x})\exp(-F(\mathbf{x}))+\hat{p}(y=-1\,|\,\mathbf{x})\exp(F(\mathbf{x}))\\ &=\frac{\hat{p}(y=1)p(\mathbf{x}\,|\,y=1)}{p(\mathbf{x})}\exp(-F(\mathbf{x}))+\frac{\hat{p}(y=-1)p(\mathbf{x}\,|\,y=-1)}{p(\mathbf{x})}\exp(F(\mathbf{x}))\\ &\to\frac{\hat{p}(y=1)p(\mathbf{x}\,|\,y=1)}{p(\mathbf{x})}\frac{p(y=1)}{\hat{p}(y=1)}\exp(-F(\mathbf{x}))+\frac{\hat{p}(y=-1)p(\mathbf{x}\,|\,y=-1)}{p(\mathbf{x})}\frac{p(y=-1)}{\hat{p}(y=-1)}\\ &\qquad)\end{aligned}$$

$$\text{where}\quad p(y=1)=p(y=-1)=\frac{1}{2}. \tag{14}$$

Then, to obtain the derivatives of F(x),

$$\frac{\partial E[\exp(-yF(\mathbf{x}))\,|\,\mathbf{x}]}{\partial F(\mathbf{x})}\to-\hat{p}(y=1\,|\,\mathbf{x})\frac{p(y=1)}{\hat{p}(y=1)}\exp(-F(\mathbf{x}))+\hat{p}(y=-1\,|\,\mathbf{x})\frac{p(y=-1)}{\hat{p}(y=-1)}\exp(F(\mathbf{x}))$$

(15)

F(x) that minimizes the above equation is obtained by using the logistic asymmetric transformation:

$$\hat{p}(y=1\,|\,\mathbf{x}).$$

$$F(\mathbf{x})=\frac{1}{2}\log\left(\frac{\hat{p}(y=1\,|\,\mathbf{x})}{\hat{p}(y=-1\,|\,\mathbf{x})}\cdot\frac{\hat{p}(y=-1)}{\hat{p}(y=1)}\right) \tag{16}$$

Hence, we can represent,

, p^ (y = 1| x) and p^(y = −1| x) by
p^(y=−1) , p^(y = 1) and) F(x).

$$\hat{p}(y=1\,|\,\mathbf{x})=\frac{\exp(F(\mathbf{x}))}{\frac{\hat{p}(y=-1)}{\hat{p}(y=1)}\exp(-F(\mathbf{x}))+\exp(F(\mathbf{x}))}$$

$$\hat{p}(y=-1\,|\,\mathbf{x})=\frac{\exp(-F(\mathbf{x}))}{\exp(-F(\mathbf{x}))+\frac{\hat{p}(y=1)}{\hat{p}(y=-1)}\exp(F(\mathbf{x}))} \tag{17}$$

Asymmetric weights denote that during the training of disease diagnosis model, an empirical class distribution is used. They become one and usual asymmetric costs, if these weights are equal. Then, we take the gradients of Levelboost, by the derivation by F(x).

$$r_{n|y=1}=-\frac{\partial L_{Bala|y=1}}{\partial F(\mathbf{x}_n)}=\frac{2}{1+\frac{\hat{p}(y_n=1)}{\hat{p}(y_n=-1)}\exp(2F(\mathbf{x}_n))}$$

$$r_{n|y=-1}=-\frac{\partial L_{Bala|y=-1}}{\partial F(\mathbf{x}_n)}=-\frac{2}{1+\frac{\hat{p}(y_n=-1)}{\hat{p}(y_n=1)}\exp(-2F(\mathbf{x}_n))} \tag{18}$$

LevelBoost Algorithm:

Given: D: Training set, L: the learner, F: the decision function, I:the number of iteration, S: total samples, v: regularization term.

```
START
Initialize F=0
Initialize i=0
Loop until i=I
    Initialize s=0
    Loop until s=S
        Compute:
            $\begin{cases} r_{n|y=1} = \dfrac{2}{1+\dfrac{\hat{p}(y_n=1)}{\hat{p}(y_n=-1)}\exp(2F_{s-1}(\mathbf{x}_n))} \\ r_{n|y=-1} = \dfrac{-2}{1+\dfrac{\hat{p}(y_n=-1)}{\hat{p}(y_n=1)}\exp(-2F_{s-1}(\mathbf{x}_n))} \end{cases}$
        Increment s by 1
    End loop
    Fit a weak learner, Li for train dataset
        $F_s(L_1(x),\cdots,L_s(x)) = F_{s-1} + v \cdot L_s(x)$
    Increment i by 1
End loop
STOP
```

The output gained is:

$$h(\mathbf{x}) = sign\left(\sum_{i=1}^{I} vF_i(\mathbf{x})\right) \quad \textbf{(19)}$$

VIII.CONCLUSION

The enhanced model provides a better result in prediction in accuracy 99.35%. This model enhanced from GBM by considering its weight and cost of samples. This model is framed by avoiding the drawbacks present in algorithms like, KNN, Naïve Bayes, Logistic regression, SVM, Decision Tree, Random forest. This model has high performance value in general since it is created in the platform h2o.

REFERENCES

[1] Al Jarullah, A.A., "Decision tree discovery for the diagnosis of type II diabetes," Innovations in Information Technology (IIT), 2011 International Conference on, vol., no., pp.303,307, 25-27 April 2011

[2] Folorunsho, Olaiya. "Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database." International Journal 3, no. 3 (2013).

[3] Huang, Feixiang; Wang, Shengyong; Chan, Chien Chung, "Predicting disease by using data mining based on healthcare information system," Granular Computing (GrC), 2012 IEEE International Conference on, vol., no., pp.191,194, 11-13 Aug. 2012

[4] Marcano-Cedeno, Alexis; Andina, Diego, "Data mining for the diagnosis of type 2 diabetes," World Automation Congress (WAC), 2012, vol., no., pp.1,6, 24-28 June 2012.

[5] Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol.6, no., pp.161,165, 810 April 2011

[6] Nincevic, I.; Cukusic, M.; Garaca, Z., "Mining demographic data with decision trees," MIPRO, 2010 Proceedings of the 33rd International Convention, vol., no., pp.1288,1293, 24-28 May 2010

[7] Robu, R.; Hora, C., "Medical data mining with extended WEKA," Intelligent Engineering Systems (INES), 2012 IEEE 6th International Conference on, vol., no., pp.347,350, 13-15 June 2012

[8] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, vol., no., pp.180,185, 27-29 Nov.,2012

[9] Arianna Dagliati, PhD1,2,3, Simone Marini, PhD1,2,3, Lucia Sacchi, PhD1,2, Giulia Cogni, MD3, Marsida Teliti, MD3, Valentina Tibollo, MS3, Pasquale De Cata, MD3, Luca Chiovato, PhD3, and Riccardo Bellazzi, PhD1 "Machine Learning Methods to Predict Diabetes Complications"

[10] Bradley AP," The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition". 1997;30(7):1145-1159.

[11] AMERICAN DIABETES ASSOCIATION: "Diagnosis and Classification of Diabetes Mellitus" journal.

[12] "Classification and Regression by randomForest" by Andy Liaw and Matthew Wiener

[13] Mahesh Pal ; Giles M. Foody., "Feature Selection for Classification of Hyperspectral Data by SVM"

[14] I. Rish, T.J. Watson Research Center, "An empirical study of the naive Bayes classifier"

[15] Y. LeCun et al., "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] P. Latinne, O. Debeir, and C. Decaestecker, "Limiting the number of trees in random forests," in Multiple Classifier Systems. Springer, 2001, pp. 178–187.