International Journal of Medical and Health Sciences

Journal Home Page: http://www.ijmhs.net ISSN:2277-4505

Original article

Diabetes Prediction Using Ensemble Classifier

Shawni Dutta[1] and Bandyopadhyay Kumar Samir[2*]

[1]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.

[2] Academic Advisor, The Bhawanipur Education Society College, Kolkata, India.

ABSTRACT

Diabetes is one of the impactful diseases that affect humans' health rigorously. Early diagnosis of diabetes will assist health care systems to decide and act according to counter measures. This paper focuses on obtaining an automated tool that will predict diabetic tendency of a patient. The system proposed by this paper contains two ensemble classifiers- Voting ensemble classifier and Stacking Ensemble classifier. Both of these methods exhibits better results while compared to other classifiers. Stacking ensemble classifier even performs better than voting ensemble classifier with an accuracy of 79.87%.

KEYWORDS: Diabetes, Automated tool, Prediction, Machine Learning, Ensemble Classifiers.

1. Introduction

Diabetes occurs when blood glucose or blood sugar of human is too high. It is the main source of energy. It originated from the food when human will take it. Insulin is a hormone made by the pancreas. It helps glucose from food into human cells for creation of energy. If insulin is not created well then Glucose stays within blood and doesn't reach respective cells. It creates health problem and till now it is not possible to cure diabetes without medicine/ injecting insulin for making a human healthy. Common types of diabetes are mostly type 1, type 2, and gestational diabetes. In type 1 diabetes, body produces very little or no insulin. It means that body needs daily insulin injections to maintain blood glucose levels under control. Type 2 diabetes, means body does not make good use of the insulin that it requires to produce. The treatment of type 2 diabetes is healthy lifestyle, including increased physical activity and healthy diet. However, over time most people with type 2 diabetes may require oral drugs and/or insulin to keep their blood glucose levels under control.

Daily generated electronic health care records can be utilised in early prediction of any disease. Timely detection and screening play leading role in prevention of disease like diabetes. This paper focuses on predicting diabetes beforehand so that counter measures can be suggested. Prediction of diabetes will assist health care systems to handle this disease carefully.

Data mining and knowledge discovery approaches are considered for obtaining an automated tool for diagnosing diabetes disease. This paper explores and applies Machine learning (ML) algorithms to differentiate between diabetic and non-diabetic patients. An automatic and adaptive technique is used by machine learning methods to achieve that classification. Given a set of messages, ML methods can obtain information and later can use the acquired information to classify unknown new messages.Supervised machine learning approaches are utilized in this field that takes patient's record as input and detect diabetes tendency beforehand. To address the problem of diabetes detection, classification techniques are implemented that maps input variable to target classes by considering training data.

The input variables include several parameters such as, number of pregnancies, patient's age, blood pressure, insulin taken, BMI percentage, Glucose level, Diabetes Pedigree function, Skin Thickness. All these data turn out to be good predictors while identifying diabetic patient. The predictive models can act as a tool to analyse the information of patients about their past health history records and their chances of being diabetic which in turn help the doctors to take informed decisions and prescribe medicines accordingly.

This paper works for improving the machine learning approaches in order to enhance the efficiency in diabetes prediction using medical data. In this paper, numerous classifier models such as Multi-layer Perceptron, Naïve Bayes Classifiers, Decision Tree Classifiers and K-nearest neighbors are employed to predict diabetic tendency of patients. These classifiers are capable enough to indicate promising results in terms of prediction. However, objective of this paper is to increase efficiency of prediction tool. Among the aforementioned classifiers, the best two models are selected for obtaining ensemble classifiers. For implementing these ensemble classifiers, two strategies such as Voting based and Stacking based are presented in this paper. Finally, this study has shown that Stacking ensemble classifier performs well over its peer classifiers.

2.Related Works

Mishra et. al. in [1], filter based feature selection methods were used based on Diabetes disease. Filter based feature selection methods include Chi-square method, Information gain method, Cluster Variation method and Correlation method and three classifiers such as RBF, IBK and JRip accompanied to estimate the performance of the algorithms. Filter based feature selection method has drawn special attention since it enhances the performance of the learning algorithms.

Gnana et. al. in [2], for diagnosing diabetes, several machine learning approaches such as probabilistic-based naïve Bayes (NB), function-based multilayer perceptron (MLP), and decision tree-based random forests (RF) are used. Different testing methods such as 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (UTD) are implemented to evaluate the performance of the machine learning model in terms of accuracy.

Hypoglycemia prediction models using machine learning approaches is developed in [3]. Several machine learning approaches such as random forest, SVM, k-nearest neighbor, and naïve Bayes are used for this purpose. Self-monitored blood glucose (SMBG) measurements with its sparse nature are used for training and testing purpose along with cross-validation method.

Zidian Xie et. al. in [4], have carried out prediction of diabetes disease considering six machine learning algorithms such as K-Nearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) and Random Forest (RF). Experimental results indicated that both SVM and KNN reach highest accuracy of 77% for predicting diabetes.

With Deep Neural Network (DNN) and Support Vector Machine (SVM) a system was developed in [5] to predict the diabetes. They have taken total 8 significant attributes of patients such as Age, Number of Times Pregnant, Plasma Glucose Concentration, Diastolic Blood Pressure, Body Mass Index etc. and got 77.86% accuracy.

An experiment on Diabetes diagnosis is conducted in [6] where Naive Bayes (NB), J48, SMO, MLP, and REP Tree algorithms are employed. Results have indicated that SMO gives 76.80% accurate results on diabetes dataset. Aljarullah et al. in [7] carried out a research on detecting type-2 diabetes. For this purpose, J48 algorithm is proposed and implementation of this algorithm is constructed using decision tree. The accuracy of the model is 78.1768%.

3.Proposed Methodology

The system flow diagram is shown in Figure.1. In subsequent sections different modules related to the proposed method are explained.

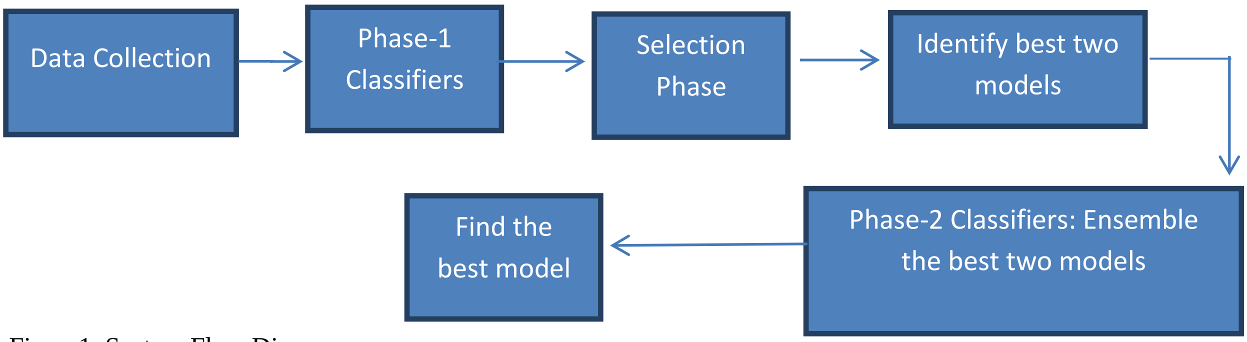

Figure1. System Flow Diagram

3.1 Dataset Collection

In this framework, Pima Indian women dataset from uci machine learning repository [8] is utilized for predicting diabetic tendency of a patient. The dataset can be formulated as collection of attributes that include several criterions for detecting diabetic tendency such as number of pregnancies, patient's age, blood pressure, insulin taken, BMI percentage, Glucose level, Diabetes Pedigree function, Skin Thickness, Outcome (Diabetic/Non-diabetic). However, the attribute 'outcome' is utilized as a target class of the prediction.

The dataset is partitioned into training set and testing dataset. 80% of the dataset is allocated for the training set whereas testing dataset need 20% of the dataset.

3.2. Classifiers Models

A classifier maps input variable to target classes by considering training data. The objective of using classifier is to predict whether a patient has diabetic tendency or not. The training set is fitted to classifier model and later prediction is obtained for test set.

3.2.1 Description of Phase-1 Classifiers

a. Naive Bayes Classifier

The Bayesian classification [9] acts as a fundamental probabilistic model that exploits the concept of Bayes Theorem [10] of Conditional Probability. By demonstrating the use of statistical method along with supervised technique this method obtains classification result. Prediction results are provided by this method by influencing the probabilities of the results. The result provided by this classifier is quite effective in practice even if its probability estimates are inaccurate. This classifier acquires a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy [9].

b. Multi-Layer Perceptron Classifier

Multi-layer perceptron [11] can be used as supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture [11].

c. K-nearest Neighbour Classifier

K-Nearest Neighbour Classifiers [12], often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k [12].

d. Decision Tree Classifier

A Decision Tree (DT) [13] is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each goal variable is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Classification results are obtained from this model by starting from the beginning at the root this tree are going through it until a leaf node is reached. This can be useful for forecasting the goal based on some criterion by implementing and training this model.

3.2.2 Implementation of the Phase-1 classifiers

In this framework classifiers are trained using appropriate parameters. For maximizing the performance of these models, default parameters may not be sufficient enough. Adjusting these parameters will produce enhanced predictive models which may be regarded as the optimised one while detecting recruitment possibilities.

This framework utilised MLP classifier as a collection of 5 hidden layers of size 128, 64, 32, 16 and 8 respectively. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. For naïve bayes classifier, multinomial naïve bayes classifier is employed. The decision tree classifier implemented in this paper uses Gini index while choosing objects from dataset. The nodes of the decision tree are expanded until all leaves are pure or until all leaves contain less than minimum number of samples. In this case, minimum number of samples is assigned a value as 2.

3.3 Selection Phase and Performance Measure Metrics

While evaluating performance skill of a model, it is necessary to employ some metrics to justify the evaluation. These metrics are used as selectors in the Selection phase. The purpose of selectors is to pick up top models based on their performances. During selection phase, the abovementioned classifier models are compared with respect to the following performance evaluation metrics. Use of these metrics will assist in identifying best problem solving approach.

Accuracy [14] is a metric that ascertains the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model’s performance since it does not consider wrong predicted cases. For compensating the above mentioned problem, we consider two more metrics known as, Recall[14] and Precision[14]. *Precision* [14] identifies the ratio of correct positive results over the number of positive results predicted by the classifier. *Recall* [14] denotes the number of correct positive results divided by the number of all relevant samples. *F1-Score* or *F-measure* [14] is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision and recall. *Cohen-Kappa Score* [15] is a statistical measure that has been considered as an evaluating metric in this paper. For classification problem this metric finds out inter-rate agreement for qualitative items. *Mean Squared Error* (MSE) [14] is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples. A better performing model is selected based on lower value of MSE and higher values of accuracy, F1-Score, and Cohen-kappa score.

This selection phase identifies top two models from phase-1 classifiers based on their performance. These classifiers are given as input for phase-2 classifiers.

3.4 Implementation of Phase-2 Classifiers

A group of dissimilar classifiers can be accommodated into a single platform to further improve the classification accuracy of the entire system. A couple of machine learning algorithms work on the identical problem to obtain enhanced classification accuracy. In phase-2 classifiers, two different ensemble strategies are employed. These are described as follows-

3.4.1 Voting Ensemble Classifier

Using voting strategy, it is potential to make a good choice out of multiple possible solutions. Hence, multiple classifiers may cast their preference for one or more solutions. Considering majority preferences, final decision is drawn for problem-solving approach. It is possible to obtain a better solution when several potential algorithms work towards the same problem domain. Using ensembles of different classifiers has the advantage that not all of them will make the same mistake [16].

3.4.2 Stacking Ensemble Classifier

Stacked generalization ensembles dissimilar machine learning methods, each of them estimates its own generalizing biases based on given training set and then filter out those biases. This model consists of two types of models- level 0 models and level 1 model. Several base models are accommodated in level 0 whereas level 1 model consists of one meta-model. Stacked generalization ensures that meta-model learns from the predictions of level-0 models. Generally introducing meta-model over the level-0 models obtains more accurate prediction over the best level-0 model [17].

The stacking ensemble classifier differs from voting ensemble strategy in the sense that the first one requires two levels to be present within the model. Use of level-1 model makes stacked generalization method more efficient.

In this methodology, selection phase provides two best models which are assembled in these ensemble classifiers. Using voting classifier, the prediction of two models from phase 1 are combined using hard voting strategy. In case of stacked generalization method, these two models are used as level-0 base models and one of them is used as meta-model. Both of these ensemble methods are implemented on the same learning set and predictions are compared in order to identify the best predictive model. The best predictive model is recommended for attaining early predicting diabetes diagnosis system.

3.5 Algorithm

The following algorithm briefs about the entire methodology executed in this paper. The execution of this algorithm starts with collecting dataset and ends up with providing enhanced prediction for diabetes diagnosis.

Algorithm for Diabetes Prediction:
Step-1: Obtain the dataset as D= { fi } where each fi denotes attribute and i ranges from 1 to n.
Step-2: Split D into D1 and D2 such that (0.8*D) $\in$ D1 and (D-D1) $\in$ D2.
Step-3: Use classifier set C={C1,C2,C3,C4} and fine-tune each Ci's if necessary.
Step-4: Fit D1 as the training set to each Ci.
Step-6: Use dataset D2 for testing and prediction.
Step-7: a. Compare actual and obtained prediction results of each Ci's with respect to performance evaluation metric set P={P1,P2,P3,P4}
    b. Prepare set Score as {S1, S2, S3, S4} such that each Si indicates score for each Ci's.
Step-8: Choose one classifier Cj that reaches Sj=max (Pi).
Step-9: Choose another classifier C_K that reaches highest SK such that SK < Sj.
Step-10: a. Obtain a Voting Ensemble Classifier C_V that ensembles prediction of Cj and C_k.
    b. Calculate Score S_V with respect to set P.
Step-11: a. Obtain Stacking Ensemble Classifier C_{SE} that ensembles prediction of Cj and C_K.
    b. Calculate Score S_{SE} with respect to set P.
Step-12: Compare S_V and S_{SE} and find the best performing model.

4. Experimental Results

In this section, performance of each of the implemented classifiers in this paper is shown with respect to performance evaluation metrics.

Table1. Performance of all implemented Phase-1 Classifiers

Performance Measure Metric	Accuracy	F1-Score	Cohen-kappa Score	MSE
Multinomial Naïve Bayes Classifier	63.64%	0.64	0.17	0.36
Multilayer Perceptron Classifier	71.43%	0.71	0.35	0.29
K-Nearest Neighbour Classifier	75.32%	0.75	0.43	0.25
Decision Tree Classifier	76.62%	0.77	0.46	0.21

Table2. Performance of all ensemble Phase-2 Classifiers

Performance Measure Metric	Accuracy	F1-Score	Cohen-kappa Score	MSE
Voting Ensemble Method	78.57%	0.79	0.44	0.21
Stacking Ensemble Method	79.87%	0.8	0.51	0.2

5.Discussion

From the aforementioned comparative analysis in Table 1, it is quite visible that the Decision Tree Classifier provides best result with respect to all the evaluation metrics. Next, K-Nearest Neighbour Classifier also provides relatively better result than other

Phase-1 classifiers except Decision Tree Classifier. These two classifiers are chosen for implementing phase-2 classifiers. In this phase, two ensemble methods are proposed in this paper. One is Voting Ensemble method and the other one is Stacking Ensemble method. Both of these proposed strategy ensembles the prediction of top two classifiers obtained during selection pane. The target of both of these classifiers is to enhance the performance of existing classifiers.

The purpose the proposed strategy is to boost the prediction outcome by combining the best two classifiers obtained during selection pane. In this case, predictions of Decision Tree classifiers and K-Nearest Neighbour Classifier are combined using ensemble methods to maximize the performance of prediction tool. Following table2 shows the performance the proposed method in terms of evaluation metrics. The results shown in table 2 indicate that voting ensemble method maximizes the efficiency of prediction over other phase 1 classifiers. But this method is not giving better result with respect to Cohen-kappa score over Decision Tree classifier. So another phase-2 ensemble method comes into play. Stacking ensemble method outperforms even better than voting ensemble strategy and finally this method can be used while detecting diabetic patients.

6. CONCLUSION

Early detection of diabetes plays significant role in the field of health care systems. In this context to assist diagnostic system an automated prediction has been carried out using classification algorithms. Several classifiers are utilised while predicting diabetic tendency of a patient. Each of these classifiers considers several interfering factors while detecting diabetic tendency. Top two classifiers are selected and utilised for ensemble method. In ensemble method, multiple algorithms are trained on same learning set and prediction of each participating class is considered while obtaining the final prediction results.

Two different approaches of ensemble methods, i.e., Voting Ensemble and Stacked generalized ensemble methods are offered in this paper. Each of these combines Decision Tree and K-Nearest Neighbour classifier since they outperform well over their peer phase-1 classifiers. Both of these ensemble methods indicate significantly better result over other classifier models. However, stacked generalization method shows superior result than voting classifier model. The combined prediction provided by stacked generalization method reaches accuracy of 79.87%, F1-Score 0.8, Cohen-Kappa Score 0.51 and MSE 0.2 which is quite efficient in terms of diabetic tendency prediction.

7. REFERENCES

1. Mishra S, Chaudhury P, Mishra BK, Tripathy HK. An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. ACM Int Conf Proceeding Ser. 2016;04-05-Marc:3–5.

2. Gnana A, Leavline E, Baig B. Diabetes Prediction Using Medical Data. J Comput Intell Bioinforma. 2017;10(January):1–8.

3. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol. 2015;9(1):86–90.

4. Xie Z, Nikolayeva O, Luo J, Li D. Building risk prediction models for type 2 diabetes using machine learning techniques. Prev Chronic Dis. 2019;16(9):1–9.

5. Wei S, Zhao X, Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. IEEE World Forum Internet Things, WF-IoT 2018 - Proc. 2018;2018-Janua:291–5.

6. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. Proc Int Conf Intell Sustain Syst ICISS 2017. 2018;(Iciss):533–8.

7. AlJarullah AA. Decision tree discovery for the diagnosis of type II diabetes. 2011 Int Conf Innov Inf Technol IIT 2011. 2011;303–7.

8.Michael Kahn, St. Louis. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

9.Kaviani P, Dhotre S. International Journal of Advance Engineering and Research Short Survey on Naive Bayes Algorithm. Int J Adv Eng Res Dev. 2017;4(11):607–11.

10.Walters DE. Bayes's Theorem and the Analysis of Binomial Random Variables. Biometrical J. 1988;30(7):817–25.

11. Marius-Constantin P, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Trans Circuits Syst. 2009;8(7):579–88.

12. Cunningham P, Delany SJ. K -Nearest Neighbour Classifiers. Mult Classif Syst. 2007;(May):1–17.

13. Sharma H, Kumar S. A Survey on Decision Tree Algorithms of Classification in Data Mining. Int J Sci Res. 2016;5(4):2094–7.

14. M H, M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag Process. 2015;5(2):01–11.

15. Vieira SM, Kaymak U, Sousa JMC. Cohen's kappa coefficient as a performance measure for feature selection. 2010 IEEE World Congr Comput Intell WCCI 2010. 2010;(May 2016).

16. Leon F, Floria SA, Badica C. Evaluating the effect of voting methods on ensemble-based classification. Proc - 2017 IEEE Int Conf Innov Intell Syst Appl INISTA 2017. 2017;(February 2018):1–6.

17. Ma Z, Wang P, Gao Z, Wang R, Khalighi K. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. PLoS One. 2018;13(10):1–12.

*Corresponding author: Bandyopadhyay Kumar Samir
E-Mail: 1954samir@gmail.com