Diabetes Prediction Using Machine Learning Algorithms

1st Md. Uzzal Rahman
Id: 2016-2-60-151
Dept. of computer Science and Engineering
East West University
Dhaka, Bangladesh
2016-2-60-151@std.ewubd.edu

1st Sabrina Islam
Id:2016-2-60-144
Dept. of computer Science and Engineering
East West University
Dhaka, Bangladesh
2016-2-60-144@std.ewubd.edu

1st Swakshar Debnath
Id:2017-2-60-034
Dept. of computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-2-60-034@std.ewubd.edu

Abstract*— This paper helps in predicting diabetes by applying machine learning technique. Thus, the most important issue is the prediction to be very accurate and to use a reliable method for that. Two different machine learning algorithms are used in this research work. For experiment purpose, a dataset of patient's medical record is obtained and three different machine learning algorithms are applied on the dataset. Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. The results show that the machine learning algorithms can able to produce highly accurate diabetes predictive healthcare systems.

Keywords*— Machine learning algorithms, Diabetes prediction, Decision trees, Random Forest, Data classification.

1. Introduction

Diabetes is a long-lasting disease that happens when the pancreas fails to create enough insulin, or when the body cannot use the insulin produced efficiently. Insulin is a hormone that controls the level of sugar in the blood. Hyperglycemia or hyperglycemia is a common result of uncontrolled diabetes and, over time, causes severe damage to many organs, particularly nerves and blood vessels.With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels [2,3]. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors [4]. For machine learning method, how to select the valid features and the correct classifier are the most important problems.

Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method [1], such as decision tree (DT), random forest and so on.

Machine learning methods are widely used in predicting diabetes, and they get preferable results. Random Forest (RF) and Decision tree are one of popular machine learning methods in medical field, which has grateful classification power. These algorithms has better performance in many aspects. So, in this study, we used decision tree and random forest (RF) to predict the diabetes.

2. Literature review

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data [19]. Moustakas and Charisse's' work [20] surveyed the role of machine learning in medical decision making and provided an extensive literature review on various ML applications in medicine that could be useful to practitioners interested in applying ML methods to improve the efficiency and quality of decision-making systems in medical applications.

Apart from the works mentioned above, a lot of research has been done specifically using ANN in diagnosing diabetes mellitus and some approaches are discussed below.
Siti Farhanah, Bt Jaffar and Dannawaty Mohd [18] proposed a method for diagnosing diabetes. The diagnosis is accomplished using back propagation neural network algorithm. The inputs to the system are plasma glucose concentration, blood pressure, triceps skin fold, serum insulin, Body Mass Index (BMI), diabetes pedigree function, number of times a person was pregnant and age. The biggest challenge to this method was the missing values in the data set. This system was later modified and presented by T.Jayalakshmi and Dr.A.Santhakumaran[21].They have proposed an idea to overcome the missing values that was not addressed by Siti Farhanah Bt Jaafar [18] and this included constructing the data sets with reconstructed missing values, thereby increasing the classification accuracy[21]. They have also proposed an alternate method

to overcome missing value by performing data pre-processing, which also speeds up the training process by reducing the actual learning time. Various missing value techniques and pre-processing methods were analyzed. By adopting these modifications, the results improved and achieved a classification accuracy of 78% [18].

Machine learning Methods:

Data Requirements:

Dataset: We have required data set of Diabetes for the simulations. Its data collected from the Kaggle. In this data set we have 768 row and 9 columns.

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **Blood Pressure:** Diastolic blood pressure (mm Hg)
- **Skin Thickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age:** Age (years)
- **Outcome:** Class variable (0 if non-diabetic, 1 if diabetic)

Diabetes or diabetes mellitus is a metabolic disorder (metabolic) in the body. This disease destroys the ability to produce insulin in the patient's body or the body develops resistance to insulin the and consequently the produced insulin cannot achieve its normal job. The main role of the produced insulin is to decrees blood sugar by different instruments. There are two key types of diabetes. In Type I diabetes, obliteration of beta pancreatic cells damage insulin construction and in type II, there is a progressive insulin confrontation in the body and ultimately may yield to the obliteration of pancreatic beta cells and faults in insulin production. In type II diabetes, it is known that genetic issues, obesity and lack of physical activity have a vital part in a person [5].

Even though the precise cause of type I diabetes is unidentified, issues that may indicate a greater risk comprise the followings [6]:

Family history**.** A person risk upsurges if his parent or sibling has history of type I diabetes.

Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.

The existence of harmful immune system cells. Occasionally family members of a person with type I diabetes are examined for the existence of diabetes autoantibodies. If a person has these autoantibodies, he/she has a chance of increased risk for evolving type I diabetes. Nonetheless not every person who has these autoantibodies gets diabetes.

Geography. Some countries, like Sweden, have bigger rates of type I diabetes.

Environmental factors. Situations for example contact with a viral illness probably play some role in type I diabetes.

Researchers don't completely comprehend why certain people develop pre-diabetes and type II diabetes and others don't. It's sure that some factors upsurge the risk like [6]:

Weight. The more fatty tissue you have, the more resilient a person cells to insulin.

Inactivity. The less energetic a person is, the more a person has risk. Physical activity assists a person control of his/her weight, consumes glucose as energy and makes a person cells more sensitive to insulin. Family history. A person risk upsurges if his parent or sibling has history of type II diabetes.

Race. Even though it's uncertain why, people of specific races are at higher risk.

Age. A person risk upsurges as he/she gets older. This may be because a person has a habit to exercise less, lose muscle mass and add weight as he/she gets older. Nonetheless type II diabetes is likewise growing among children, youths and adults.

Gestational diabetes. If a person developed gestational diabetes when she was pregnant, her risk of emerging pre-diabetes and type II diabetes far ahead upsurges. If she gives birth to a baby weighing more than 4 kilograms, she is also at risk of type II diabetes.

Polycystic ovary syndrome. For females, having polycystic ovary syndrome increases the risk of getting diabetes.

High blood pressure. Having blood pressure more than 140/90 millimeters of mercury (mm Hg) is connected to an augmented risk of type II diabetes.

Abnormal cholesterol and triglyceride levels. If a person has low levels of high-density lipoprotein, or good cholesterol, his/her risk of type II diabetes is going to be higher. Triglycerides are additional type of fat passed in the blood. A person with greater levels of triglycerides has an augmented risk of type II diabetes.

A number of disease prediction models are used in medical diagnosis system which are using data mining and machine leaning techniques like Bayesian classification, Decision Tree, Regression model, Neural Network etc. We choose neural-network, logistic regression and decision tree for solving this problem.

2.1 Decision tree:

Decision tree is a basic classification and regression method. Decision tree model has a tree structure, which can describe the process of classification instances based on features [7]. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space.

Decision tree uses tree structure and the tree begins with a single node representing the training samples [8]. If the samples are all in the same class, the node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into serval subsets, each of which forms a branch, and there are serval values that form serval branches [9]. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples [9,10].

The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the J48 decision tree in WEKA. J48 another name is C4.8, which is an upgrade of C4.5. J48[11] is a top-down, recursive divide and conquer strategy. This method selects an attribute to be root node, generates a branch for each possible attribute value, divides the instance into multiple subsets, and each subset corresponds to a branch of the root node, and then repeats the process recursively on each branch [12]. When all instances have the same classification, the algorithm stop. In J48, the nodes are decided by information gain. According to the following formulas, in each iteration, J48 calculates the information gain of each attribute, and selects the attribute with the largest value of information gain as the node of this iteration [13].

Attribute A information gain:
Gain(A)=Info(D)−InfoA(D)
Pre-segmentation information entropy:
Info(D)=Entropy(D)=−Σjp(j|D)logp(j|||d)
Distributed information entropy:
InfoA(D)=Σi=1vninInfo(Di)

2.2 *Random Forests:*

Random Forest developed by Leo Breiman [4] is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in [24]: By Sampling N randomly, If the number of cases in the training set is N but with replacement, from the original data. This sample will be used as the training set for growing the tree. For M number of input variables, the variable m is selected such that m<<M is specified at each node, m variables are selected at random out of the M and the best split on these m is used for splitting the node. During the forest growing, the value of m is held constant. Each tree is grown to the largest possible extent. No pruning is used. Random Forest generally exhibits a significant performance improvement as compared to single tree classifier such as C4.5. The generalization error rate that it yields compares favorably to Adaboost, however it is more robust to noise.

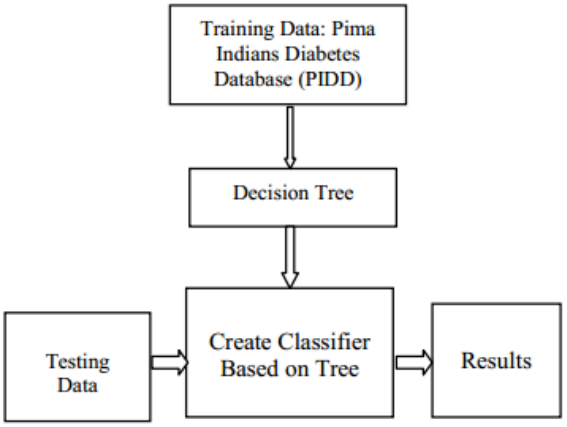

Figure 1: Algorithmic Sequence of Random Forest Classifier

3. Result Analysis

Our Diabetes dataset contains 770 rows in total. We splitted our dataset into test and train datasets. The Ratio is 80% train data and 20% for data to be tested. 29 internal trees were generated in our experiment. Using Random Forest we got 80.52% accuracy and error rate was 0.19%. We also calculated precision, recall, f1-score, support values. The resulting values are following:

	precision	recall	f1-score	support
0	0.85	0.83	0.84	107
1	0.63	0.66	0.65	47
accuracy			0.78	154
macro avg	0.74	0.75	0.74	154
weighted avg	0.78	0.78	0.78	154

Figure 2: Result analysis of Random Forest

We also generated the Confusion matrix of Random forest implementation. The result is shown below:

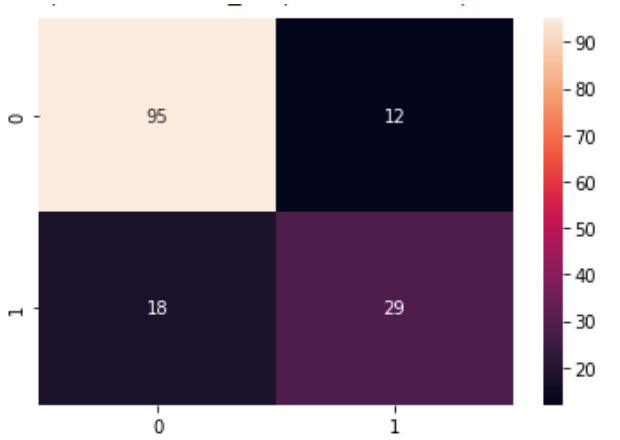

Figure 3: Confusion Matrix of Random Forest

In this paper, we also experimented with Decision trees . And compared with Random Forest. Using Decision trees we got 77.92% accuracy and error rate was 0.22%. We also calculated precision, recall, f1-score, and support values. The resulting values are following:

	precision	recall	f1-score	support
0	0.85	0.83	0.84	107
1	0.63	0.66	0.65	47
accuracy			0.78	154
macro avg	0.74	0.75	0.74	154
weighted avg	0.78	0.78	0.78	154

Figure 4: Result analysis of Decision trees

We also generated the Confusion matrix of Decision Trees implementation. The result is shown below:

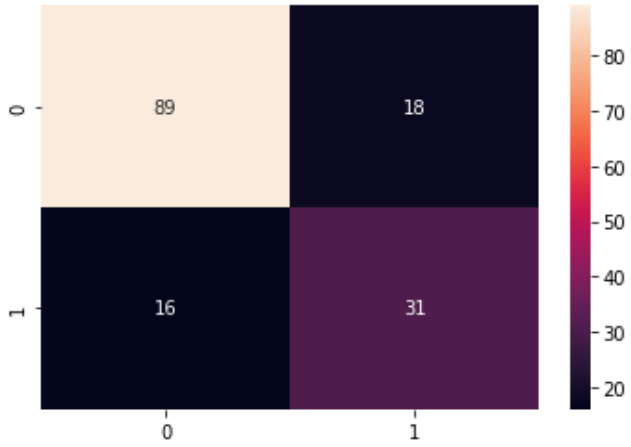

Figure 5: Confusion Matrix of Decision Trees

Finally, we compared the performance of both algorithms. And from the aforementioned result analysis, we can conclude that Random forest outperforms Decision trees in both accuracy and error rate.

	Algorithm	Accuracy	Error_rate
0	Randon Forest	80.519481	0.194805
1	Decision Tree	77.922078	0.220779

Figure 6: Comparison between Random Forest and Decision Trees

4. Conclusion and Future Work

In this paper, the issue of current medical diagnosis system and various machine learning algorithms are used for the medical prediction is explained. The focus is on using different algorithms and consolidation of certain target attributes to predict lung cancer effectively using machine learning algorithms. For predicting lung cancer, significantly 6 attributes are listed and give priories using information gain and we have already applied the machine learning techniques like CNN, logistic regression on diabetes actual data to get the optimal outputs. The proposed work will be further increased developed for the automation of the diabetes disease prediction more accurately.

5. REFERENCES

[1].Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116. doi: 10.1016/j.csbj.2016.12.005

[2].American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. doi: 10.2337/dc12-s064

[3].Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27, 132–138. doi: 10.2337/diaclin.27.4.132

[4].Lee, B. J., and Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20, 39–46. doi: 10.1109/JBHI.2015.2396520

[5] World Health Organization (WHO), "Definition, Diagnosis, and classification of diabetes mellitus and its complications", part 1. WHO/NCD/NCS/2016.2, (2016).

[6] H. Temurtas, N. Yumusak and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", Expert System, vol. 36, (2009), pp. 8610–15.

[7]Quinlan, J. R. (1996a). “Bagging, boosting, and C4.5,” in Proceedings of the Thirteenth National Conference on

Artificial Intelligence (Menlo Park, CA: AAAI Press), 725–730.

[8]Liao, Z. J., Wan, S., He, Y., and Zou, Q. (2018). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. Curr. Bioinform. 13, 492–500. doi: 10.2174/1574893612666171121162552

[9]Quinlan, J. R. (1996a). "Bagging, boosting, and C4.5," in Proceedings of the Thirteenth National Conference on Artificial Intelligence (Menlo Park, CA: AAAI Press), 725–730.

[10]Habibi, S., Ahmadi, M., and Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob. J. Health Sci. 7, 304–310. doi: 10.5539/gjhs.v7n5p304

[11]Salzberg, S. L. (1994). C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, Inc., 1993. Mach. Learn. 16, 235–240.

[12]Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal

[13]Sharma, S., Agrawal, J., and Sharma, S. (2014). classification through machine learning technique: C4. 5 algorithm based on various entropies. Int. J. Comput. Appl. 82, 28–32

[19] Ranjit Abraham, Jay B.Simha, Iyengar S. "Medical datamining with a new algorithm for feature selection and Naïve Bayesian classifier", 10th International Conference on Information Technology.

[20] Moustakis, V. and Charissis, G. (1999). Machine learning and medical decision making, In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence- ACAI99, Chania, Greece, 1-19.

[18] Siti Farhanah Bt Jaafar and Dannawaty Mohd Ali, "Diabetes mellitus forecast using artificial neural networks", Asian conference of paramedical research proceedings, 5-7, September, 2005, Kuala Lumpur, MALAYSIA

[21] T.Jayalakshmi and Dr.A.Santhakumaran, "A novel classification method for classification of diabetes mellitus using artificial neural networks". 2010 International Conference on Data Storage and Data Engineering.