CLASSIFICATION OF DIABETES DISEASE USING MACHINE LEARNING ALGORITHMS

Nirjharini Mohanty[1], Soumen Nayak[2], Monarch Saha[3], Vishal Baral[4] and Imlee Rout[5]
[1,3,4,5]*Department of Computer Science and Information Technology, Institute of Technical Education and Research, India*
[2]*Department of Computer Science and Engineering, Institute of Technical Education and Research, India*

Abstract

The inclusion of Information Technology in various fields like that of health care has proven to boost the existing progressions at all times, like Machine Learning algorithms that automate the manual effort of people. The advancements in the health care sector are always in demand and are studied upon. With the growing population and reckless lifestyles, the urgency to find proper tools and precautions for substantial diseases is increasing. The prediction algorithms have been providing the sufferer with a head start to adequate treatment and diagnosis. A significant part of the human population suffers from diabetes. The disease can affect the immune system costly by inducing various other diseases. Diabetes is the disease that is focused on in this paper, by using two Machine Learning Classification Algorithms to conclude with the better performing classifier while predicting if a person might have diabetes by a given set of data along with the deliberations on the main characteristics and significance of the two classifiers namely Gradient Boosting (GrB) classfier and Extra Trees (ExT) Classifiers. The latter section will explain why GrB Classifier surpasses ExT Classifier when it comes to predicting diabetes in a patient. As Accuracy percentage of GrB is 73.3%.

Keywords:

Machine Learning, GrB, ExT, Classifier, Diabetes

1. INTRODUCTION

In accordance with the reports presented by the World Health Organization, worldwide 422 Million people suffer from Diabetes which is 1 in every 11 people. A spike can be clearly observed when the cases recorded were 108 million in 1980 to 422 million in 2014. The tragic consequences are only known to those who look into the matter. Heart attacks, complete loss of eyesight, stroke, failure of kidneys, and limb amputations are only some of the major ways of how Diabetes affects humankind. Diabetes can be lethal and extremely overpriced when it comes to treatment. While it is difficult to assess people on an individual level on maintaining a healthy living and healthful lifestyle including a constant form of exercise as it is expected from the increase in the obesity rates worldwide to increase the risk of diabetes and diseases related to the vascular system [1], what is necessary is an adequate prognosis of the ailment. When diagnosed beforehand, it will help the patient to avoid long term misery.

Machine learning, in numerous ways, has been aiding the health care sector for years. As the understanding of precautions and diseases increases over time, technology becomes more sought after. Like the method of prediction of diseases from a list of attributes, has been playing a major role with the accurate prediction of diseases like kidney diseases [3], heart attacks, risk of chronic diseases [2], cancers [5] [8], etc. Machine Learning, in particular, will help come up with classifiers that will be able to predict the status of the disease in a patient in near future with the highest accuracy. It has also been contemplated for machine learning to be a constituent of health-care with organizations insisting to exert efforts further in this sphere to serve and diminish human indulgence by several contributors and researchers in this delicate yet significant part indifferent of the field of medicine [7]. The applications of such algorithms have even assisted people to examine the elucidation of conditions or ailments from the pathological reports and data to understand the necessity of patients and support when medical specialists are unavailable [6].

The point of convergence of the paper is to figure out the better classifier among the two, namely GrB and ExT Classifiers in the prognosis of diabetes in an individual. The data-set fed to the classifiers has a various number of attributes that might play a role in contributing to the growth and development of the ailment in a person. The paper even helps with the understanding of the two algorithms and draws a clear conclusion by using figure and numbers in the end.

2. RELATED WORKS

The assistance of Machine Learning in bio-informatics is not a new idea, over time, the interest in these domains together has increased gradually, increasing the number of people researching on the same.

Bhardwaj et al. [7] address the potential of employing various technologies in the healthcare sector and outline industry actions applying machine learning in the medical management sector. Finally, concluding how technologies like machine learning and big data have the potential to help patients and health care professionals equally in terms of ampler care and lower charges as the expenditure by the government on healthcare has attained an unsurpassed high. They even believe that there are information and data available and all that has to be done to achieve the aim is to find a way to interpret them. Definitely, machine learning presents a fruitful opportunity for humankind to do so.

Ravi et al. [4] shed a light upon the emerging and promising popularity of deep learning, as an efficacious mechanism for Machine Learning. Their study presented a thorough evaluation of research exercising deep learning in bioinformatics and health informatics. The focus of the paper was the chief implementations of deep learning that serve in the field of bioinformatics in a few niches like that of medical imaging, informatics, etc.

The research work conducted by Bottaci et al. [5] was upon the prediction of results for individual patients treated for colorectal cancer as they addressed that prognosis in a large group is not beneficial on an individual level because the outcomes had a basis of population statistics. So the application of Deep Learning was to inscribe the training of neural networks for prediction of the result for each and every patient.

Kaveeshwar and Cornwall [9] concluded how Diabetes is in potential epidemic proportions in India. While it is not only a burden on a personal and economical level, the alarming fact is that the younger age groups are even getting diseased with the same, which might be costlier than anticipated. They even stress the urgency to raise the efforts put into the research. The roots of diabetes are dug deep due to lifestyle changes, unhealthy practices, etc. are to be eliminated, or at least the increase in the number of patients should be precautious decreased for the upcoming years on a national and regional level too.

The framework proposed by Zheng et al. [11] would identify if patients suffer from Type II Diabetes from a dataset of 300 patients and their respective records, by implementing machine learning and feature engineering that takes various features like self-diagnosis reports, medicines that are taken by the patient, etc into consideration. There were over 5 machine learning models to do the necessity like Decision Tree, Logistic Regression, etc. The outcome was significant with a high performance of ~0.98 on average AUC.

The machine learning strategy by Pliset al. [12] was for the prediction of blood glucose levels and exhibited the outcomes of the experiments in hypoglycemia prediction. They trained a Support Vector Regression on patient-specific data by utilizing the information generated by the support of a physiological prototype of blood glucose dynamics to finally conclude with their model to be capable of anticipating blood glucose levels in a better manner.

3. BACKGROUND

Supervised Machine Learning (SuML) is the method of building algorithms that are able to deliver customary patterns and conditions by using supplied occurrences for anticipating the result. The data input for SuML is always labeled as opposed to that of UnSuML methods.

Classification falls under the SuMLtechnique and aims at categorizing data from prior information that is split into testing and training data based on necessity. It is executed to distinguish the data components into corresponding classes that show common characteristics [10].

Here two of the classification algorithms are used to predict if the person is suffering from diabetes or not. The algorithms are:

3.1 EXTRA TREES CLASSIFIER

It is an ensemble learning technique which group the outcomes of multiple non-identical decision trees which is combined to form forest and then it return the required result. The ExT has an optional parameter for selection of subsamples that is bootstrap. In this classification the dataset split is done based on random cut points for each sub trees. The structure of ExT classifier is show in Fig.1.

3.2 GRADIENT BOOSTING CLASSIFIER

It is the combination of many machine learning algorithms that are many weak learning models together, and get a strong model which can be used for prediction of the required dataset.

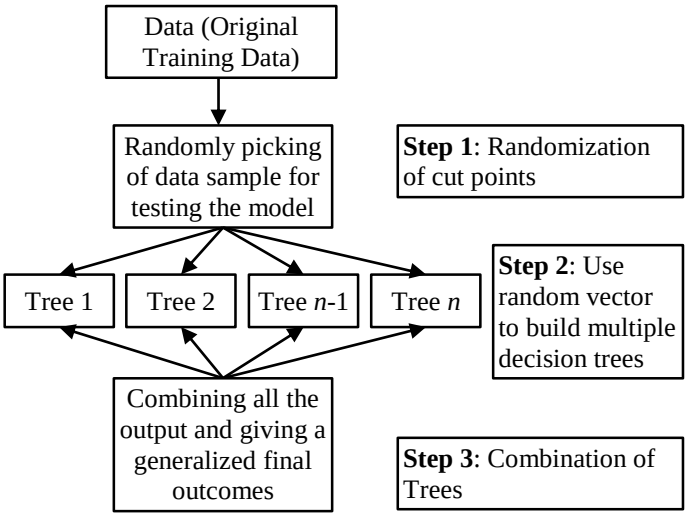

Fig.1. Structural Representation of ExT

4. RESULT AND DISCUSSION

This paper gives the outcome of Diabetes prediction in order to get the training set and testing set from dataset for the preparation and testing of the models which will be used for the prediction test of diabetes.

The dataset which is used for the testing of the model is taken from the online source (www.kaggle.com), and the data consist of 768 cases and each case has 8 attributes for each individual, the attributes are:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI (Body Mass Index)
- Diabetes Pedigree Function
- Age

The attributes will help the models return the binary values from the prediction test, the values that would be returned are:

- **Diabetes**: (Positive (1) or Negative (0))
- **Positive** (1): The patient is suffering from Diabetes.
- **Negative** (0): The patient is not suffering from Diabetes.

In this paper, two classification models are utilized (GrB Classifier and ExT Classifier) will accept the data of each patient and give us the result put together from the training dataset that is, whether the person has a chance of Diabetes or not. The training dataset will be brought into play to make the model acquainted with all the attributes and their values. Then the testing dataset will verify the accuracy of the output given by the model using different data value and their actual outcomes. The train-test split of this dataset is done as 75% of training and 25% of testing.

The techniques which are used are GrB Classifier and ExT Classifier. To settle upon the better classifier in predicting diabetes, some factors are evaluated. The main factor to evaluate the models is Accuracy.

Accuracy: It is the basic criteria for the evaluation of any model or algorithm. It is defined as the number of accurate responses given by the model concerning the total number of data cases given for the prediction.

Accuracy = Number of correct predictions/Total number of Predictions (1)

Table.1. Accuracy Score of the models

Classification Algorithm	Accuracy
GrB Classifier	0.733
ExT Classifier	0.732

From Table.1 we can deduce that the accuracy value of GrB Classifier and ExT Classifier is almost same. From the Table.1, we cannot conclude which model is better as the value of each model is almost equal. So, there is a need to evaluate other factors too.

Table.2. Average Precision, Precision and Recall of the models

Classification Algorithm	Average Precision	Precision	Recall
GrB classifier	0.675	0.646	0.539
ExT classifier	0.606	0.672	0.469

In the Table.2, factors corresponding to models are:

Precision: It checks all the True positive values with respect to the total positive cases. Where total positive cases includes both true positive and false positive.

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

Recall: It checks the True Positive cases with total case predicted by the model of the classifier.

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

Average Precision: It is the area under the curves of Precision and Recall. The value of average precision is always between 0 and 1 and the model which has the higher average precision is better model for prediction.

$$AP = \int_0^1 rdr \quad (4)$$

From the Figure 3 we can say the Avg Precision value of Gradient ishigher compared tothat of ExT. But the difference of the Avg Precision of both the classifiers is minimal (0.675 - 0.606 = 0.069) which is approx to 0.07.

So let us deliberate upon more factors that can be considered for the selection of the classifier. The factors which are shown in the Table.3, are F1-measure, log loss, ROC AUC, build time (s).

Table.3. Representation of F1, Log Loss, ROC AUC of the model

Classification Algorithm	F1	Log loss	ROC AUC	Build Time (s)
GrB Classifier	0.586	0.536	0.805	4
ExT Classifier	0.545	0.846	0.757	1

The F1-score is used when we seek a balance in between the value of precision and recall.

$$F1 = 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

Log Loss: It is the most essential factor for classification based on the probabilities.

ROC AUC: It is a curve which shows the performance of the classification problems at different threshold settings. ROC AUC are two distinguished parts, ROC is the probability curve and AUC is used to measure the separability or it represent the degree of separability.

Build Time: It is the time taken to create the prediction model.

From all the Table.3, it is clear that GrB classifier is better than ExT. As it clearly evident in Table.3, the Log loss value of GrB is less as compared to that of extra tress. And the time consumed by the classifier to generating the model for prediction is less of GrB than that of ExT Classifier.

5. CONCLUSION

After observations of all the results and comparing all the factors of the classifiers it is conclusive that the GrB classifier is better at predicting if a person has diabetes or not than the ExT Classifier. As the Accuracy of the GrB classifier is 73.3% making it greater than the ExT classifier's accuracy value. From Table.3 we can say that the log loss value of GrB (0.536) is comparatively less than that of ExT classifier (0.846), as we know lesser the log loss more.

REFERENCES

[1] Emerging Risk Factors Collaboration, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies", *The Lancet*, Vol. 375, No. 9733, pp. 2215-2222, 2010.

[2] L. Jena, S. Nayak and R. Swain, "Chronic Disease Risk (CDR) Prediction in Biomedical Data using Machine Learning Approach", *Advances in Intelligent Computing and Communication*, Vol. 109, pp. 1-13, 2021.

[3] L. Jena, B. Patra, S. Nayak and S. Mishra, "Risk Prediction of Kidney Disease using Machine Learning Strategie", *Intelligent and Cloud Computing Smart Innovation, Systems and Technologies*, pp. 1-12, 2021.

[4] D. Ravì, "Deep Learning for Health Informatics", *IEEE Journal of Biomedical and Health Informatics*, Vol. 21, No. 1, pp. 4-21, 2017.

[5] L. Bottaci, P.J. Drew and J.E. Hartley, "Artificial Neural Networks are Applied to Outcome Prediction for Colorectal Cancer Patients in Separate Institutions", *The Lancet*, Vol. 350, No. 9076, pp. 469-472, 1997.

[6] C. Singh, N. Cheggoju and V.R. Satpute, "Implementing Classification algorithms in Medical Report Analysis for Helping Patient During Unavailability of Medical Expertise", *Proceedings of International Conference on Computing, Communication, and Networking Technologies*, pp. 123-132, 2018.

[7] R. Bhardwaj, A.R. Nambiar and D. Dutta, "A Study of Machine Learning in Healthcare", *Proceedings of IEEE*

Annual Conference on Computer Software and Applications, pp. 1-14, 2017.

[8] B. Patra, L. Jena and S. Bhutia, "Evolutionary Hybrid Feature Selection for Cancer Diagnosis", *Proceedings of International Conference on Intelligent and Cloud Computing*, pp. 253-263, 2021.

[9] S.A. Kaveeshwar and J. Cornwall, "The Current State of Diabetes Mellitus in India", *Australasian Medical Journal*, Vol. 7, No. 1, pp. 45-48, 2014.

[10] A. Singh, N. Thakur and A. Sharma, "A Review of SuML Algorithms", *Proceedings of International Conference on Computing for Sustainable Global Development*, pp. 1310-1315, 2016.

[11] T. Zheng, W. Xie and L. Xu, L., "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records", *International Journal of Medical Informatics*, Vol. 97, pp. 120-127, 2017.

[12] K. Plis, R. Bunescu and C. Marling, "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-12, 2014.