Int J Adv Life Sci Res. Volume 5(1) 29-34
https://doi.org/10.31632/ijalsr.2022.v05i01.005

International Journal of Advancement in Life Sciences Research

Online ISSN: 2581-4877

Journal homepage http://ijalsr.org

Original Article

Assessing Machine Learning Methods for Predicting Diabetes among Pregnant Women

Zahura Zaman, Ashrak Al Arif Shohas*, Mahedi Hasan Bijoy, Meherab Hossain,, Shakawat Al Sakib

Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

**Correspondence E-mail :* ashraks97@gmail.com

Abstract

Machine Learning has a big impact on a lot of different scientific and technical disciplines, including medical research and Biophysics. Diabetes is a chronic condition marked by abnormally high glucose levels and the body's inefficient utilization of insulin. Diabetes is now becoming a leading cause of death all over the world. The objective of this article was to use multiple. Machine Learning methods are used to create a model with a limited number of dependencies, which could be used to study diabetic patients and diagnose diabetes using the PIMA dataset. Some of the most well-known prediction algorithms employed in this system are SVM (support vector machine), Multinomial Naive Bayes, Random forest, and Decision tree. Use these algorithms to construct a gathering of models by combining multiple combinations into one. This will enhance performance and accuracy.

Keywords: Diabetes prediction; dataset of PIMA; train test split; machine learning; classification.

Introduction

In our present world, Diabetes is a very kin word and crucial challenge in both developed and developing countries. High blood sugar is the most common symptom of diabetes. High blood sugar causes symptoms such as increased thirst, increased hunger, and frequent urination. If diabetes is not treated, it can lead to a variety of complications (Barrett-Connor, 2003).

According to a report (Report of 2016) of the world health organization (WHO), On World Diabetes Day, Eye on Diabetes reports that there are Diabetes affects 422 million individuals, resulting in 1.6 million fatalities, and the document notes that it's not hard to guess how severe chronic diabetes (Collaboration, 2010).

High blood glucose (HBG) levels were a problem in 2012, which was the reason behind 2.2 million people have died in 2014, it is found that Diabetes was diagnosed in 8.5 percent of persons aged 18 and up. That is not something we expect normally (Hawksworth D.L 2001).

CDCP (Centers for Disease Control and Prevention) says, type 1 diabetes climbed by 23 percent in the United States over the next nine years, from 2001 to 2009. Countries, organizations, and various health sectors are afraid of controlling and preventing this chronic disease before a person dies (Hawksworth D.L& Lücking R, 2017).

The lack of insulin chemicals caused the pancreas to break down, resulting in trance-like states, obsessive obliteration of pancreatic beta cells, renal and retinal disappointment, cerebral vascular brokenness, cardiovascular obstruction, fringe vascular sickness, weakness, weight loss, ulcers, joint disappointment, and pathogenic resistant reactions (Leong A.*et.al* 2014). The pancreas produces the chemical insulin, which allows glucose to flow from the diet to the circulatory system in humans. Many factors influence diabetes, including stature, weight, genetic characteristics, and insulin, but the sugar level is the most important of them. The only way to avoid this is to notice it as soon as possible (Pranto B *et al*., 2020).

Our research is focused on diabetes-affected pregnant mothers. In this study, we used and assessed some machine learning classification algorithms on the PIMA dataset which will find the risk of diabetes. On a variety of measures, the experimental performance of these two algorithms is compared, and good precision is achieved.

Proposed System

These days Machine learning techniques are generally used to foresee diabetes. AI empowers a machine to gain naturally as a matter of fact and to anticipate effectively when new occurrences happen. We explored different avenues regarding four distinct calculations to fabricate our prescient model. The primary calculation that we utilized was SVM (Support Vector Machines), next we utilized Decision Tree Classifier, then, at that point Multinomial Naive Bayes, and finally, we attempted with Random Forest.

The data collected from the Kaggle repository's PIMA Indians dataset. The dataset comprises information on females who are at least 21 years old, with a total of 768 records.

1 Train Test Split

An approach for assessing the performance of a machine learning is the train-test split. It may be used for classification or regression issues, and any supervised learning method can be used with it. A dataset is divided into two subsets as part of the method.

20% data for test

80% data for train

random_state = 1

2 Random forest classifier:

Random forest combines the results of many decision trees to provide a more accurate and consistent forecast. The parameters of a random forest are quite similar to decision tree. While growing the trees, this algorithm adds more unpredictability to the model.

3 Decision Tree classifier:

The decision tree model's tree structure may be used to depict the process of categorizing instances based on their features (Radja M & Emanuel A W R, 2019).

A decision tree's tree structure is employed, with the tree starting with a single node that represents the training samples (Team, 2011). The node becomes the leaf if all of the samples belong to the same class, and the class is used to identify it. If that does not do then the algorithm selects the discriminating attribute as the current node of the decision tree.

4 Naive Bayes:

The Bayes theorem is the foundation of Naive Bayes, which states that characteristics in a dataset are mutually independent. The occurrence of one trait has no bearing on the likelihood of the occurrence of the other. Naive Bayes can outperform the most powerful alternatives for small sample sets. It is utilized in a variety of disciplines due to its relative robustness, ease of implementation, speed, and accuracy.

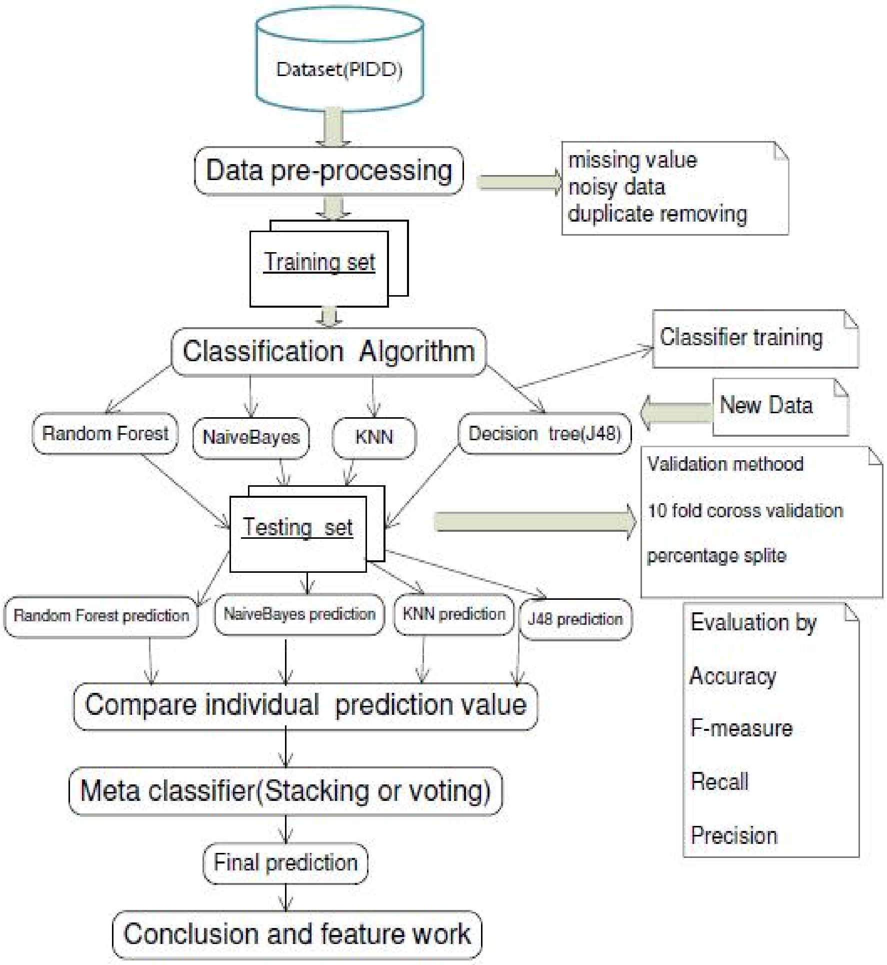

Figure 1: Detail Architecture of work flow

Implementations

Machine learning algorithms were applied in this section. Some results come out after implement.

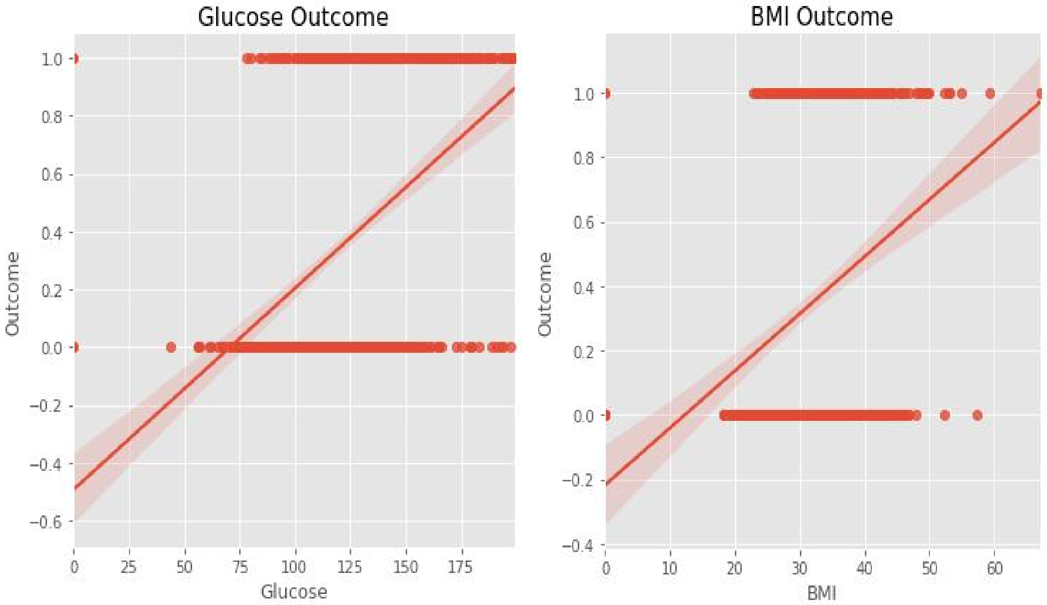

Figure 2: Glucose and BMI outcome.

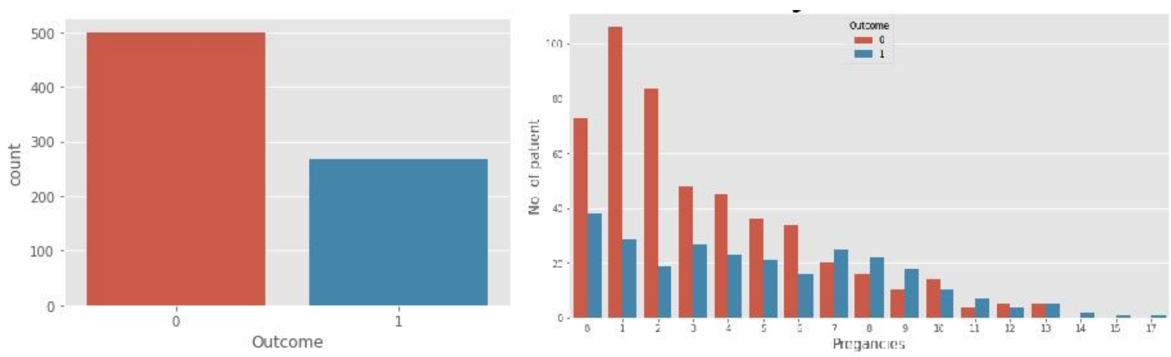

Figure 3: Pregnancies outcome

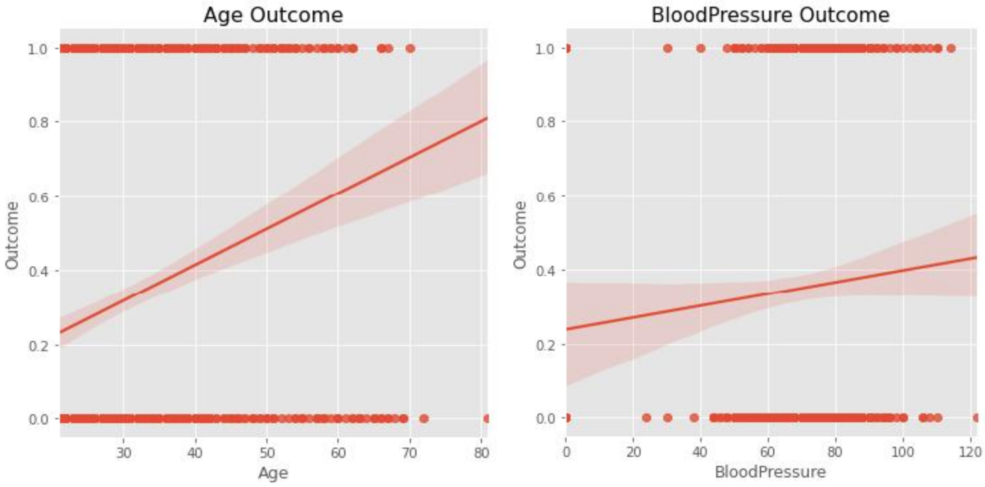

Figure 4: Age & Blood Pressure outcome

Results

Table 1: Accuracy with Confusion Matrix.

Name of algorithms	"Accuracy Score"	"True Positive"	"False Positive"	"False Negative"	"True Negative"
SVM	"78"	"93"	"27"	"6"	"28"
NB	"81"	"90"	"20"	"9"	"35"
DT Classifier	"70"	"77"	"24"	"22"	"31"
MNB	"51"	"59"	"35"	"40"	"20"

Table 2: Accuracy with Confusion Matrix of all algorithms

		Support Vector Machine (SVM)		
Accuracy			0.79	154
Macro Average	0.72	0.80	0.74	154
Weighted Average	0.84	0.79	0.80	154
		Naïve Bayes		
Accuracy			0.81	154
Macro Average	0.77	0.81	0.78	154
Weighted Average	0.83	0.81	0.82	154
		Decision Tree Classifier		
Accuracy			0.70	154
Macro Average	0.67	0.67	0.67	154
Weighted Average	0.70	0.70	0.70	154
		Multinomial Naïve Bayes		
Accuracy			0.51	154
Macro Average	0.48	0.48	0.48	154
Weighted Average	0.51	0.51	0.51	154

Discussion

The PIMA dataset, which is accessible at the UCI Machine Learning repository, was utilized in several of the articles evaluated here. We concentrated on these publications since we rely on this dataset for our research (Vaishali R. *et.al*. 2017). In one study, the major computation was SVM (Support Vector Machines), followed by Decision Tree Classifier, Multinomial Naive Bayes, and Random Forest. The tree structure of the decision tree model may be used to illustrate the process of classifying instances depending on their features, according to Friedl and Brodley (Andallu B. et.al.2001). It may be regarded as a collection of if-then rules in feature and class space, as well as conditional probability distributions (Mahapatra D.K. *et.al*.2015). The tree structure of a decision tree is used, using a single node representing the training samples at the root. The equation was employed in one study to show how the Nave Bayes classifier arrived at its result. In another investigation, we observed that although several factors impact diabetes, such as height, weight, genetics, and insulin, the sugar level is the most important. The paper of Radja and Emanuel undertook a series of tests to find out assess the effectiveness of supervised Machine Learning algorithms for diabetes prediction using various data sizes. One study used comparative analysis to predict diabetes with a high accuracy outcome utilizing several Machine Learning method (Hung H.Y. et.al. 2012).

The most accurate method was determined to be logistic regression and this resulted in an accuracy of 78.01 percent. According to their findings, logistic regression combined with sequential minimum optimization (SMO) and multilayer perceptron (MLP) can accurately predict the onset of diabetes by 78 percent. Bagging, Nave Bayes, and the Support Vector Machine (SVM) all yielded positive results.

Conclusion

This research method is designed for women who are pregnant and who experience non-periodic fluctuations in blood sugar levels throughout pregnancy. By entering a few parameters, our technology can help you forecast your current diabetes level. Various data mining techniques and their applications have been researched and reviewed. Machine learning techniques were used to analyze various medical data sets. Machine learning algorithms have varying degrees of power depending on the data set. The single method was less accurate than the ensemble approach. The decision tree demonstrated good accuracy in the majority of studies. Random Forest provides the utmost accuracy in this project. We built a model that can predict diabetes of pregnant women where these methods to engage in a technique or other necessary means where that person may adhere to a tight plan to decrease their risks of contracting this terrible condition or the impact of this disease.

In the future, We plan to collect more information on In which data set that can lead us to a higher accuracy rate.

Acknowledgments

The authors would like to thank the department of Biotechnology and Genetic Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh for supporting and discussing about this research article.

Disclosure statement

The authors declare no conflict of interest.

Author contributions

AAAS, ZZ, SAS, MH and MHB designed the project and performed the experiments; AAAS, ZZ, SAS evaluated and interpreted the data; SAS, MH and MHB prepared the draft manuscript and finalized the manuscript. All authors approved the final version of the manuscript.

References

Andallu, B., Suryakantham, V., Srikanthi, B. L., & Reddy, G. K. (2001). Effect of mulberry (Morus indica L.) therapy on plasma and erythrocyte membrane lipids in patients with type 2 diabetes. *Clinica Chimica Acta*, *314*(1-2), 47-53. https://doi.org/10.1016/S0009-8981(01)00632-5

Barrett-Connor, E. (2003). Diabetes and heart disease. *Diabetes Care, 26*(10), 2947-2958. https://doi.org/10.2337/diacare.26.10.2947

Emerging Risk Factors Collaboration. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet*, *375*(9733), 2215-2222. https://doi.org/10.1016/S0140-6736(10)60484-9

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1· 5 million species estimate revisited. *Mycological research, 105*(12), 1422-1432. DOI: https://doi.org/10.1017/S0953756201004725

Hawksworth, D. L., & Lücking, R. (2017). Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology spectrum, 5*(4), 5.4. 10. DOI:https://doi.org/10.1128/microbiolspec.FUNK-0052-2016

Hung, H. Y., Qian, K., Morris-Natschke, S. L., Hsu, C. S., & Lee, K. H. (2012). Recent discovery of plant-derived anti-diabetic natural products. *Natural product reports*, *29*(5), 580-606. https://doi.org/10.1039/C2NP00074A

Leong, A., Rahme, E., & Dasgupta, K. (2014). Spousal diabetes as a diabetes risk factor: a systematic review and meta-analysis. *BMC medicine, 12*(1), 1-12. https://doi.org/10.1186/1741-7015-12-12

Mahapatra, D. K., Asati, V., & Bharti, S. K. (2015). Chalcones and their therapeutic targets for the management of diabetes: structural and pharmacological perspectives. *European journal of medicinal chemistry*, *92*, 839-865. https://doi.org/10.1016/j.ejmech.2015.01.051

Pranto, B., Mehnaz, S., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information, 11*(8), 374. https://doi.org/10.3390/info11080374

Radja, M., & Emanuel, A. W. R. (2019, October). Performance evaluation of supervised machine learning algorithms using different data set sizes for diabetes prediction. In *2019 5th international conference on science in information technology (ICSITech)* (pp. 252-258). IEEE. https://doi.org/10.1109/ICSITech46713.2019.8987479

National Lung Screening Trial Research Team. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, *365*(5), 395-409. https://www.nejm.org/doi/10.1056/NEJMoa1102873.

Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017, October). Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCNI.2017.8123815