International Journal of Applied Engineering and Technology ISSN: 2277-212X
An Open Access, Online International Journal Available at http://www.cibtech.org/jet.htm
2022, Vol.12, pp.89-94/ Infantsahayafuela et al.
Research Article

A DATA-DRIVEN APPROACH TO PREDICT DIABETES EARLY WITH MACHINE LEARNING ALGORITHMS

***S. Infantsahayafuela, S. Rajeswari and K. Yazhini**
Department of Information Technology,
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Thiruvarur
**Author for Correspondence*

ABSTRACT
Diabetes is a common, long-term disease. Diabetes prediction at an early stage can lead to better treatment. Data mining techniques are widely used for early disease prediction. Diabetes is predicted using significant attributes in this research paper, and the relationship between the various attributes is also described. For diabetes, various tools are used to determine significant attribute selection, as well as clustering, prediction, and association rule mining. The principal component analysis method was used to select significant attributes. Our findings show a strong link between diabetes and body mass index (BMI) and glucose. The gradient boosting classifier algorithm was used to predict diabetes.

Keywords*: Diabetes, Machine Learning, Supervised Learning Algorithms, Prediction Accuracy*

INTRODUCTION

Diabetes is one of the most dangerous diseases in the world. Diabetes is caused by obesity, high blood glucose levels, and so on. It affects the hormone insulin, causing abnormal crab metabolism and improving blood sugar levels. Diabetes develops when the body does not produce enough insulin. According to the World Health Organization, approximately 422 million people in low or middle-income countries suffer from diabetes. This figure could be increased to 490 billion by 2030. Diabetes is prevalent in several countries, including Canada, China, and India. With India's population now exceeding 100 million, the actual number of diabetics in the country is 40 million. Diabetes is a leading cause of death worldwide. Diabetes, for example, can be controlled and saved by early detection. To that end, this work investigates diabetes prediction using various diabetes disease-related attributes. We use the Pima Indian Diabetes Dataset for this purpose, and we use various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning is a method for explicitly training computers or machines. Various Machine Learning Techniques provide efficient results for knowledge collection by constructing various classification and ensemble models from collected datasets. This type of data can be used to predict diabetes. Various Machine learning techniques are capable of prediction, but selecting the best technique is difficult.As a result, we use popular classification and ensemble methods on datasets for prediction.

PROBLEM STATEMENT

Diagnosis of diabetes is one of the most challenging and considerable duties in medicine. Several traits need to be acquired to forecast the illness, such as plasma glucose concentration, diastolic blood pressure, triceps, pores and skin fold thickness, serum insulin, physique mass, and age, which can also take time to analyse and make the ultimate judgement. The project's intention is to discover and enforce diabetic conditions the use of computer getting to know algorithms. To enhance diabetes prediction, a computer studying method used to be developed here.

EXISTING SYSTEM

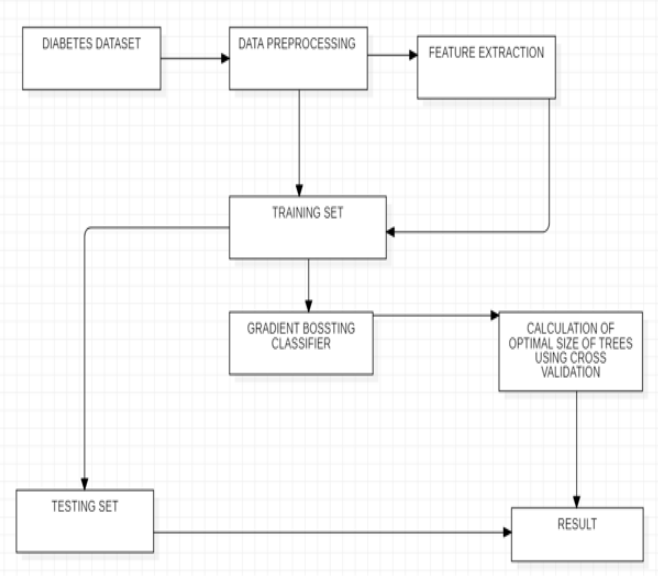
Many studies have been undertaken to identify diabetes. Data was investigated in an existing system using mining approaches such as clustering and classification. Diabetes prediction techniques such as k-NN, k-means, and the branch and bound algorithm were suggested. For the comparison study, a basic

diabetes dataset is used. The significance of feature analysis for diabetes prediction using a machine learning approach is explored.The accuracy of classification and prediction is not very high.

PROPOSED SYSTEM

The proposed system study is a binary classification and regression problem classification of the kaggle dataset for diabetes. This is proposed to be accomplished using machine learning and a classification algorithm. The gradient boosting classifier algorithm is proposed for machine learning. Through machine learning techniques, the proposed system improves prediction accuracy.

SYSTEM ARCHITECTURE DIAGRAM



H/W SYSTEM CONFIGURATION

Processor - Pentium – IV
RAM - 4 GB (min)
Hard Disk - 20 GB

S/W SYSTEM CONFIGURATION

Operating System: Windows 7 or 8
Software : python Idle

GRADIENT BOOSTING CLASSIFIER

Gradient Boosting is the most powerful ensemble technique for prediction and classification. It combines week learner models to create powerful learner models for prediction. It employs the Decision Tree model. It is a very effective and popular method for classifying complex data sets. The performance of the gradient boosting model improves over iterations.

MATERIALS AND METHODS

DATA COLLECTION

Diabetes is one of the most dangerous chronic diseases, affecting 422 million people worldwide, according to a WHO report from 2018. The dataset contains 520 people's records of diabetes-related symptoms. It includes records of people, as well as symptoms that may indicate diabetes. The dataset includes 17 attributes that include information about diabetes symptoms and patient conditions.

DATA PREPROCESSING

The most crucial process is data preparation. Most healthcare data has missing values and other contaminants that might reduce data effectiveness. Data preprocessing is performed to increase the quality

and effectiveness of the results acquired from the mining process. The method is critical for accurate results and good prediction when using Machine Learning Techniques on a dataset.

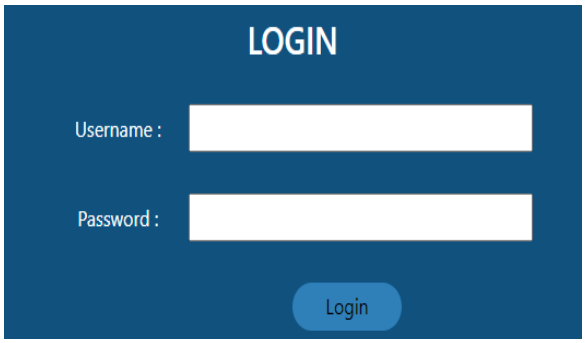
FEATURE EXTRACTION

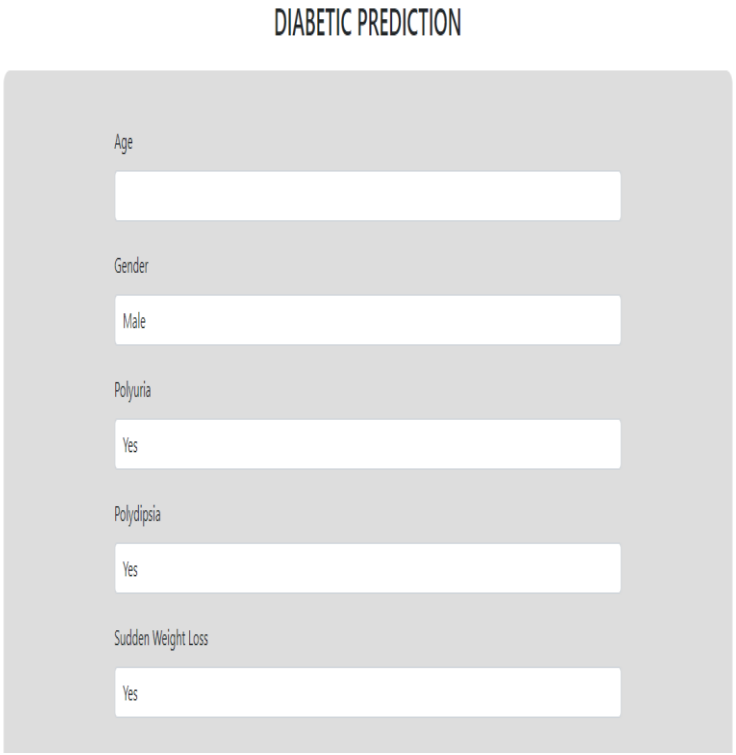
The feature extraction algorithm is used to extract the patient feature values individual feature condition of the patient. The attribute-based feature extraction techniques extracted the attribute-based feature extraction algorithm.

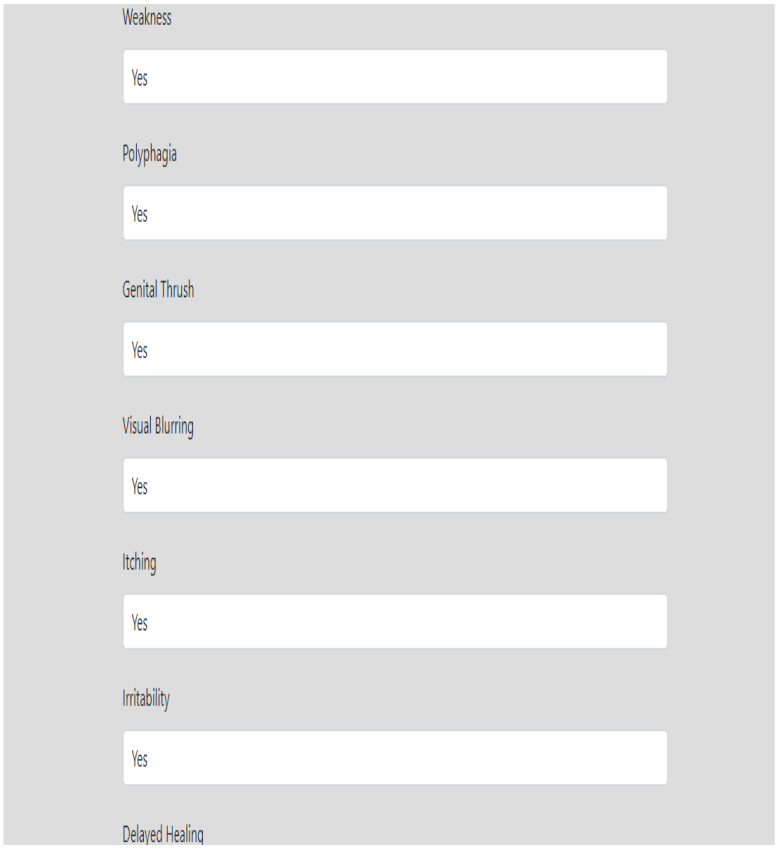
CLASSIFICATION

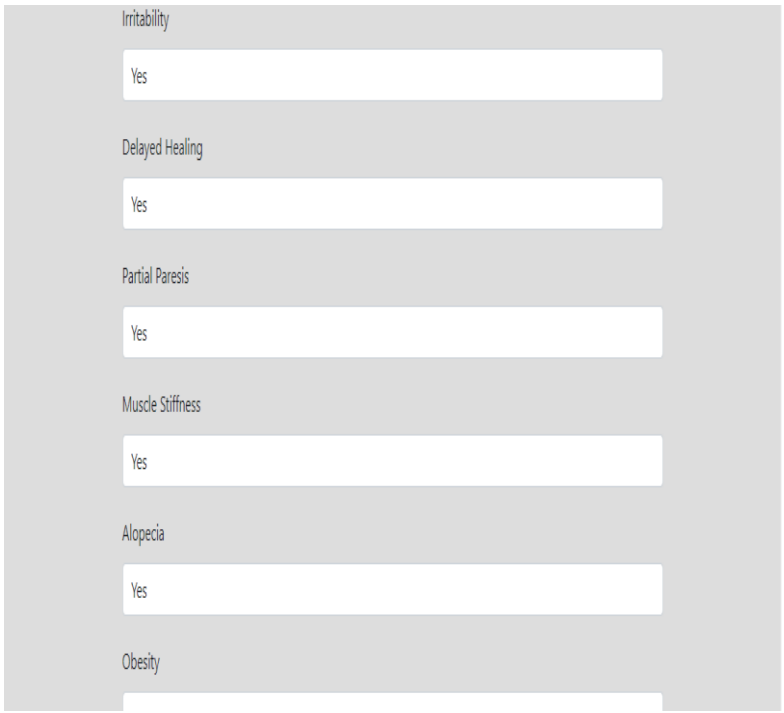
After training the data set on feature values, we performed classification based on dataset preprocessing. Gradient boosting classifier algorithms were used after the data preprocessing and selection processes were completed. Finally, we calculated the system's prediction accuracy.

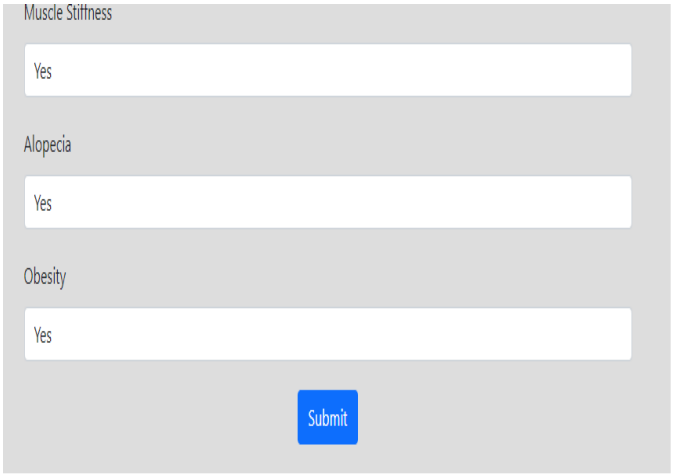
SCREEN SHOTS






Weakness
Yes
Polyphagia
Yes
Genital Thrush
Yes
Visual Blurring
Yes
Itching
Yes
Irritability
Yes
Delayed Healing


Irritability
Yes
Delayed Healing
Yes
Partial Paresis
Yes
Muscle Stiffness
Yes
Alopecia
Yes
Obesity



RESULTS AND DISCUSSION

Several stages were done in this work. The suggested approach employs various classification and ensemble algorithms and is written in Python. These are common Machine Learning approaches for obtaining the highest accuracy from data. In this paper, we see that the gradient boost classifier outperforms the others. Furthermore, to attain high performance prediction.

CONCLUSION

This project's main goal was to design and implement Diabetes Prediction Using Machine Learning Methods, and it was a success. The proposed technique includes various classification and ensemble learning methods, including the Gradient Boosting classifier algorithm. In addition, 77 percent classification accuracy was achieved. The experimental results can help health care providers make early predictions and decisions to cure diabetes and save people's lives.

REFERENCES

[1] D. Falco and B. E. Holland, Medical and Psychosocial Aspects of Chronic Illness and Disability. Burlington, MA, USA: Jones & Bartlett Learning, 2017.
[2] G. Klöppel, M. Löhr, K. Hibachi, M. Oberholzer, and P. U. Heitz, ``Islet pathology and the pathogenesis of type 1 and type 2 diabetes mellitus revisited, 1985.
[3] International Diabetes Federation Facts & Figures. Accessed: Dec. 24, 2020.
[4] C. S. Dangare and S. S. Apte, ``A data mining approach for prediction of heart disease using neural networks,2018.
[5] S. Smiley. (Jan. 12, 2020). Diagnostic for Heart Disease with Machine Learning.Medium.Accessed: Sep. 19, 2020.
[6] R. E. Wright, ``Logistic regression," in Reading and Understanding Multivariate Statistics. Washington, DC, US: American Psychological Association, 1995.
[7] An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression: The American Statistician. Accessed: Sep. 6, 2020.
[8] K. M. Ting and Z. Zheng, ``Improving the performance of boosting for naive Bayesian classification," in Proc. Methodol. Knowl. Discovery Data Mining, Berlin, Germany, 1999.
[9] N. V. Vapnik, Statistical Learning Theory. Hoboken, NJ, USA: Wiley, Sep. 1998. Accessed Sep. 6, 2020.
[10] J. R. Quinlan, ``Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81106, Mar. 1986.
[11] L. Bremen, ``Random forests," Mach. Learn., vol. 45, no. 1, pp. 532, Oct. 2001.

[12] T. Zheng, W. Xie, L. Xu, X. He, Y.Zhang, M. You, G. Yang, and Y. Chen, ``A machine learning-based framework to identify type 2 diabetes through electronic health records," Int. J. Med. Informal., vol. 97, pp. 120127, Jan. 2017.
[13] D. Sisodia and D. S. Sisodia, ``Prediction of diabetes using classification algorithms," ProcediaCompute. Sci., vol. 132, pp. 15781585, Jan. 2018.