Prediction of Diabetes at Early Stage Using Machine Learning Algorithms

Jieping Luo[1, *]

[1]Department of Mathematics, UCLA, Los Angeles, 90024, United States

* Corresponding author: jpluo@g.ucla.edu

Abstract. Diabetes is a chronic illness caused by a lack of insulin produced in the pancreas, which leads to an elevated blood glucose level. Diabetes leads to complications, such as loss of vision and kidney function, heart attacks, and strokes, which could ultimately shorten life. Diabetes has many different symptoms, for example, obesity, polydipsia, and itching. This study aimed to use these symptoms to predict diabetes using machine learning algorithms and to find the best classifier. The result is that random forest had the highest accuracy of 97% on this dataset.

Keywords: Diabetes; early stage; machine learning.

1. Introduction

An article published by WHO in 2022 stated that the number of deaths before the age of 70 caused by diabetes increased by 5% between 2000 and 2016. Of the deaths in 2019, one and a half million people died directly from diabetes, making it one of the deadliest conditions in the world [1,2]. Unlike other diseases, the early signs of diabetes are very subtle, making it difficult to detect, and many patients miss the best treatment period. About half of people with diabetes are asymptomatic and undiagnosed [3,4]. Therefore, detecting diabetes at the earliest stage is crucial to its treatment and prevention. The study used machine learning algorithms, which are widely used in predictive analytics since they make predictions more accurate.

The dataset was obtained from UCI Machine Learning Repository, containing 520 instances from Sylhet Diabetes Hospital. In order to better detect various symptoms and enable patients to receive timely treatment, it is vital to develop a deeper understanding of the various symptoms, so a heatmap was created to understand the correlation of various symptoms with diabetes. In this dataset, the symptoms most associated with diabetes were polyuria and polydipsia. Then eXtreme Gradient Boosting (XG Boost), decision trees, logistic regression, random forest, and K-Nearest Neighbors (KNN) were used to predict diabetes.

2. Related Works

There are many studies on diabetes, for example, studies of symptoms with specific types of diabetes, prediction of diabetes development in different regions, and studies of diabetes-related complications. This study is based on diabetes-related symptoms to predict diabetes at an early stage, and there are relatively few similar studies. Unlike similar studies that mostly use Bayesian, random forest, and logistic regression algorithms, the algorithms in this study include but are not limited to XG boost and KNN, which have higher accuracy.

3. Method

3.1 Data

3.1.1 Data Collection

This dataset was obtained from the UCI Machine Learning Repository, and it was collected directly from patient questionnaires at the Sylhet Diabetes Hospital, Sylhet, Bangladesh. This dataset contains 520 instances of people who newly have diabetes or will become diabetic patients, as well as 17 attributes, which are signs and symptoms data of these people, such as age, gender, and presence of symptoms, for example, obesity, polydipsia, irritability, or itching.

In this dataset, the data in the age column are numbers, the data in the gender column are 'Male' or 'Female,' and the data in the class column are 'Positive' or 'Negative,' where 'Positive' means the person is diagnosed with diabetes and 'Negative' means the person is not diagnosed with diabetes. The person may develop diabetes, or the person is healthy. In the other columns, the data are 'Yes' or 'No,' respectively, representing whether the person has the symptom.

3.1.2 Data Cleaning

Read data and get the data information, such as the column name, the number of null values, and the data type, then drop all the null values. Keep the Age column unchanged, then replace 'Male' with 0 and 'Female' with 1 in the Gender column, replace 'Positive' with 0 and 'Negative' with 1 in the Class column and replace 'Yes' with 0 and 'No' with 1 in the other columns. Then the data became comprehensive, well-organized, and simple to utilize.

3.2 Method and Algorithms

A heatmap was created to know the correlation between symptoms and diabetes, with each cell describing the relationship between the intersecting variables. The closer the number is to 1, the higher the correlation between the two intersecting variables.

In order to forecast diabetes at the early stage, the clean data was divided into test and training data with a test size of 0.20, and analyze the data using algorithms, XG Boost, decision tree, logistic regression, random forest, and KNN, and then calculate their accuracy. Because the accuracy scores of some algorithms were very close, ROC curves were created to better differentiate the models.

3.2.1 XG Boost (Xtreme Gradient Boosting)

It is a machine learning algorithm built on the framework of GBDT (Gradient Boosting Decision Tree) and is commonly used in regression and classification problems. It has the advantages of scalability and fast execution, and is the most useful, straightforward and powerful algorithm.

The bjective function of XG Boost is

$$obj(\theta) = \sum_{i}^{n} l(y_i - \hat{y}_i) + \sum_{j=1}^{j} \Omega(f_j) \tag{1}$$

where $l(y_i - \hat{y}_i)$ represents the loss function, $\Omega(f_j)$ represents the regularization function, y_i represents the i^{th} sample, and f_j represents the j^{th} tree [5].

3.2.2 Decision Tree

It is a non-parametric supervised learning algorithm that is commonly used in classification and regression problems. The principle of this algorithm is to infer decision rules from data characteristics, and then predict the target variable based on these rules. A decision tree can be seen as mimicking the human decision-making process. Its structure is close to a flowchart, in which internal nodes are tests of different attributes, branches are test results, and leaf nodes have class labels.

The formula of decision tree is:

$$G(x) = \sum_{c=1}^{c} [[b(x) = c]] G_c(x) \tag{2}$$

where G(x) represents the full-tree hypothesis, b(x) represents the branching criteria, and $G_c(x)$ is the sub-tree hypothesis at the c^{th} branch [6].

3.2.3 Logistic Regression

It is a machine learning algorithm based on statistical knowledge, similar to linear regression. It is often used to calculate the probability of an incident happening according to a given set of variables or to classify new data based on discrete datasets. The main advantage of this algorithm is that it can classify observations using different data types and easily determine the most influential variables used for the classification [7].

3.2.4 Random Forest

It is a Supervised Machine Learning Algorithm consisting of many decision trees built on different samples, and it is widely used for Classification and Regression problems. Its advantage is that it can process datasets containing continuous variables. The steps of the random forest algorithm are to take n random records from a dataset with k records; build individual decision trees for each sample; generate an output from each decision tree; obtain the final result according to the majority vote of classification or the average of regression [8].

3.2.5 KNN (K-Nearest Neighbors)

It is one of the simplest supervised machine learning algorithms and is widely used for classification and regression prediction problems due to its short computation time. This algorithm is based on the assumption that similar data will be clustered together. The steps of KNN algorithm are as follows: initialize the parameter K to the number of neighbors; measure the distance between all classified data and each new unclassified data; sort the distances from small to large to get the first K distances; get the class with the most occurrences in the first k distances, which is the predicted class [9].

4. Results

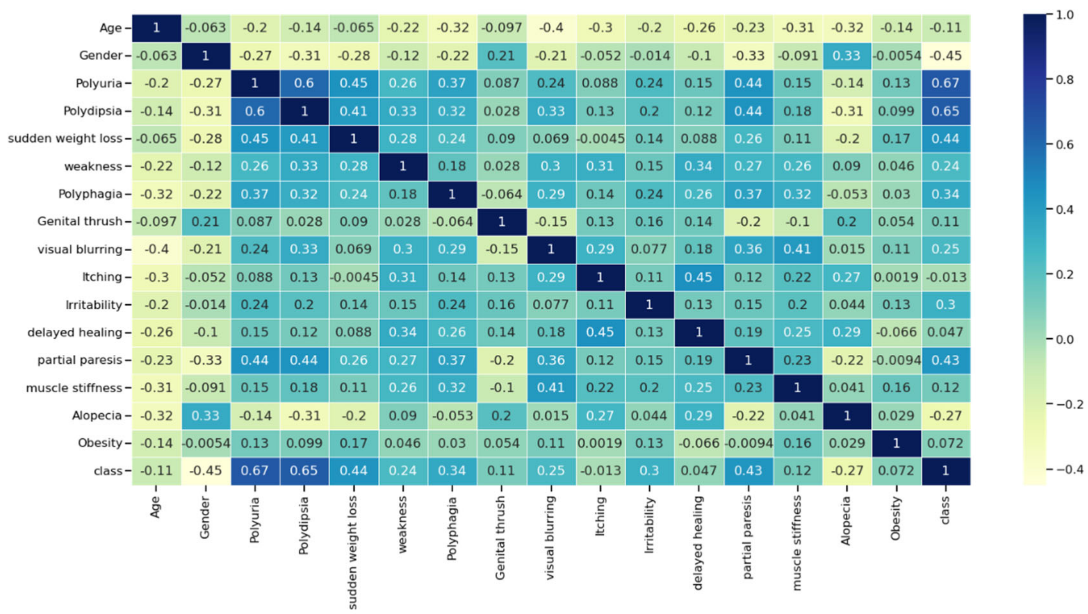

Fig. 1 Heatmap

Figure 1 is the heatmap used to represent the relationship between the intersecting variables. The last column is the intersection of class and other variables, representing the correlation between symptoms and diabetes. It can be seen from this figure that age, gender, itching, and alopecia have a negative linear relationship with diabetes since the corresponding values are negative, and the rest are positive linear relationships with diabetes.

Polyuria and polydipsia had the highest correlations with diabetes at values of 0.67 and 0.65, respectively. The lowest associations with diabetes were Itching, delayed healing, and Obesity, with values of -0.013, 0.047, and 0.072, respectively.

Table 1. Accuracy Scores of Algorithms

Models	Accuracy
Random Forest	0.971154
Logistic Regression	0.951923
XG Boost	0.942308
Decision Tree	0.942308
KNN	0.913462

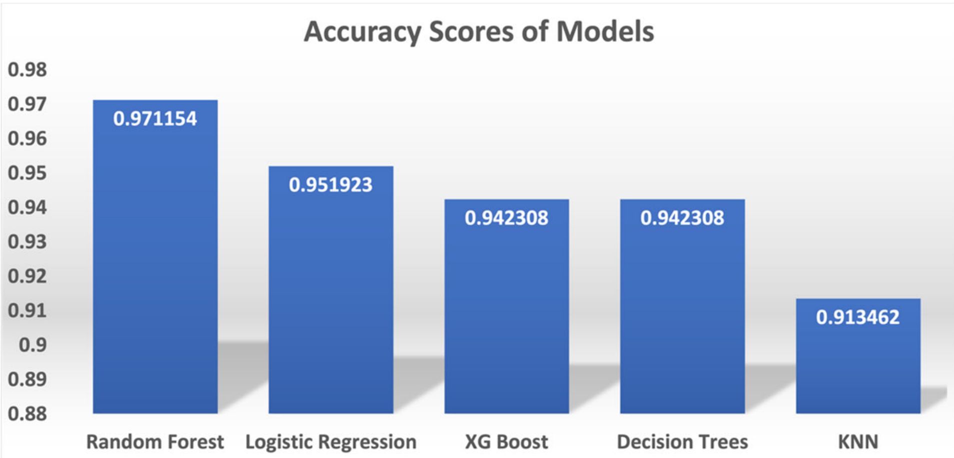

Fig. 2 Histogram of Accuracy Scores

Table 1 shows the accuracy scores for models and the figure 2 is a histogram of accuracy scores. Separate the clean data into training and testing data, with a test size of 0.20. Import data into models of XG Boost, decision tree, logistic regression, random forest, and KNN, then calculate their accuracy scores. According to figure 2, all models had an accuracy above 90%, while random forest had the greatest accuracy at 97%.

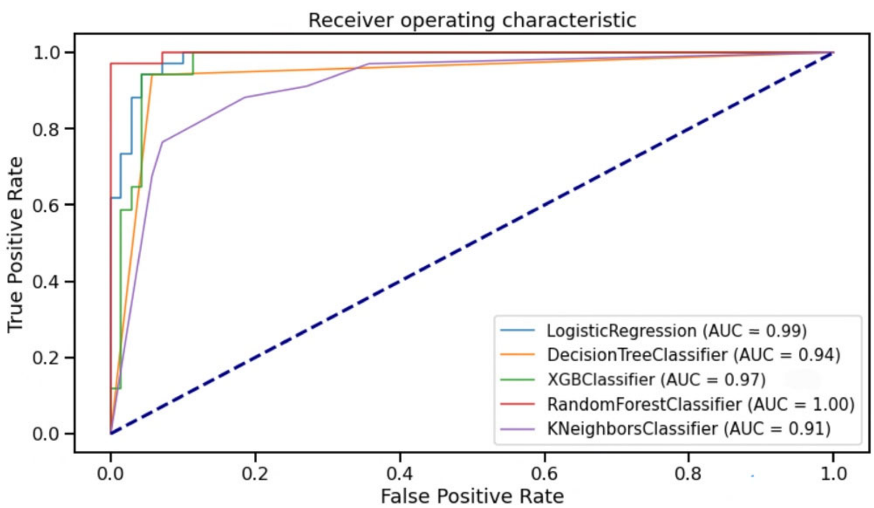

Fig. 3 ROC

Figure 3 shows the ROC curves, which provide a more intuitive visual representation of the accuracy of different models. Based on ROC curves and AUC values, random forest was the most accurate model, followed by logistic regression, XG Boost, decision tree, and KNN, and this was consistent with the results in Table 1.

5. Discussion

In the heatmap, the highest associations with diabetes were polyuria and polydipsia, and the lowest associations with diabetes were itching, delayed healing, and obesity.

As all we know, polyuria and polydipsia are highly correlated with diabetes since they are the typical symptoms of type 1 diabetes and are present in all age categories [10,11]. Itching is a common symptom of diabetes, but it can also be caused by skin conditions and lymphomas, so its causal relationship with diabetes is controversial [12].

Moreover, obesity and delayed healing are highly correlated with diabetes. Obesity increases the risk of developing diabetes, and the incidence of wound complications is higher in diabetic surgical patients than in other patients, especially type II diabetes [13, 14].

In the heatmap, obesity and delayed healing were significantly less correlated with diabetes than other symptoms. For models, the random forest had the highest accuracy score and was the best model, but the accuracy scores of XG Boost and decision tree were too close.

This dataset came from one hospital and only included 520 instances. The three major types of diabetes are type 1, type 2, and gestational diabetes, and different types of diabetes can cause different symptoms. This dataset did not distinguish the diabetes type and will likely affect the accuracy of the models. The accuracy of our prediction may be improved if we collect more data from different hospitals.

6. Conclusion

In this paper, using 520 instances collected from Sylhet Diabetes Hospital, I successfully concluded that polyuria and polydipsia had the highest correlation with diabetes in this data; I used XG Boost, decision tree, logistic regression, random forest, and KNN to analyze the dataset and finally found that random forest had the highest accuracy of 97% on this data.

In this paper, by using XG Boost, Decision Tree, Logistic Regression, Random Forest, and KNN to predict the dataset of 520 instances collected from Sylhet Diabetes Hospital, the study purposes were successfully achieved, and the following conclusions were obtained: polyuria and polydipsia had the highest correlation with diabetes, and random forest had the highest accuracy rate of 97% in this dataset.

According to the outcomes of this study, we may develop diabetes prediction software in the future. Once the user inputs their symptoms, the software will calculate the probability of the user having diabetes and make corresponding suggestions to the user to achieve the goal of preventing diabetes.

References

[1] Anjali D Deshpande, Marcie Harris-Hayes, Mario Schootman, Epidemiology of Diabetes and Diabetes-Related Complications, Physical Therapy, Volume 88, Issue 11, 1 November 2008, Pages 1254–1264, https://doi.org/10.2522/ptj.20080020.

[2] "Diabetes." World Health Organization, 16 Sep, 2022, https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] "What is diabetes?" cdc.gov, https://www.cdc.gov/diabetes/basics/diabetes.html. Ac- cessed 14 August 2022.

[4] Factors, Risk. "Early Warning: A Look at the First Signs of Diabetes." TrustCare, 27 Sep, 2021, https://trustcarehealth.com/blog/early-warning-a-look-at-the-first-signs-of-diabetes.

[5] Latha, Asha. “XGBoost as Regressor and Classifier.” Inblog, 20 Jan, 2021, blog.ineuron.ai/XGBoost- as-Regressor-and-Classifier-4IVqLERTn8.

[6] ”9–DecisionTree.”https://wizardforcel.gitbooks.io/ntu-hsuantienlin-ml/content/27.html. Accessed 11 September 2022.

[7] ”LogisticRegressioninMachineLearning.”https://www.javatpoint.com/logistic-regression- in-machine-learning. Accessed 11 September 2022.

[8] R, Sruthi. ”Understanding Random Forest.” Analytics Vidhya, 21 June, 2022, https://www.analyticsvidhya.com random-forest/.

[9] Harrison, Onel. ”Machine Learning Basics with the K-Nearest Neighbors Algorithm.” To- wards Data Science, 10 September, 2022, https://towardsdatascience.com/machine-learning- basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[10] Roche, Edna Frances, et al. “Clinical Presentation of Type 1 Diabetes.” Pediatric Diabetes, vol. 6, no. 2, 2005, pp. 75–78., https://doi.org/10.1111/j.1399-543x.2005.00110.x.

[11] Catherine C. Cowie, Keith F. Rust, Earl S. Ford, Mark S. Eberhardt, Danita D. Byrd- Holt, Chaoyang Li, Desmond E. Williams, Edward W. Gregg, Kathleen E. Bainbridge, Sharon H. Saydah, Linda S. Geiss; Full Accounting of Diabetes and Pre-Diabetes in the U.S. Population in 1988–1994 and 2005–2006. Diabetes Care 1 February 2009; 32 (2): 287–294. https://doi.org/10.2337/dc08-1296.

[12] Babakinejad, Parastoo, and Shernaz Walton. “Diabetes and Pruritus.” British Journal of Diabetes, vol. 16, no. 4, 2016, p. 154., https://doi.org/10.15277/bjd.2016.095.

[13] Hjart˚aker, A., Langseth, H., Weiderpass, E. (2008). Obesity and Diabetes Epidemics:. In: Berstein, L.M., Santen, R.J. (eds) Innovative Endocrinology of Cancer. Advances in Exper- imental Medicine and Biology, vol 630. Springer, New York, NY. https://doi.org/10.1007/978- 0-387-78818-0 6.

[14] Babakinejad, Parastoo, and Shernaz Walton. “Diabetes and Pruritus.” British Journal of Diabetes, vol. 16, no. 4, 2016, p. 154., https://doi.org/10.15277/bjd.2016.095.