JOIN (Jurnal Online Informatika)
Volume 7 No. 2 | December 2022: 244-250
DOI: 10.15575/join.v7i2.970
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)

Kartina Diah Kusuma W[1], Memen Akbar[2]
[1]Teknik Informatika, Politeknik Caltex Riau, Indonesia
[2]Teknik Rekayasa Komputer, Politeknik Caltex Riau, Indonesia

Article Info

Article history:

Received November 05, 2022
Revised November 24, 2022
Accepted November 24, 2022
Published December 26, 2022

Keywords:

Diabetes
Prediction
Machine Learning
XGBoost

ABSTRACT

One of the uses of medical data from diabetes patients is to produce models that can be used by medical personnel to predict and identify diabetes in patients. Various techniques are used to be able to provide a diabetes model as early as possible based on the symptoms experienced by diabetic patients, including using machine learning. The machine learning technique used to predict diabetes in this study is extreme gradient boosting (XGBoost). XGBoost is an advanced implementation of gradient boosting along with multiple regularization factors to accurately predict target variables by combining simpler and weaker model set estimations. Errors made by the previous model are tried to be corrected by the next model by adding some weight to the model. The diabetes prediction model using XGBoost is shown in the form of a tree, with the accuracy of the model produced in this study of 98.71%.

Corresponding Author:

Kartina Diah Kusuma W,
Jurusan Teknologi Informasi, Teknik Informatika, Politeknik Caltex Riau, Indonesia
Jl. Umbansari No.1. Pekanbaru – Riau – Indonesia. 28265
Email: diah@pcr.ac.id

1. INTRODUCTION

Diabetes is a degenerative disease. Diabetes caused by impaired function of the pancreas as a producer of the hormones insulin and glucagon which functions to regulate glucose levels in the blood. In the journal [1] it is stated that Diabetes is a disease that causes many other complications in the body such as cardiovascular and kidney, retinopathy, damage to the nervous system and others. Based on data from the American Diabetes Association, there are more than 387 million people with diabetes in the age range of 20 to 79 years, and there are still 46% who have not been identified [2]. Diabetes is divided into 3 types, including gestational diabetes, type-1 or juvenile diabetes and type-2 diabetes [3][4]. Several factors causing diabetes in research [5] include Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, delayed healing, Partial paresis, Muscle rigidity, Alopecia, and Obesity. People with diabetes that is left untreated and undetected can damage vital organs including the eyes, kidneys, nerves, heart, and feet as well as cause death[6]. In the study [7] it is stated that several environmental factors that have a risk of diabetes include: an environment that can be passed on foot, the availability of resources for physical activity around the environment, the food environment starting from the production phase to being consumed by the community in the environment. , availability of green space, residential noise level, traffic, and proximity to roads, air pollution, as well as environmental conditions, safety, and other environmental characteristics. Early prediction of diabetes can make it easier for doctors and patients to intervene as soon as possible so that the risk of complications can be reduced [8].

The development of technology in the world of health is in line with the growth of medical data that can be utilized in such a way for the benefit of mankind. One of the uses of medical data from diabetic patients is to produce a model that can be used by medical staff to predict and identify diabetes in patients. One of the branches of Machine Learning is Supervised Learning. One of its functions is to produce models from historical data to classify or predict. This study implements one of the supervised learning algorithms, namely extreme gradient boosting (XGBoost) to make predictions on diabetic patient data.

Many studies have been conducted to predict diabetes using various machine learning algorithms, including by [1], [3], [8]–[12]. Research by [3] predicts diabetes using several machine learning algorithms

such as Naïve Bayes, Random Forest, Support Vector Machine, and Multilayer Perceptron. This study shows that the Random Forest algorithm has the best accuracy value compared to others. Research by [10] conducted a classification for Gestational Diabetes using a modification of the Fuzzy C-Means Algorithm, K-Means Algorithm and Naïve Bayes MFCM Algortima. In this study, it was found that the Naïve Bayes MFCM Algorithm has the highest accuracy value compared to other algorithms. Research by [9] predicts diabetes using several supervised learning algorithms, including Logistic Regression, Neural Network, Random Forest, kNN, Tree, SVM, Naïve Bayes, and AdaBoost. Research by [8] predicts early risk for diabetes using the GA-Stacking Algorithm. This study shows a comparison of the performance of several supervised learning algorithms with GA-Stacking. The implementation of GA-Stacking to predict the early risk of diabetes showed better performance on accuracy, precision, sensitivity, specificity and F1-score. Research [11] uses the K-Nearest Neighbor Algorithm to detect diabetes mellitus. Research by [1] detects diabetes using the CNN (Convolutional Neural Network) Algorithm. This study creates a model that can be used to predict diabetes using the XGBoost Algorithm.

2. METHOD

The stages of the research will be carried out according to the chart at figure 1:

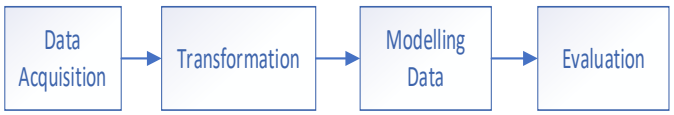

Figure 1. Research Stages

Details of the stages of the research to be carried out are as follows:

1) Data Acquisition
2) The diabetes data used will be crawled from the web which provides specific health data for diabetes. The dataset used in this study was taken from the Early-stage diabetes risk prediction published by UCI Machine Learning, totaling 520 records and 16 attributes and 1 class label
3) Transformation
   Transformation is purpose to change the categorical value of some attributes into numerical value
4) Data modeling
   Data modeling is done using a supervised learning algorithm, namely XGBoost.
5) XGBoost was first developed by Tianqi Chen in 2016. The XGBoost algorithm is a development of the GBDT (Gradient Boosting Decision Tree) algorithm which was previously discovered by Friedman. XGBoost is Supervised Learning that can be used to make predictions and classifications. XGBoost can also be applied to various disciplines such as education, health, government and others [13]. The computational stages performed on the XGBoost Algorithm are shown in Figure 2 [14].

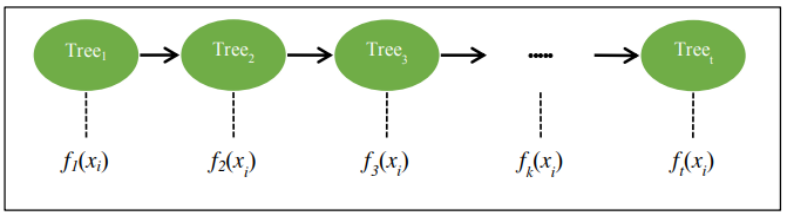

Figure 2. Diagram Skema XGBoost [14]

The predicted value at step t is likened to $\hat{y}_i^{(t)}$ with:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} fk(xi) \quad (1)$$

$f_k(x_i)$ describe the tree model. For y_i obtained from the following calculation:

$$\hat{y}_i^{(0)} = 0 \quad (2)$$

$$\hat{y}_i^{(1)} = f_1(x_1) = \hat{y}_i^{(0)} + f_1(x_1) \quad (3)$$

$$\hat{y}_i^{(2)} = f_1(x_1) + f_2(x_2) = \hat{y}_i^{(1)} + f_2(x_2) \quad (4).$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (5)$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} fk(xi) \quad (6)$$

Description:

- $\hat{y}_i^{(t)}$ = Final tree model
- $\hat{y}_i^{(t-1)}$ = Pre-generated tree model
- $f_t(x_i)$ = New model built
- t = Total number of models from base tree models

XGBoost in the process requires several parameters as a reference, including [15]:

a. colsample_bytree is a parameter to select the number of sample columns to be used, the default is 1 which means the entire column. Parameters range from 0 to 1.
b. beta is a learning rate parameter that serves to prevent the model from being overfitted. Parameters range from 0 to 1.
c. gamma is a parameter to determine the pruning of nodes in the created tree. The larger the gamma, the more conservative the model built. Parameters range from 0 to infinity.
d. max_depth is a parameter to determine the depth of the tree to be built, default 6. Parameter range from 0 to infinity.
e. min_child_weight is a parameter to determine the minimum weight limit that is owned by a node. Parameters range from 0 to infinity.
f. subsample is a parameter to select many sample rows of data to be used, the default is 1 which means all rows of data. Range from 0 to 1.
g. objective is a parameter that serves to determine the purpose of the model built such as regression or classification.
h. eval_metric is a parameter to select the evaluation size used, there are many evaluation measures such as RMSLE, RMSE, MAE, MAPE, AUC and others.

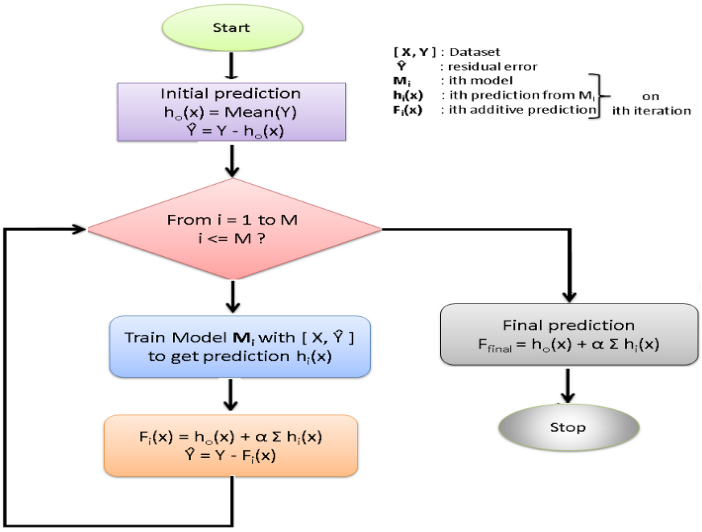

Figure 3. XGBoost Result

6) Evaluation/Testing.
The accuracy of the model will be done using a confusion matrix.

3. RESULTS AND DISCUSSION

1) Data Acquisition

The dataset used in this study was taken from the Early-stage diabetes risk prediction published by UCI Machine Learning, totaling 520 records and 16 attributes and 1 class label, as follows:

a) Age: Age in years ranging from (20 years to 90 years)
b) Gender: Male / Female
c) Polyuria: Yes / No
d) Polydipsia: Yes/ No
e) Sudden weight loss: Yes/ No
f) Weakness: Yes/ No
g) Polyphagia: Yes/ No

h) Genital Thrush: Yes/ No
i) Visual blurring: Yes/ No
j) Itching: Yes/ No
k) Irritability: Yes/No
l) Delayed healing: Yes/ No
m) Partial Paresis: Yes/ No
n) Muscle stiffness: yes/ No
o) Alopecia: Yes/ No
p) Obesity: Yes/ No
Class: Positive / Negative

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
39	Female	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Positive
48	Female	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Positive
58	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Positive
32	Female	No	No	No	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	No	Negative
42	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Negative

rows × 17 columns

Figure 4. The dataset shows as table

2) Transformation

In this research, transformation is done to change the categorical value of some attributes into numerical value. Figure 4 is the result of the transformation that has been carried out.

	Age	Gender_encode	Polyuria_encode	Polydipsia_encode	sudden weight loss_encode	weakness_encode	Polyphagia_encode	Genital thrush_encode	visual blurring_encode	Itching_en
0	40	1	0	1	0	1	0	0	0	
1	58	1	0	0	0	1	0	0	1	
2	41	1	1	0	0	1	1	0	0	
3	45	1	0	0	1	1	1	1	0	
4	60	1	1	1	1	1	1	0	1	
...	...	...	...	...	...	...	...	...	...	
515	39	0	1	1	1	0	1	0	0	
516	48	0	1	1	1	1	1	0	0	
517	58	0	1	1	1	1	1	0	1	
518	32	0	0	0	0	1	0	0	1	
519	42	1	0	0	0	0	0	0	0	

520 rows × 17 columns

Figure 5. Dataset after transformation

The following are the details of the dataset after the transformation.

The following is the diabetes dataset used.

```
#   Column              Non-Null Count  Dtype
---  ------              --------------  -----
0   Age                 520 non-null    int64
1   Gender              520 non-null    object
2   Polyuria            520 non-null    object
3   Polydipsia          520 non-null    object
4   sudden weight loss  520 non-null    object
```

```
5   weakness            520 non-null    object
6   Polyphagia          520 non-null    object
7   Genital thrush      520 non-null    object
8   visual blurring     520 non-null    object
9   Itching             520 non-null    object
10  Irritability        520 non-null    object
11  delayed healing     520 non-null    object
12  partial paresis     520 non-null    object
13  muscle stiffness    520 non-null    object
14  Alopecia            520 non-null    object
15  Obesity             520 non-null    object
16  class               520 non-null    object
```

After the transformation, the correlation between attributes in the dataset is displayed using the heatmap () function. It can be seen at figure 6.

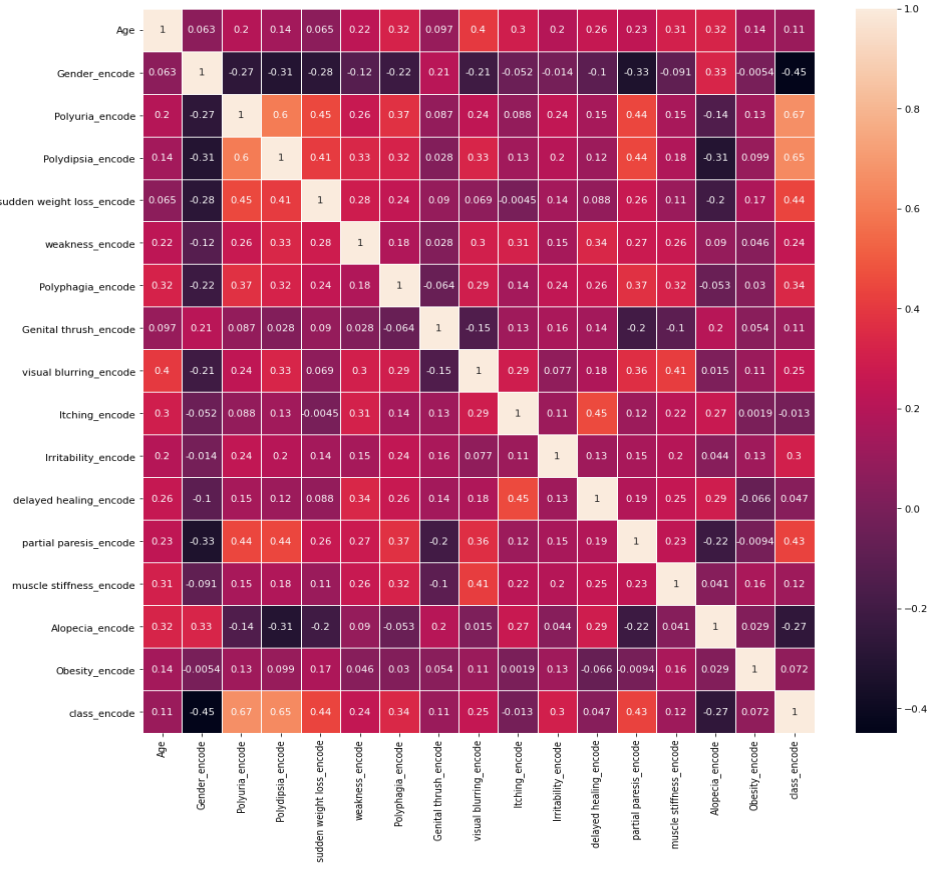

Figure 6.Correlation coefficient matrix

From the correlation coefficient matrix, we can see that polydipsia, polyuria and sudden weight loss are the most relative features to diabetes than others.

3) Modeling & Evaluation

Data modeling in this research using XGBoost displayed in the form of a tree as shown figure 7.

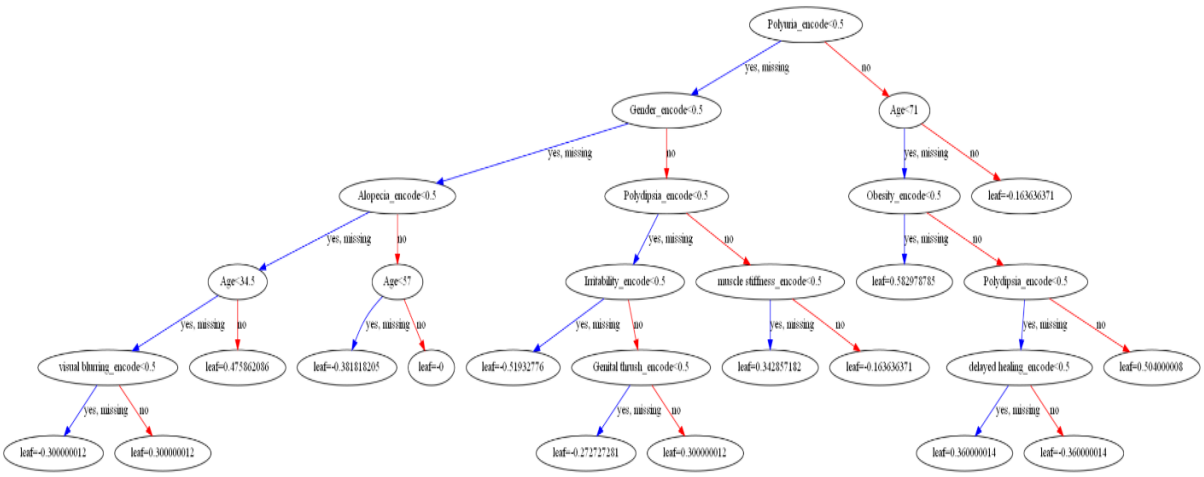

Figure 7. XGBoost Result

The evaluation model in this study using the confusion matrix is obtained as follows:

```
                        False Positive
True Positif    [ 54,    0],
False Negative [  2, 100]
                        True Negative
```

True Positive (TP) = 54 patients are predicted is positive diabetes, and they have positive diabetes
True Negative (TN) = 100 patients are predicted is negative diabetes, and they have negative diabetes
False Positive (FP) = no patients are predicted is positive diabetes, and they are
False Negative (FN) = 2 patients are negatively predicted for diabetes, but the fact is it turns out that the patient is positive for diabetes (predictions are incorrect)

$$\text{Accuracy} = \frac{54+100}{54+0+2+100} = 98.71\%$$

$$\text{Precision} = \frac{54}{54+0} = 100\%$$

$$\text{Recall} = \frac{54}{54+2} = 96.43\%$$

The model accuracy in doing the classification correctly is 98.71%.

4. CONCLUSION

From the research that has been done, it shows that Transformation data helps XGBoost modeling process more effectively. XGBoost algorithm can be implemented in diabetes dataset modeling to produce diabetes risk prediction with an accuracy of 98.71%. XGBoost has performed quite well for structured data.

5. REFERENCES

[1] V. Vaidya and L. K. Vishwamitra Scholar, "Diabetes Detection using Convolutional Neural Network through Feature Sequencing," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 2783–2789, 2021.
[2] A. D. Association, "Classification and diagnosis of diabetes," *Diabetes Care*, vol. 38 Suppl:S, 2015, doi: 10.2337/dc15-S005.
[3] S. Patel, R. Patel, N. Ganatra, and A. Patel, "Predicting a Risk of Diabetes at Early Stage using Machine Learning Approach," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 5277–5284, 2021.
[4] D. Elias and T. Maria, "Data-Driven Machine-Learning Methods for Diabetes Risk Prediction," *Sensors*, p. 5304, 2022, doi: https://doi.org/10.3390/s22145304.
[5] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," *Comput. Vis. Mach. Intell. Med. Image Anal. Springer,*

Singapore., pp. 113-125., 2020.
[6] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.
[7] T. Dendup, X. Feng, S. Clingan, and T. Astell-Burt, "Environmental risk factors for developing type 2 diabetes mellitus: A systematic review," *Int. J. Environ. Res. Public Health*, vol. 15, no. 1, 2018, doi: 10.3390/ijerph15010078.
[8] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early Risk Prediction of Diabetes Based on GA-Stacking," *Appl. Sci.*, vol. 12, no. 2, 2022, doi: 10.3390/app12020632.
[9] S. K. Bhoi *et al.*, "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 3074–3084, 2021.
[10] V. R. Geetha, N. Jayaveeran, and A. S. A. K. N, "Classification Of Gestational Diabetes Using Modified Fuzzy C Means Clustering And Machine Learning Technique," vol. 12, no. 10, pp. 2416–2427, 2021.
[11] R. Saxena and S. Kumar Sharma Manali Gupta, "Role of K-nearest neighbour in detection of Diabetes Mellitus," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 373–376, 2021.
[12] J. J. S. M. Et. al., "Predictive Modeling Framework for Diabetes Classification Using Big Data Tools and Machine Learning," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 818–823, 2021, doi: 10.17762/turcomat.v12i10.4255.
[13] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: 10.24843/jim.2020.v08.i03.p04.
[14] H. Mo, H. Sun, J. Liu, and S. Wei, "Developing window behavior models for residential buildings using XGBoost algorithm," *Energy Build.*, vol. 205, pp. 1–23, 2019, doi: 10.1016/j.enbuild.2019.109564.
[15] A. Mello, "XGBoost: theory and practice," *https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6*, 2020.