Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 4 (2023) 824 - 831 ISSN Media Electronic: 2580-0760

Diabetes Risk Prediction using Feature Importance Extreme Gradient Boosting (XGBoost)

Kartina Diah Kusuma W[1], Memen Akbar[2]
[1]Information Technology Department, Faculty of Informatics Engineering, Politeknik Caltex Riau
[2]Information Technology Department, Faculty of Computer Engineering Techniques, Politeknik Caltex Riau
[1]diah@pcr.ac.id, [2]memen@pcr.ac.id

Abstract

Diabetes results from impaired pancreas function as a producer of insulin and glucagon hormones, which regulate glucose levels in the blood. People with diabetes today are not only experienced adults, but pre-diabetes has been identified since the age of children and adolescents. 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. One of the uses of medical data from diabetes patients is used to produce a model that can be used by medical staff to predict and identify diabetes in patients. Various techniques are used to provide the earliest possible prediction of diabetes based on the symptoms experienced by diabetic patients, including using machine learning. People can use Machine Learning to generate models based on historical data of diabetic patients, and predictions are made with the model. In this study, extreme gradient boosting is the machine learning technique to predict diabetes (xgboost) using Feature Importance XGBoost. The diabetes dataset used in this study comes from the Early stage diabetes risk prediction dataset published by UCI Machine Learning, which has 520 records and 16 attributes. The diabetes prediction model using xgboost is displayed as a tree. The model accuracy result in this study was 98.71%, for the F1 score was 98.18%. While the accuracy obtained based on the best 10 attributes using the XGBoost feature importance are 98.72%.

Keywords: diabetes; prediction; machine learning; xgboost

1. Introduction

Common symptoms of diabetes include frequent urination, excessive thirst, unexplained weight loss, increased hunger, fatigue, blurred vision, slow wound healing, and recurrent infections. However, some people may have diabetes without experiencing noticeable symptoms, especially in the early stages. Patients with diabetes who are not detected and appropriately controlled can quickly damage their vital organs, such as the eyes, kidneys, heart, nerves, and feet, and even cause death [1]. 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 [2].

Diabetes is a degenerative disease. Diabetes is the result of impaired function of the pancreas organ as a producer of the hormones insulin and glucagon, which function to regulate glucose levels in the blood. In the journal [3], 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, more than 387 million people have diabetes in the age range of 20 to 79 years, and there are still 46% who have not been identified [4]. Diabetes is divided into three types, gestational diabetes, type-1 diabetes, and type-2 diabetes [5]. Some of the causes of diabetes in research [6] include Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital Thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Stiffness Musculoskeletal Alopecia, and Obesity. In research [7], they stated that several environmental factors that have a risk of diabetes include: an environment that can be traversed by foot, the availability of resources for physical activity around the environment, the food environment starting from the production phase to being consumed by the people in the environment, the availability of green spaces, residential noise levels, traffic, and proximity to roads, air pollution, and environmental conditions, safety, and other ecological characteristics.

Technological developments in the world of health are in line with the growth of medical data, which can be utilized in such a way for the benefit of humankind. One

Accepted: 07-11-2022 | Received in revised: 17-05-2023 | Published: 12-08-2023

of the uses of medical data from diabetes patients is to produce a model that can be used by medical staff to predict and identify diabetes in patients. Now days machine learning is one of the method used to analyze medical data.

Machine learning aims to produce specific patterns. This pattern is then implemented for various human purposes such as estimation, classification, prediction, clustering, forecasting, and association. Learning is carried out in 2 stages: training/exercise and testing/testing [8]. One of the branches of machine learning is supervised learning, and one of its functions is to produce models from historical data for classification or prediction [9], [10].

Many studies have been conducted to predict diabetes using various machine learning algorithms, including [2], [3], [5], [11]–[14]. Research by A. Patel et al. [5] predicts diabetes using several machine learning algorithms such as Nave Bayes, Random Forest, Support Vector Machine, and Multilayer Perceptron. This study shows that the Random Forest algorithm has the best accuracy value compared to the others. Research by Geetha [12] classified gestational diabetes using a modification of the fuzzy C-Means algorithm, the K-Means algorithm, and the MFCM Nave Bayes algorithm. In this study, the results showed that the Nave Bayes MFCM algorithm has the highest accuracy value compared to other algorithms. Research by S. K. Bhoi et al [11] made predictions for diabetes using several supervised learning algorithms, including logistic regression, neural networks, random forests, kNN, trees, SVM, Nave Bayes, and AdaBoost. Research by R. Saxena and S. Kumar Sharma Manali Gupta [13] uses the K-Nearest Neighbor Algorithm to detect diabetes mellitus. A study by V. Vaidya and L. K. Vishwamitra Scholar [3] detects diabetes using the CNN algorithm (convolutional neural network). Research by Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu [2] conducted early risk prediction for diabetes using the GA-Stacking Algorithm. This study compares the performance of several supervised learning algorithms with GA stacking. The implementation of GA-Stacking to predict the initial risk of diabetes shows better performance on the accuracy, precision, sensitivity, specificity, and F1-score.

Early risk prediction for diabetes has also been carried out in the author's previous study [15]. In previous study, modeling was carried out using the XGBoost algorithm with a model accuracy using a confusion matrix of 98.71%.

Current study implements machine learning intending to produce a model that can be used to predict the initial risk of diabetes using the XGBoost algorithm by performing feature selection using the XGBoost Feature Importance in order to increase the accuracy of model performance.

2. Research Methods

The stages of model development will be carried out according to Figure 1.

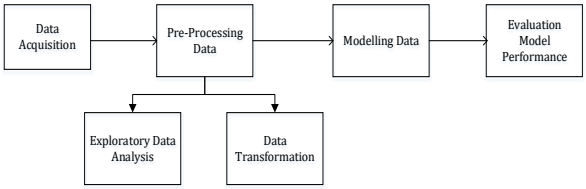

Figure 1. Model Development Stages

The details of the stages of the research to be carried out according to figure 3 are: Data Acquisition, Data Pre-Processing, Modeling Data and Evaluation Model Performance.

2.1 Data Acquisition and Pre-Processing Data

Data Acquisition for the diabetes data will be crawled from the web that provides specific health data for diabetes.

Pre-Processing Data contain 2 steps are: Exploratory Data Analysis and Transformation. Exploratory Data Analysis was first carried out. Exploratory data refers to the initial step in data analysis. Data analysts use data visualization and statistical techniques to represent characterizations of data sets, such as size, quantity, and accuracy, to better understand the nature of data [16]. Data exploration techniques include manual analysis and automated data exploration tools that visually explore and identify relationships between different data variables, the structure of data sets, and the distribution of data values to reveal patterns and enable data analysts to gain more significant insights into the raw data. In this study exploratory data analysis used to identification of typical data, shows correlations between attributes and label of data, descriptive analysis to make adjustments of variable type.

After Exploratory Data Analysis, next is Data Transformation. Data transformation is performed to adapt the data format to the modeling algorithm to be used. Many techniques can be used in data transformation such as binary transformation as machine learning algorithms cannot be directly applied to raw text [17].

2.2 Modeling Data

Data modeling is done using a supervised learning algorithm, namely XGBoost. The XGBoost algorithm is a development of the GBDT (Gradient Boosting Decision Tree) algorithm, which Friedman previously discovered. 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 [18]. The computational stages performed on the XGBoost Algorithm are shown in Figure 2 [18].

DOI: https://doi.org/10.29207/resti.v7i4.4651

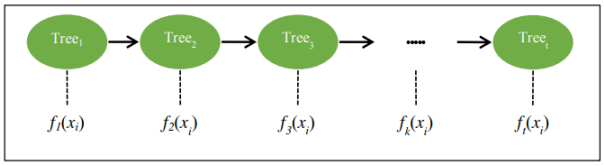

Figure 2. XGBoost Schematic Diagram (Mo et al., 2019)

The predicted value in step t is likened to $\hat{y}_i^{(t)}$ with Equation 1.

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

$f_k(x_i)$ describes the tree model. For yi obtained from Equation 2 until Equation 6 calculations:

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

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

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

.

.

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

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

$\hat{y}_i^{(t)}$ *is a* final tree model, $\hat{y}_i^{(t-1)}$ *is t*he previously generated tree model. $f_t(x_i)$ is a new model built and *t* is the total number of models from the base tree models.

In the process, XGBoost requires several parameters as a reference, including: [19]: colsample_bytree is a parameter to select the number of column samples to be used, default 1, which means the entire column. Parameter range from 0 to 1. Beta is a learning rate parameter that functions to prevent the model from experiencing overfitting. Parameter range from 0 to 1. Gamma is a parameter to determine the pruning of the nodes in the created tree. The bigger the gamma, the more conservative the model is built. Parameter range from 0 to infinity. Max_depth is a parameter to determine the depth of the tree to be made, default 6. Parameter range from 0 to infinity. Min_child_weight is a parameter to determine the minimum weight limit for a node. Parameter range from 0 to infinity. The subsample is a parameter to select the number of sample data rows to be used, the default one, which means the entire data row. Range from 0 to 1. The objective is a parameter that determines the purpose of the model built, such as regression or classification. Eval_metric is a parameter to select the evaluation size used. Many evaluation measures include RMSLE, RMSE, MAE, MAPE, AUC, and others.

The steps of the XGBoost algorithm as shown in figure 3 are: averaging the calculated target values for the initial predictions and the corresponding initial residual errors. Then a model (shallow decision tree) is trained with independent variables and residual errors as data to obtain predictions. After that, the additive prediction and residual error are calculated with several learning rates from the previous output predictions obtained from the model. Then repeat steps 2 and 3 for M several times until the required number of models is created. The final boosting prediction is the sum of all the previous predictions made by the model.

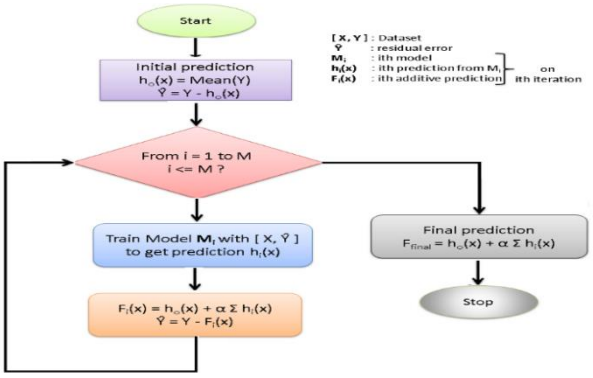

Figure 3. XGBoost Algorithm

2.3 Evaluation Model Performance

Last thing to do in this research is the Evaluation stage. Evaluation used to calculate model accuracy got from previous stage. Mode accuracy will be done using a confusion matrix. Matrix elements are characterized based on predicted labels (positive, negative) and the results of comparing predictions with actual class labels (true, false) [20]: True Positive (TP) is the total data with the exact number of positives and the number of optimistic predictions. True Negative (TN) is the real data with the actual number of positives and the number of negative predictions. False Positive (FP) is the amount of actual negative data and the amount of positive predictive data, and False Negative (FN) is the amount of actual negative data and the amount of negative predictive data. Equation 7 and 8 is the calculation of the f1-score and accuracy [9].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (8)$$

3. Results and Discussions

3.1 Data Acquisition

The dataset used in this study was taken from UCI Machine Learning which totaled 520 records and 16 attributes, and 1 class label. This data is a dataset published in the Early stage diabetes risk prediction UCI Machine Learning.

The dataset contains the sign and symptoms data of newly diabetic or would be diabetic patients. The detail information of the dataset can be found on the table 1.

Table 1 is the information from the diabetes dataset with

DOI: https://doi.org/10.29207/resti.v7i4.4651

the respective data types for each attribute. While table 2 is a representation of the diabetes dataset with values for each attribute. The dataset contains text data so it is necessary to transform the data into a numeric form.

Table 1. Diabetes Dataset Information

#	Column	Value	Non-Null Count	Dtype
1	Age	in years ranging form20 to 90 years	520 non-null	int64
2	Gender	Male/Female	521 non-null	object
3	Polyuria	Yes/No	522 non-null	object
4	Polydipsia	Yes/No	523 non-null	object
5	Sudden weight loss	Yes/No	524 non-null	object
6	Weakness	Yes/No	525 non-null	object
7	Polyphagia	Yes/No	526 non-null	object
8	Genital Thrush	Yes/No	527 non-null	object
9	Visual Blurring	Yes/No	528 non-null	object
10	Itching	Yes/No	529 non-null	object
11	Irritability	Yes/No	530 non-null	object
12	Delayed Healing	Yes/No	531 non-null	object
13	Partial Paresis	Yes/No	532 non-null	object
14	Muscle Stiffness	Yes/No	533 non-null	object
15	Alopecia	Yes/No	534 non-null	object
16	Obesity	Yes/No	535 non-null	object
17	Class	Positive/Negative	536 non-null	object

Table 2. Diabetes Dataset

Age	Gender	Polyura	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	Yes	Positive
45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	Yes	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

3.2 Data Exploration

The stages of data exploration in this study are used to analyze the representation of the data used. The following is an exploration of the data that has been visualized.

From Figure 4, it is known that the youngest age who has a risk of diabetes is 16 years, and the oldest is 90 years. The most at risk for diabetes are 35 years and 50 years.

The age distribution in the dataset ranges from 16 years to 90 years according to figure 5. The age range with most risk of diabetes is 40 years to 65 years.

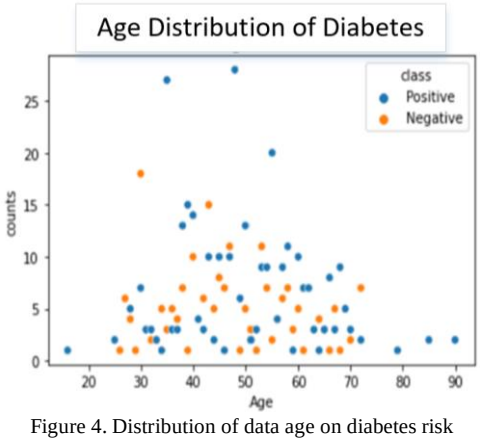

Figure 4. Distribution of data age on diabetes risk

DOI: https://doi.org/10.29207/resti.v7i4.4651

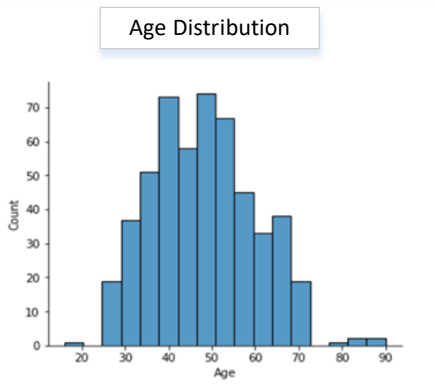

Figure 5. Age Distribution

It can be seen that in the figure 6, the number of females gender with a positive risk of diabetes is more than those who are not at risk of diabetes. Meanwhile, the number of male gender with positive and negative risks of diabetes appears to be balanced. Based on the dataset, it shows that women have a higher risk of diabetes than men.

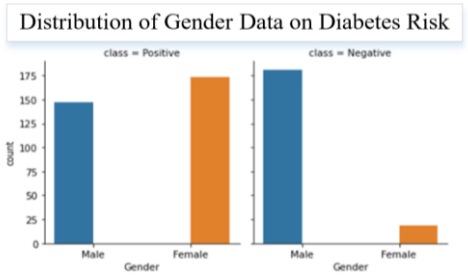

Figure 6. Distribution of Gender Data on Diabetes Risk

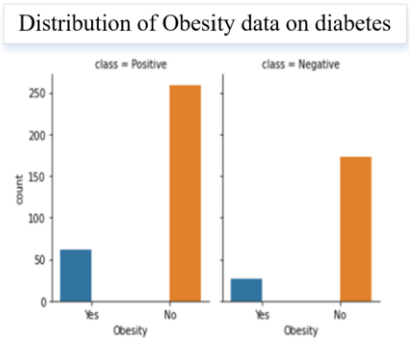

Figure 7. Distribution of Obesity data on diabetes risk

As shown in figure 7 people with obesity have a higher risk of diabetes then people who do not have obesity.

3.3 Transformation

In this study, the transformation was carried out to change the categorical values of several attributes into numerical values. Table 3 is the result of the transformation that has been done. Following are the details of the dataset after transformation.

Table 3. Diabetes Dataset Information After Transformation

#	Column	Non-Null Count	Dtype
1	Age	520 non-null	int64
2	Gender	521 non-null	int32
3	Polyuria	522 non-null	int32
4	Polydipsia	523 non-null	int32
5	Sudden weight loss	524 non-null	int32
6	Weakness	525 non-null	int32
7	Polyphagia	526 non-null	int32
8	Genital Thrush	527 non-null	int32
9	Visual Blurring	528 non-null	int32
10	Itching	529 non-null	int32
11	Irritability	530 non-null	int32
12	Delayed Healing	531 non-null	int32
13	Partial Paresis	532 non-null	int32
14	Muscle Stiffness	533 non-null	int32
15	Alopecia	534 non-null	int32
16	Obesity	535 non-null	int32
17	Class	536 non-null	int32

Table 4 is the dataset that has been transformed to fit into the XGBoost Algorithm in the next stage. After the transformation is performed, the correlation between the attributes in the dataset is displayed using the heatmap() function. Can be seen in Figure 8.

From the correlation coefficient matrix, we can see that the attribute Polydipsia, Polyuria and sudden weight lost are the attribute that have the most correlation with diabetes, with the correlation number for Polydipsia being 0.65, Polyuria 0.67 and sudden weight loss being 0.44.

3.4 Modelling & Evaluation

Data modeling is done using XGBoost displayed as a tree, as shown in Figure 9. XGBoost is an evolution of a tree-based algorithm. From Figure 9, it is known that the polydipsia attribute is a root. This is because Polydipsia is the factor that most influences the risk of diabetes in diabetic patients. Followed by gender and age. It is consistent with the results of the correlation analysis between attributes on diabetes risk in Figure 8 that Polydipsia is one of the attributes with the most correlation with diabetes, with a correlation number of 0.65.

In the previous study, the researcher obtained the model and accuracy for predicting pre-diabetes with the XGBoost algorithm without feature selection using Feature Importance. The resulting model accuracy is 98.71% and F1 score of 98.18% [15].

Current study, model testing was also carried out using the feature importance function of XGBoost by taking 10 of 15 attributes with the highest level of importance, as seen in Figure 10. Feature Importance improves the model and the resulting accuracy by selecting features. The 10 best attributes produced show the importance ranking of these attributes in the formed decision tree model.

DOI: https://doi.org/10.29207/resti.v7i4.4651

Table 4. Diabetes Dataset After Transformation

	Age	Gender_encode	Polyuria_encode	Polydipsia_encode	sudden weight loss_encode	weakness_encode	Polyphagia_encode	Genital thrush_encode	visual blurring_encode	Itching_encode	Irritability_encode	delayed healing_encode	partial paresis_encode	muscle stiffness_encode	Alopecia_encode	Obesity_encode	class_encode
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1
516	48	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1
517	58	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	1
518	32	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0
519	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

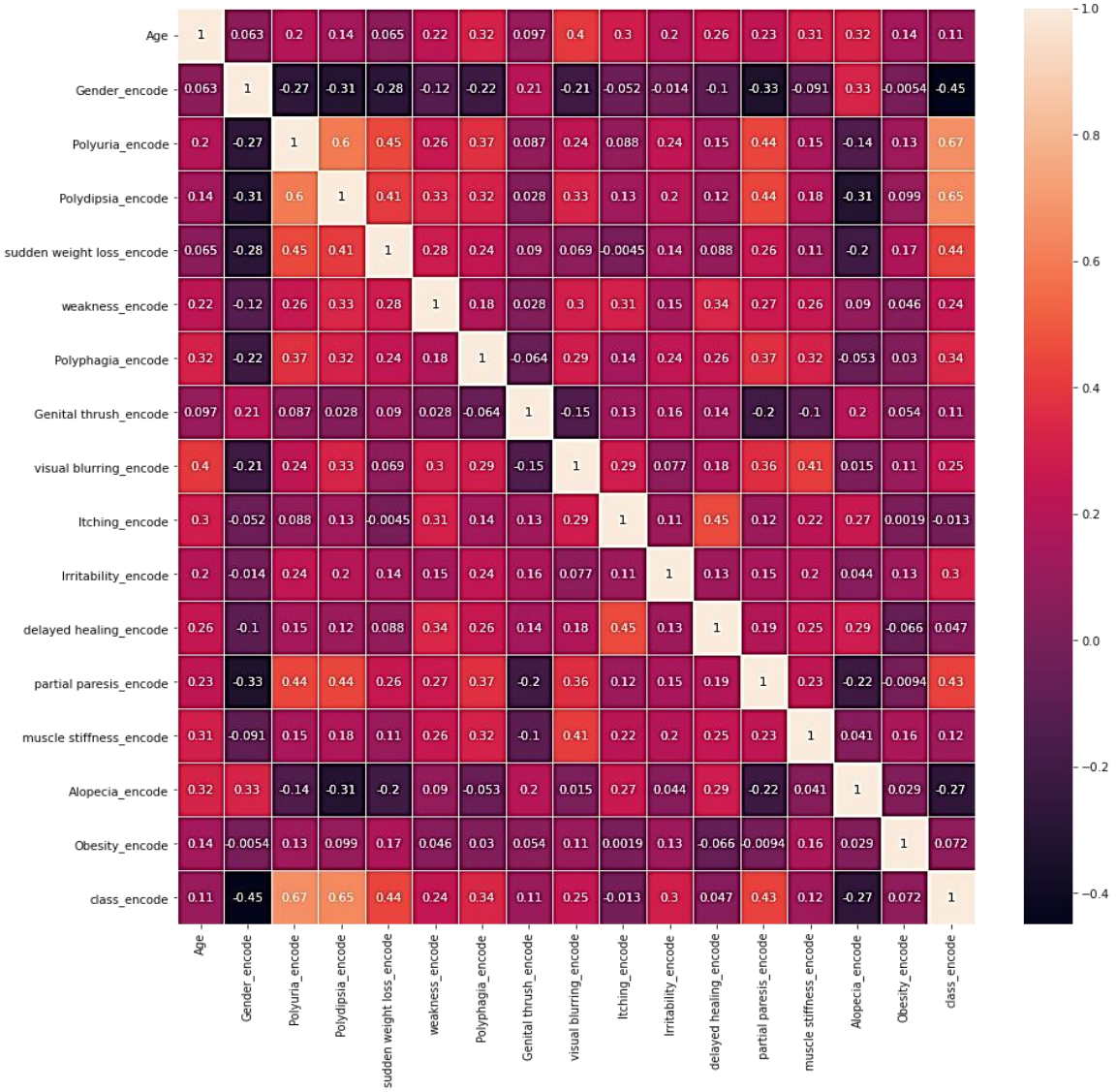

Figure 8. Attribute Correlation Matrix

The 10 important attributes of XGBoost features are Age, Alopecia, Gender, Polydipsia, Polyuria, Delayed Healing, Irritability, Muscle Stiffness, Visual Blurring, and Itching can be seen in Figure 10. While the other 5 features are considered less important, including: sudden weight loss, weakness, polyphagia, genital thrust, partial paresis, obesity.

DOI: https://doi.org/10.29207/resti.v7i4.4651

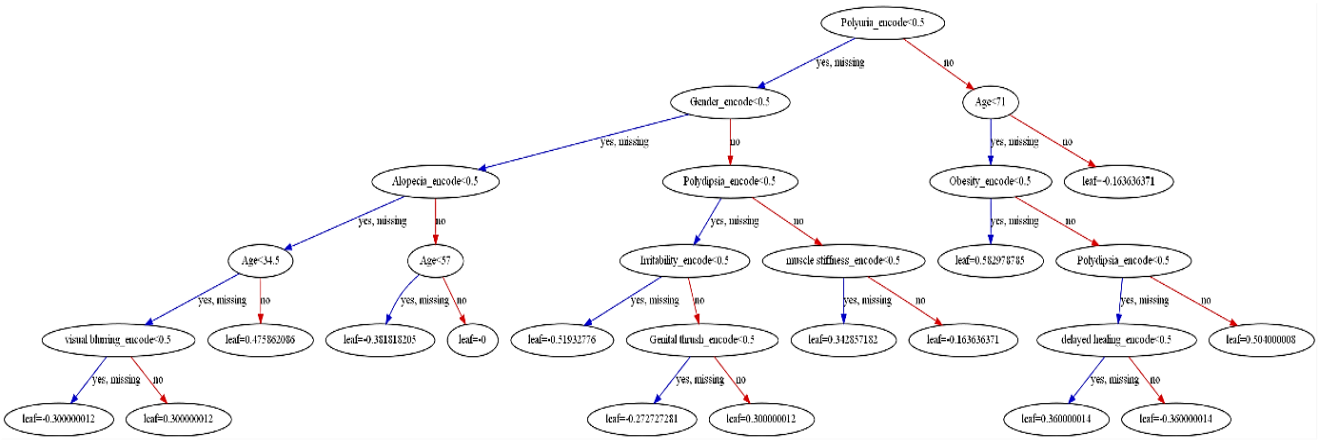

Figure 9. Modeling with the XGBoost Algorithm

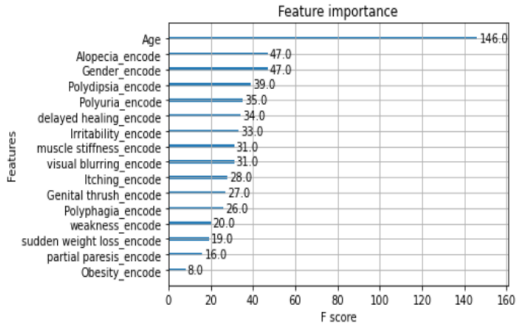

Figure 10. Feature Importance XGBoost

The evaluation model in this study uses a confusion matrix. For model evaluation with XGBoost:

False Positive
True Positif [55, 0],
False Negative [2, 100]
True Negative

While the accuracy results obtained are based on the 10 best attributes using the important feature of XGBoost according to Equation 9.

$$\frac{55+100}{55+2+100} = 98.72\% \quad (9)$$

And the value of the F1 Score according to Equation 10.

$$\frac{2 \text{ x } 96.42\% * 100\%}{96.42\% + 100\%} = 98.17\% \quad (10)$$

Evaluation result for accuration model is 98.72% and F-1 Score is 98.17%.

4. Conclusion

The research that has been done shows that the correlation between attributes from the correlation coefficient matrix is compatible with the diabetes risk factor model produced by XGBoost. The data transformation stage helps the XGBoost modeling process more effectively. The XGBoost algorithm implemented in the diabetes dataset modeling without feature selection using feature importance produces a risk prediction for diabetes with an accuracy of 98.71% and F1 score of 98.18%. Meanwhile, implementing the pre-diabetes risk prediction model with feature selection using the 10 best attributes from feature importance XGBoost increased the model's accuracy to 98.72%. Based on the results obtained, it can be concluded that using feature selection using feature importance for this study raises the model's accuracy by 0.01%.

References

[1] 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.

[2] 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.

[3] 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.

[4] A. D. Association, "Classification and diagnosis of diabetes," *Diabetes Care*, vol. 38 Su, 2015, doi: 10.2337/dc15-S005.

[5] 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.

[6] H. Y. Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," *Comput. Vis. Mach. Intell. Med. image Anal. (pp. 113-125)*, 2020, [Online]. Available: https://sreyas.ac.in/wp-content/uploads/2021/07/1.-Dr.-Rohit-Raja.pdf#page=119

[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] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006, doi: 10.1016/j.neucom.2005.12.126.

[9] Darussalam and G. Arief, "Sentiment Analysis on Social Media with Glove Using Combination CNN and RoBERTa," *J. Resti*, vol. 7 No.3, no. 1, pp. 457–463, 2023, doi: https://doi.org/10.29207/resti.v7iX3.4892.

DOI: https://doi.org/10.29207/resti.v7i4.4651

[10] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, no. December, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] K. D. K. Wardhani and M. Akbar, "Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)," *J. Online Inform. 7(2), 244-250.*, vol. Vol 7.No 2, 2022, doi: 10.15575/join.v7i2.970.

[16] F. M. Basysyar and G. Dwilestari, "House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection," *Acadlore Trans. AI Mach. Learn.*, vol. 1, no. 1, pp. 11–21, 2022, doi: 10.56578/ataiml010103.

[17] T. Sarwar *et al.*, "The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges," *ACM Comput. Surv.*, vol. 55, no. 2, 2023, doi: 10.1145/3490234.

[18] 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.

[19] A. Mello, "XGBoost: theory and practice," *https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6*, 2020. [Online]. Available: https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6

[20] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

DOI: https://doi.org/10.29207/resti.v7i4.4651