Investigation Related to Performance of KNN, Logistic Regression and XGBoost on Diabetes Prediction

Jiaguo Lin

Seaver College, Pepperdine University, Malibu, 90263, United States
jiaguo.lin@pepperdine.edu

Abstract. This study uses three different machine learning algorithms to build model for diabetes prediction and compares the accuracy of each model, and these algorithms are K Nearest Neighbors (KNN), Logistic Regression, and Extreme Gradient Boosting (XGBoost). The goal for this study is to find a precise algorithm for diabetes prediction, and this is really conductive to diagnosis of diabetes for doctors. In this way, patients can get apt treatment on time. Before building models, the dataset is pre-processed by standard scaling and Synthetic Minority Over-sampling (SMOTE) to balance the class. Then, Grid Search CV is used to find the best parameter for the model. Finally, the results show that KNN has an accuracy of 82%, followed by XGBoost which is 79.87% and Logistic Regression which is 75.5%. The advantage of KNN algorithm is that it only considers the distance between training sample and the new sample that is going to be predicted without any other computation. As a result, KNN demonstrated the best performance among these three algorithms. In the future, this study can expand the size of the dataset and try more parameters in order to achieve a higher accuracy on the model for diabetes prediction.

Keywords: Machine Learning, Algorithms, Diabetes.

1 Introduction

Diabetes is a serious global health issue that plagues a large number of patients and medical researchers since it induces significant impacts on personal health and public health. Diabetes can be caused by various reasons, including genetic factors, obesity, and insulin insufficiency. The study highlights the severity of this condition, revealing that diabetes ranked as the seventh leading cause of death in the United States in 2007, with 71,382 death certificates attributing diabetes as the underlying cause, accounting for nearly 3% of total mortality that year [1]. which indicates that diabetes has a terrifying threat to human life. Therefore, an accurate prediction algorithm for diabetes can greatly increase the chances for people to receive timely treatment, thus reducing the mortality rates.

Machine learning is an advanced prediction technology that has been widely employed in various domains in the last decades, which can be also considered as an effective way in predicting the probability of diabetes. The field of machine learning

© The Author(s) 2023
P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_70

is witnessing rapid advancements in contemporary times, emerging as a potent tool for human endeavors. In the medical field, machine learning also plays a very dominant role in promoting the rapid development of the medical field. For instance, machine learning applied to computer-aided diagnosis applications, these algorithms acquire knowledge by leveraging a substantial corpus of diagnosed samples obtained from medical test reports, coupled with expert diagnoses. This assimilation of data enables these algorithms to aid medical professionals in their prognostic and diagnostic endeavors, facilitating disease prediction and diagnosis in future cases [2]. Moreover, machine learning continues to show excellent performance in healthcare applications. In real life, it is challenging to process and analyze an abundance of heterogeneous data and information that is generated by healthcare service providers. In this way, machine learning methods play a pivotal role in efficiently analyzing diverse data types within the healthcare domain, thereby extracting valuable insights that can be acted upon [3]. These methods leverage heterogeneous sources of data, including but not limited to genomics, medical records, social media data, and environmental data [3]. This amalgamation of data from varied origins enriches the scope of healthcare data, facilitating comprehensive and multi-dimensional analyses. These practical applications of machine learning become persuasive evidence that shows the development in the medical field is greatly aided by machine learning. The use of machine learning not only can improve reliability and performance but also proffer a high accuracy result in many specific and complex tasks. In this way, making use of machine learning to build models for diagnosis of diabetes can offer doctors a precise diagnostic tool and help them to make decisions. Additionally, patients can gain more opportunities to receive timely treatment and achieve an increase in their chance of survival.

This study will focus on three machine learning algorithms to build classification models in order to predict whether the patient has diabetes or not. The three algorithms that are going to be used are K-Nearest Neighbors (KNN), Logistic Regression, and Extreme Gradient Boosting (XGBoost). The dataset that is considered in this study is a diabetes dataset from Kaggle [4]. The dataset is composed of 768 samples and 8 features such as pregnancies and glucose level. Each algorithm will be used to build a model. Based on the 8 features, the model will be trained by using eighty percent of the samples in the dataset. After that, twenty percent of the samples will be used to test the model, and the model will return its accuracy. By comparing the accuracy, the model that returns the highest accuracy will be the most suitable model for diagnosis of diabetes.

2 Method

2.1 Data Preparation

The dataset utilized in this study is sourced from Kaggle, consisting of 768 samples in the dataset and 8 features for each sample such as times of pregnancies and glucose level. This dataset is a part of the original dataset that is derived from the National Institute of Diabetes and Digestive and Kidney Diseases [4]. This dataset focuses on

the patients who are at least 21 years old of Pima Indian heritage women. The main aim of the dataset is to predict the existence of diabetes in a patient by using certain diagnostic measurements that are already provided in the dataset. There are two categories, where the "1" category stands for having diabetes, and the "0" category stands for not having diabetes.

Data preprocessing is a vital part in building and testing models. The first stage of data preprocessing is to ensure that all the features have similar scales, and this involves using the StandardScalar method to standardize the data in this study. However, the StandardScalar method is not applied to the XGBoost algorithm later since this algorithm has its own feature processing method. Secondly, the dataset is observed that there is a class imbalance between the samples who have diabetes and samples who do not have diabetes. There are 500 non-diabetes samples and 268 diabetes samples. To tackle this challenge, Synthetic Minority Over-sampling Technique (SMOTE) is adopted. SMOTE allows to synthetically increase the number of diabetes samples by generating new synthetic samples that are similar to the existing minority class instances. Subsequently, the dataset is splitted into training set and testing set, and the ratio between these two sets is 80 to 20.

2.2 Machine Learning Models

K-Nearest Neighbors (KNN). K-Nearest Neighbors has been studied for a long time and used broadly in many domains, and it is one of the most intuitive, oldest, and precise algorithms for classification tasks [5]. In many methods of supervised statistical classification, the K Nearest Neighbor maintains an outstanding performance and does not require the a priori assumptions of the distributions from the training set [6]. In the KNN classification approach, an essential step involves determining an appropriate value for K, which represents the number of neighboring samples considered during classification. For every new sample, the KNN algorithm calculates the distances between that sample and all samples in the training set. There are several formulas that are used to calculate the distances between samples like Euclidean distance or Manhattan distance. After that, the KNN algorithm will select the K nearest neighbor based on these distances. Eventually, the algorithm will classify the new sample by the majority vote in the labels of the K nearest neighbors.

Logistic Regression. Logistics Regression is a highly general supervised learning algorithm. It focuses on solving binary classification issues, and it has a strong similarity to linear regression [7]. This algorithm utilizes a logistic function to limit the results of linear regression on a range between 0 and 1, which is convenient to classify samples. Sigmoid function is commonly applied as the logistic function in logistic regression algorithm. Sigmoid function is defined as a mathematical formula shown in Formula (1), and it is able to create a curve that forms an "S" shape [8]. First of all, the logistic regression will use the maximum likelihood estimation method to gauge the model parameters for the provided training data. Secondly, logistic regression is going to operate a weighted sum to the samples that is going to be

predicted by using a linear function. Then, it will pass the result into the logistic function, and the logistic function will return a probability value. After that, based on the threshold, samples which have probability values that are greater than the threshold are classified as "1", and samples which have probability values that are smaller than the threshold are classified as "0".

$$s(x) = 1 / (1 + e^{\wedge} - x) \quad (1)$$

Extreme Gradient Boosting (XGBoost). XGBoost is an extremely strong and efficient algorithm based on gradient boosting, and it is described as a flexible machine learning system for tree boosting [9]. This algorithm is a powerful supervised learning algorithm, and it is intensely good at dealing with the problem of prediction with big training data and missing values [10]. XGBoost is composed of several weak learners. The main principle of the XGBoost algorithm is gradient boosting trees, and it trains a sequence of decision trees repeatedly. Moreover, it allows the error correction of the next tree according to the predictions of the previous tree. For each iteration, XGBoost is able to optimize the model by making the loss function smaller and renewing the weights of every decision tree by using gradient descent algorithm.

Implementation Details. For the KNN model, the parameter that needs to be tuned is "n_neighbors" which is the value of K that is mentioned previously. It is set from 3 to 200 with a step of 10, the best parameter is found to be 3 by using the GridSearchCV. In terms of the Logistic Regression model, 2 parameters are tuned, which are penalty and parameter "C". Penalty is the regularization term, and its ability is to prevent overfitting. Parameter "C" can control the strength of the regularization. The first step is to narrow down the range of C by applying RandomSearchCV method, and the result is 0.02154. Subsequently, GridSearchCV is used, and it is set with two penalty types which are l1 and l2 and a range of 30 values from 0.02154 / 20 to 0.02154 x 20 to find the best parameter "C". The final parameter is determined to be penalty as l1 and C as 0.16411. For the XGBoost model, multiple parameters were adjusted including max_depth, learning_rate, n_estimators, subsample, and colsample_bytree. Eventually, the best parameter is determined to be colsample_bytree as 0.8, learning_rate as 0.001, max_depth as 4, n_estimators as 1000, and subsamples as 0.8.

3 Results and Discussion

3.1 The Performance of Various Models

Following the development of three distinct models, each of them undergoes separate testing procedures in order to evaluate their respective performances. Accuracy is used as a metric to measure and compare the performance of the model in this study. The result is shown in Fig. 1. Based on the figure, the K-Nearest Neighbors (KNN) classifier attains an accuracy of 82%, while the Logistic Regression model achieves

75.5% accuracy, and the XGBoost classifier records an accuracy of 79.87%. Even though three models all show close accuracy, KNN still has the highest accuracy in this study, and XGBoost follows up.

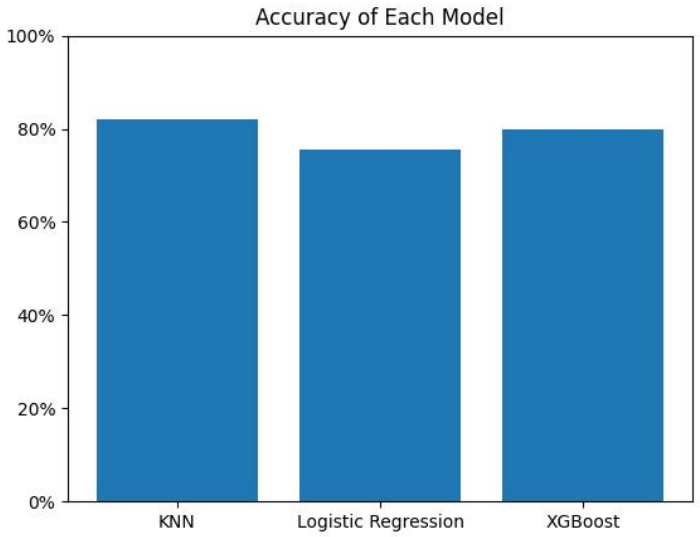

Fig. 1. The performance comparison among machine learning models.

3.2 Discussion

According to the result, the KNN algorithm has better performance in terms of accuracy. KNN classifies the new sample by measuring the distances between the training samples. In this way, its advantage is that it can consider more about the proximity between samples. This may explain why the KNN model returns the highest accuracy in predicting diabetes. However, the limitation of the KNN algorithm is also obvious because it is sensitive to the size and dimensionality of the dataset. If the size and dimensionality of the dataset increases, its execution time will increase as well when it has to calculate the distance for every prediction.

Logistic Regression algorithm is based on linear regression which classifies the new sample by using a logistic function. Logistic Regression has a slightly lower accuracy probably due to the fact that the Logistic Regression algorithm is based on the assumption of linear relationships in the data, but it is hard to find the relationships of the non-linear data. In order to improve the accuracy of the Logistic Regression algorithm, feature engineering and parameter tuning can be explored more. The advantage of the Logistic Regression is that it has a strong ability to interpret.

The XGBoost algorithm combines many predictions of weak learners to enhance performance. In this way, the XGBoost algorithm is able to handle complex datasets and incorporates optimization techniques, regularization, and parallel processing to

get better performance. For this reason, it displays a relatively high accuracy from the result. Regardless, training and parameter tuning for the XGBoost model is challenging and it needs a large number of computational resources in order to get a high accuracy. In the future, exploring more on parameter tuning can increase the accuracy of this model.

Nevertheless, this study has many limitations including the use of the datasets and limited feature engineering. The size and quality of the datasets also have a large influence on the models' performance. To improve the model's performances, having larger and better-quality datasets is really important. In addition, there are still many other machine learning algorithms that are also powerful. Trying more algorithms such as neural network also can help to predict diabetes precisely due to their excellent performance in other tasks [11, 12].

4 Conclusion

In conclusion, the main goal for this study is to analyze and compare the performance of three machine learning algorithms including KNN, Logistic Regression, XGBoost in diabetes prediction. After feature engineering and parameter tuning, the result provided by each model indicates that the KNN algorithm has an accuracy of 82%; the Logistic Regression algorithm has an accuracy of 75.5%; the XGBoost algorithm has an accuracy of 79.87%. The KNN algorithm performs better than the other two models in diabetes prediction. However, the fact that this study has to admit is the existence of limitations. This is because this study uses a specific dataset to test the model, and the result could be influenced by the features of another dataset. Moreover, in spite of the fact that the study includes the step for feature engineering and parameter tuning, there may be other feature engineering ways that are better and other combinations of parameters for each model, and this could further improve the performance. In order to overcome these limitations, future research should focus on expanding the size of the dataset such as adding new features and expanding the number of samples. Furthermore, integrating other strong machine learning algorithms like Support Vector Machines or Random Forest can promote the performance on diabetes predictions.

References

1. Dailey, G.: Overall mortality in diabetes mellitus: where do we stand today? Diabetes technology & therapeutics, 13(S1): S-65-S-74 (2011).
2. Shehab, M., Abualigah, L., Shambour, Q., et al.: Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine, 145: 105458 (2022).
3. Qayyum, A., Qadir, J., Bilal, M., et al.: Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering, 14: 156-180 (2020).
4. Kaggle Diabetes Dataset (2022) https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

5. Abu, A. H. A., Hassanat, A. B. A., Lasassmeh, O., et al.: Effects of distance measure choice on k-nearest neighbor classifier performance: a review. Big data, 7(4): 221-248 (2019).
6. Islam, M. J., Wu, Q. M. J., Ahmadi, M., et al.: Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. 2007 international conference on convergence information technology (ICCIT 2007). IEEE (2007) 1541-1546.
7. Sperandei, S.: Understanding logistic regression analysis. Biochemia medica, 24(1): 12-18 (2014).
8. Jónás, T.: Sigmoid functions in reliability based management. Periodica Polytechnica Social and Management Sciences, 15(2): 67-72 (2007).
9. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794 (2016).
10. Fauzan, M. A., Murfi, H.: The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2): 159-171 (2018).
11. Qiu, Y., Yang, Y., Lin, Z., et al.: Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications, 17(3): 46-57 (2020).
12. Ogundokun, R. O., Misra, S., Douglas, M., et al.: Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. Future Internet, 14(5): 153 (2022).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.