GANIT

Available online at https://ganitjournal.bdmathsociety.org/

GANIT: Journal of Bangladesh Mathematical Society

GANIT *J. Bangladesh Math. Soc.* 43.2 (2023) 01–07
DOI: https://doi.org/10.3329/ganit.v43i2.70794.

A Comparative Analysis Of Logistic Regression and K-Nearest Neighbors Algorithms In Diagnosis Of Diabetes

Hasan Mahdi Mahi and Adeeb Shahriar Zaman *

Department of Mathematics, Dhaka University, Dhaka-1000, Bangladesh

ABSTRACT

Machine Learning techniques have gained prominence in medical diagnosis due to their ability to uncover patterns in complex data-sets, thereby giving accurate disease classification. In this study, we mainly focus on the application of two widely used Machine Learning algorithms, Logistic Regression and K-Nearest-neighbors(KNN), for the purpose of distinguishing patients with diabetes from those without. Our research aims to shed light on the comparative accuracy and performance of these algorithms in a medical context. The methodology section outlines experimental setup, detailing data processing, algorithm training and testing procedures. A comprehensive data-set comprising medical attributes is utilized for evaluation and accuracy metrics are employed to quantify the performance of the algorithms. Results has shown efficacy of both the algorithms and our findings showcase the strengths and limitations of each approach, contributing on the applicability in medical decision making. By offering a nuanced comparison, we illuminate a path for more robust and accurate disease identification techniques, further enhancing patient care and medical outcomes.

Received: August 18, 2023 **Accepted:** October 28, 2023 **Published Online:** December 31, 2023

Keywords: Logistic Regression; K-Nearest-Neighbors; Machine Learning; Diabetes Diagnosis
AMS Subject Classifications 2020: 68T10, 68W01, 90C30, 90C90.

1 Introduction

In the era of advancing technology, machine learning has emerged as a pivotal tool revolutionizing various domains, such as healthcare. Harnessing the ability to learn patterns from data, machine learning algorithms play an important role in aiding medical diagnostics. One of the crucial areas where machine learning has shown remarkable potential is in the accurate classification of medical conditions, such as diabetes. This paper delves into the application of two prominent machine learning algorithms, Logistic Regression and K-Nearest Neighbors (KNN), to distinguish patients with diabetes from those without.

Machine learning algorithms are computational models designed to learn from data patterns and make informed decisions without being explicitly programmed. In the realm of medical diagnosis, they offer the possibility of improving accuracy, efficiency, and ultimately patient care. Logistic Regression, a statistical method, estimates the probability of a binary outcome by analyzing input features. On the other hand, KNN is a

*Adeeb Shahriar Zaman *E-mail address:* adeeb.math@gmail.com

non-parametric algorithm that classifies a sample by comparing it to the k-nearest neighbors in the training data-set. Both methods have been extensively utilized in medical contexts, demonstrating their potential to aid clinicians in making accurate diagnoses.

This study focuses on evaluating the performance of these algorithms within the specific context of diabetes diagnosis. Diabetes has always been a major health issue, particularly because it does not heal entirely if it is not diagnosed early. Also for leading a better life by having a control on higher blood pressure and hypertension as a consequence of diabetes, necessary steps to be taken [2, 3]. Several research-studies have already been conducted in application of different machine learning algorithms in diagnosis of diabetes, and many many research works are still ongoing in this field as there is still a lot of room for improvements [1]. We aim to provide insights into the strengths and limitations of Logistic Regression and KNN in identifying patients with diabetes accurately. By comparing their accuracy, we can contribute to the broader dialogue on enhancing medical diagnosis using machine learning techniques.

The subsequent sections of this paper provide a comprehensive overview of the methodology employed for training and testing the algorithms, details about the data-set used for evaluation, the obtained results, and a discussion of the implications of our findings. Through this research, we strive to illuminate the potential of machine learning algorithms in the area of diabetes diagnosis and their wider applications in healthcare.

2 Preliminaries

2.1 Machine Learning

Machine Learning is a process where a computer learns patterns from given data and do some specific tasks by itself. Those tasks are done by the experience it gets form the training data.

Consider a spam filter program. It learns by using samples of both spam and regular emails. These samples are training set and each one is training instance. The goal is to identify spam in new emails, using the experience it got from the training data-set. To measure how well it performs, we can check the accuracy on how many emails it can correctly classify [4, 5, 6, 7].

There are various types of Machine Learning systems that can be categorised. But in this paper, we will be discussing the type where they are trained by human supervision. Here, we will mostly focus on Supervised Learning and it's classification.

2.2 Supervised and Unsupervised Learning

Machine Learning algorithms can be classified into two classes, namely Supervised and Unsupervised Machine Learning algorithm. Supervised Learning is the process where it uses the labeled data-set to train algorithms for classifying data or predicting outcomes accurately. The model learns from a training set that encompasses input data and corresponding accurate outputs, facilitating gradual learning [5].

Unsupervised Learning involves machine learning algorithms to analyze and cluster unlabeled data, without requiring a predefined training data-set. Instead, it automatically discovers hidden patterns and insights. This mirrors human learning process where knowledge is gained from data without supervision. For example, if we give the algorithm a picture to identify if it is a cat or a dog, it won't need any training data, rather it will just adapt by itself using the basic features distinguishing those two animals.

Here, we won't go further on Unsupervised Learning; we will be focusing mainly on Supervised Learning. In the aspect of data analysis, there can be two types of Supervised Learning problems. They are classification and regression.

- Classification involves employing algorithms that consider specific categories to test the data enhancing better accuracy for the algorithm. However, not all types of issues can be benefited from this approach for

improved precision, which we will explore later. Some of the classification algorithms are linear classifier, support vector machine(SVM), decision trees, k-nearest neighbors, random forest and so on.

- Regression comes into play where data displays a strong relation between dependent and independent variables. Regression identifies patterns with the training data-set to align with trial data, ensuring that when test data is introduced, it conforms to the established patterns. Some of the popular regression algorithms are linear regression, logistic regression and polynomial regression [6].

Now we will discuss on a regression algorithm, specifically, Logistic Regression and a classification algorithm, specifically, K-Nearest-Neighbors(KNN). Later in the paper we will make a comparison between these two algorithms in terms of specific conditions and a specific problem.

2.2.1 Logistic Regression

Logistic Regression is a regression algorithm where we will train the data-set with the help of a special function called the sigmoid function. The sigmoid function is given by,

$$f(x) = \frac{1}{1 + e^{-x}}$$

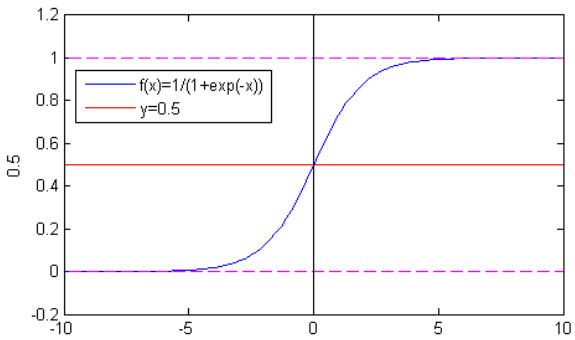

Figure 2.1: Sigomid Function

Logistic model is widely used to solve classification problems. This model involves probability to take binary decisions, that is it solves classification problems that involve exactly two classes. So the output of the algorithm is either 1 or 0 (or true/false or yes/no).

Logistic Regression is developed based on the idea of Linear Regression. The Linear Regression model is given by

$$y = \theta^T x \tag{2.1}$$

In Logistic Regression, we pass this output of the Linear Regression model through the sigmoid function to get the Logistic Regression model estimated probability,

$$\hat{p} = h_\theta(x) = f(\theta^T x) \tag{2.2}$$

In Linear Regression model, the Mean Squared Error(MSE) cost function is,

$$MSE(X, h_\theta) = \frac{1}{m_i} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 \tag{2.3}$$

In case of Logistic Regression, the term $\theta^T x^{(i)}$ is simply replaced by $f(\theta^T x^{(i)})$ or $h_\theta(x^{(i)})$. Hence, the cost function for Logistic Regression model is

$$MSE(X, h_\theta) = \frac{1}{m_i} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{2.4}$$

The main goal in this model is to find the vector θ^T that minimizes the cost function in equation (2.4). After the choice of θ^T is made, all that remains is to check whether $\hat{p} > 0.50$. We define,

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} > 0.5 \end{cases} \tag{2.5}$$

Since the Logistic Regression model helps to solve only binary decision making problems, it can be used to get a very good prediction on cases such as predicting win or lose for a candidate in an election, predicting the chance of a student to get admission, diagnosis of a specific disease etc [5].

2.2.2 K-Nearest-Neighbors Algorithm

K-nearest-neighbors(KNN) method is a supervised learning method which is worked off the assumption that the similar points can be found near one another. It can be used for both classification and regression problems but mainly used in classification where a binary decision is to be made. KNN method contains distance matrices for the test data to be compared with the trial data. Some of the distance matrices are :

Euclidean Distance :

$$d(x, y) = \sqrt{\sum_{i=1}^{\infty}(x_i - y_i)^2}$$

Minkowski Distance :

$$d(x, y) = (\sum_{i=1}^{n}|x_i - y_i|)^{1/p}$$

Manhattan Distance :

$$d(x, y) = \sum_{i=1}^{m}|x_i - y_i|$$

For KNN method, if we consider the test data(some unclassified points on the graph), we can assign it to the group it is least distant from. The idea being used here is the fact that a point close to a cluster of points classified as 'Red' has a higher probability of getting classified as 'Red' [7].

Choosing the value of k is crucial here. The choice of k mainly depends on the input data. For a data-set containing clear and strong pattern, $k = 1$ will work fine. But in case of dealing with data-sets that contain ambiguities, choosing $k = 1$ will no longer work. In those cases, we need to choose a suitable value of $k > 1$. In case of ties between two groups, we may have to improvise. Integers that are not divisible by the first few primes can be good choices for k to avoid such situations. Cross-Validation is a method for selecting the best value of k from input data.

3 Model Framework

We shall work with a data-set that contains information about 768 female patients aged at least 21 years from Pima Indian heritage. This data-set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [8]. The data-set contains eight information about the 768 patients. Those are:

Pregnancies: Expressing the Number of pregnancies
Glucose: Expressing the Glucose level in blood
Blood Pressure: Expressing the Blood pressure measurement
Skin Thickness: Expressing the thickness of the skin
Insulin: Expressing the Insulin level in blood
BMI: Expressing the Body mass index
Diabetes Pedigree Function: Expressing the Diabetes percentage
Age: Expressing the age
Diagnosis: Expressing the diagnosis

We have used the first seven information as features and the last one as a binary variable that outputs 0 or 1. Here, 1 implies that the person has diabetes and 0 implies that she does not have diabetes. 85% of the data that has been given were used as the trial data. The remaining 15% were used as test data. Those 15% were chosen randomly. We used Logistic Regression first then K-Nearest-Neighbors(KNN) algorithm and compared the accuracy level of the two methods in predicting whether the test patients have diabetes or not.

4 Result and Analysis

We observed that in 76% cases, the Logistic Regression successfully predicted the correct diagnosis. That is, the accuracy is 76% for Logistic Regression.

Now, it is not possible to have a proper visualisation of all the seven featured constraints and so many data to see how they follow sigmoid function. So we randomly take two of the constraints among the seven to perfectly judge the visual representation of logistic regression in 3D. We will get the picture as Figure 4.1. In this figure, we have taken Age and Glucose level of first 100 patients from 768 patients [8]. Here, we can see that it resonates with the sigmoid function.

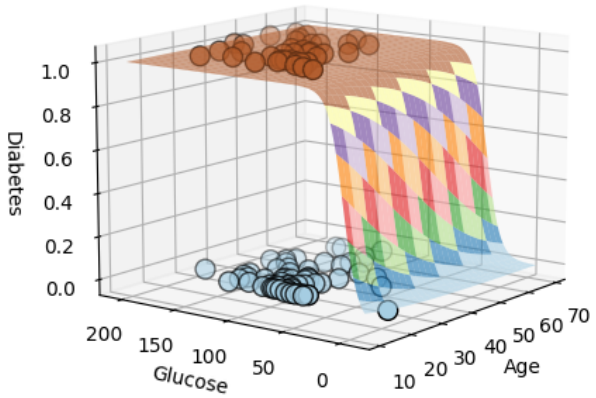

Figure 4.1: Logistic Regression while taking Age and Glucose constraints for first 100 patients [8]

The accuracy of the K-Nearest-Neighbors(KNN) algorithm was 70% for $k = 4$. For $k = 3$ it gave 69% accurate output and we have got 66% for $k = 9$ and $k = 10$. We can see from Figure 4.2 that for $k \geq 14$, there is a noticeable spike in the graph, indicating a rapid increase in accuracy.

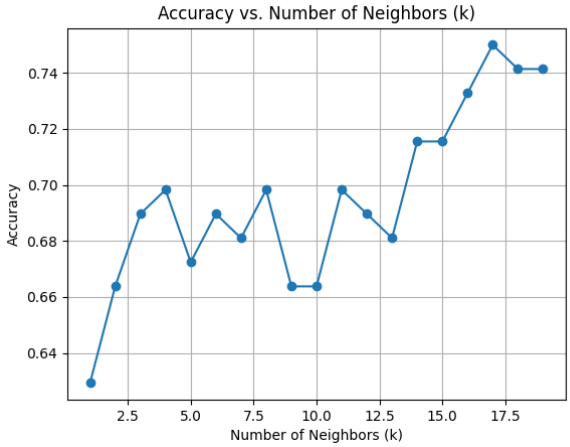

Figure 4.2: Accuracy vs Number of Neighbors for different values of k up to 20

Figure 4.2 suggests a positive correlation between the value of k and its corresponding accuracy. This gives an impression that the accuracy may keep increasing if we keep taking larger values of k. In theory, this should not be the case since at some point, the accuracy has to start decreasing with k. This is mainly due to the fact that taking large value of k would imply grouping our test data with a lot of train data that are not so close to it. This contradiction with theory motivated us to experiment with larger values of k. Following is the resulting graph we got.

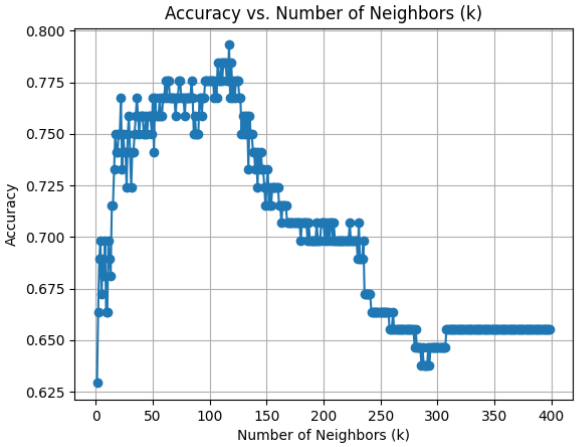

Figure 4.3: Accuracy vs Number of Neighbors for different values of k up to 400

Figure 4.3 clearly shows that the accuracy finally starts dropping at around 120 (117 to be exact). This result is coherent with the theoretical ideas. A smaller k might make the model more sensitive to individual noisy data points. Increasing k can help to mitigate the impact on noisy data points. But as we have reasoned earlier, there has to be an upper bound for the effective value of k as well. In our case, the most effective value of k was found to be 117, producing an accuracy of 79.26%.

Finally, we can propose that for the given features in the data-set we have used, when the value of k is optimized, k-nearest-neighbors(KNN) method gives more accurate result compared to the Logistic Regression model [8].

5 Conclusion

In this study, we embarked on a journey to explore the application of machine learning algorithms in diagnosis of diabetes. Comparing the two methods, Logistic Regression and K-Nearest-Neighbors(KNN), we aimed to ascertain their comparative accuracy and potential in aiding medical decision making. Our findings reveal that both Logistic Regression and KNN exhibit commendable accuracy rates in distinguishing patients with diabetes from those without. Logistic Regression demonstrates its strength in establishing linear decision boundaries. On the other hand, KNN excels in local pattern recognition, pointing out relationships within the data that might be overlooked by other methods. Though the KNN method provided slightly better accuracy, results of both the algorithms suggest that there is possible room for improvements. As we look for the future, the connection of machine learning and healthcare holds immense promise. Continued research in this area, exploring diverse algorithms and refining methodologies will undoubtedly contribute to enhance diagnostic accuracy. This study serves a stepping stone in this journey providing insights into the potential of machine learning algorithms in medical diagnostics.

References

[1] Neha P.T., Shruti G. *Prediction of Type 2 Diabetes using Machine Learning Classification Methods*,International Conference on Computational Intelligence and Data Science, Procedia Computer Science, vol 167,P(706-716),2020.

[2] Islam, M., Chaudhuri, I. , Mobin, M. , Islam, M., Mahmud, M., KutubUddin, M. , Kabir, K. and Kamrujjaman, M. (2021) The Perspective of Acquired Immunity to Combat against Infectious Diseases: An Overview. Health, 13, 1020-1044. doi: 10.4236/health.2021.139077.

[3] Mobashara Islam, Irfan Chaudhuri, Md. Shahidul Islam, Md. Kamrujjaman, 2023. "A Review on Hypertension: Practice and Diagnosis," Journal of Biology and Life Science, Macrothink Institute, vol. 14(2), pages 18-38, August.

[4] Aurelién Geron, 2019 *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, Inc, 2nd edition.

[5] Supervised Machine Learning, https://www.javatpoint.com/supervised-machine-learning;

[6] Shalev-Shawartz S. and S. Ben-David,2014. *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, New Work,NY,USA.

[7] K-Nearest-Neighbors Algorithm, https://www.geeksforgeeks.org/k-nearest-neighbours/

[8] Data of Diabetes Patients, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset