Journal of Techniques, ISSN: 2708-8383, Vol. 6, No. 3, 2024, Pages 19-25
DOI: https://doi.org/10.51173/jt.v6i3.2587

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - COMPUTER ENGINEERING

Improving Diabetes Prediction by Selecting Optimal K and Distance Measures in KNN Classifier

Emad Majeed Hameed[1], Hardik Joshi[1*]

[1]Department of Computer Science, Gujarat University, Ahmedabad, India

[*] Corresponding author E-mail: hardikjoshi@gujaratuniversity.ac.in

Article Info.

Article history:

Received
24 April 2024

Accepted
16 June 2024

Publishing
30 September 2024

Abstract

Diabetes is an illness that is widespread throughout the world and is considered a health concern, which requires work to explore advanced predictive techniques for early diagnosis of the illness. This paper discusses diabetes prediction by using the K-Nearest Neighbors (KNN) classifier, which is a widely used algorithm in machine learning. Most studies only dealt with investigating the optimal value of k in the KNN algorithm and did not address the best method to measure distance alone or together with the optimal value of k to improve the efficiency of diabetes prediction. This study simultaneously investigates both the optimal value of k and the optimal method for measuring distance to improve the performance of the KNN technique in predicting diabetes. By using and analyzing the Indian Diabetes PIMA dataset, this study seeks to discover the extent to which different parameters, especially the optimal value of K and distance metrics, affect the performance of the classifier. Through experiments that included applying different values for the K factor and using various distance measures, the study reached insights into maximizing the classifier's accuracy. The study shows that choosing the distance measure greatly affects the accuracy of classification and selecting the optimal K value helps eliminate problems of overfitting and underfitting, which is a feature of robust models for diabetes prediction. The research results showed that the best performance achieved was 80.5% when k=35 and the Euclidean distance measure was used.

Publisher: Middle Technical University

Keywords: Diabetes; Prediction; Feature Selection; KNN.

1. Introduction

The World Health Organization classifies diabetes as a disease of modern society and a major global health concern. About 25.8 million individuals, making up 8.3% of the US population, suffer from diabetes. The projected healthcare expenditure for diabetes is estimated to reach $490 billion by 2030, making up 11.6% of the total global healthcare spending. Diabetes Mellitus denotes a metabolic disorder resulting from abnormalities in insulin production and function. It is marked by hyperglycemia.

In diabetic individuals, hyperglycemia results from either the ineffective production or absence of insulin. Failure to diagnose and treat diabetes adequately in its initial stages can lead to significant consequences such as cardiovascular disease, visual impairments, renal failure, delayed wound healing, and the potential for leg amputation. Additional indicators of the severity of diabetes include numbness or tingling in the hands or legs[1, 2]. Three main forms of diabetes exist., namely Type 1 diabetes (T1D), Type 2 diabetes (T2D), and Gestational diabetes. Type 1 diabetes is often diagnosed in individuals under the age of 30 and is characterized by signs including constant thirst, increased urination, and high levels of blood sugar. This type of diabetes is usually treated with medications. On the other hand, there are several factors associated with type 2 diabetes, the most important of which are weight gain, high blood pressure, atherosclerosis, and other health problems. It tends to affect middle-aged and older adults. Lifestyle choices, lack of physical activity, dietary habits, obesity, smoking, high cholesterol (hyperlipidemia), and hypertension (hyperglycemia) are known contributors to the development of Type 2 diabetes. In gestational diabetes, Diabetes detected in pregnant during the sixth or seventh month of pregnancy tends to resolve after the baby is delivered. Diabetes cannot be cured, but adopting a nutritious diet, using medication, doing physical exercise, maintaining a healthy body weight, and undergoing regular screenings can help postpone or prevent its associated complications [3, 4].

Over the past few years, considerable researches have been conducted to predict diabetes using machine learning methods. Regarding the field of machine learning, classification stands out as a crucial method in constructing predictive models. Various machine learning approaches prove beneficial in analyzing data from various angles and condensing it into valuable insights. Various intelligent learning techniques involving KNN, Naive Bayes, logistic regression, SVM, random forests, decision trees, and ANN are employed for disease diagnosis. This study presents a predictive model utilizing the KNN algorithm for diagnosing patients into diabetic and non-diabetic groups. The effectiveness of the classifier is assessed through accuracy metrics. Additionally, a study is conducted to determine the best value of K and distance measures for maximizing the performance of the KNN classifier. Despite being a straightforward and "lazy" algorithm, KNN demonstrates superior performance, particularly with smaller datasets. Most studies only dealt with investigating the optimal value of k in the KNN algorithm and did not address the best method to measure distance alone or together with the optimal value of k to improve the efficiency of diabetes prediction.

Nomenclature & Symbols			
KNN	K-Nearest Neighbors	KNNB1R	K-Nearest-Neighbor-Based-OneR
T1D	Type 1 Diabetes	AF-KNN	Adaptable Fuzzified K-Nearest Neighborhood
T2D	Type 2 Diabetes	SSE	Sum of the Squared Differences
m	Count of Constraints	K	Signifies the Number of Independent Variables

By using and analyzing the Indian Diabetes PIMA dataset, this study seeks to discover the extent to which different parameters, especially the optimal value of K and distance metrics, affect the performance of the classifier.

2. Related Works

The task of predicting diabetes and classifying diabetic patients has been extensively explored in the medical literature. Numerous researchers have investigated different methodologies to enhance the performance and accuracy of predictive models. In this section, we review several notable studies in this domain. Bano et al. [5], focused on the PIMA dataset and implemented models using both KNN and SVM classifiers in WEKA. Their KNN-based model achieved an accuracy of 85.8%. Sneha et al. [6], explored random forest, Naïve Bayes, KNN, SVM, and decision tree algorithms for this purpose. Among these classifiers, KNN demonstrated an accuracy of 63.04%. Jianping Gou et al. [7], proposed a classifier utilizing a dual-weighted voting function to mitigate the influence of outliers in the KNN neighborhood. This work contributed to improving the performance of prediction of KNN-based models, especially in datasets involving outliers. Kumar et al. [8], used CART, LDA, random forest, KNN, and SVM algorithms on a dataset obtained from a diagnosis lab. In this study, the Random Forest algorithm gave the best performance, and the performance of the KNN algorithm was 59.69%. In the study presented by Aishwarya J. et al. [9], the authors applied different intelligent techniques to the same dataset. Among these techniques were logistic regression and KNN which had the highest accuracy of 77.6% and 73.43 respectively. A novel KNN algorithm for diabetes prediction was developed by Christobel et al. [10]. The authors called their algorithm CKNN. This algorithm achieved an accuracy of 78.16% which is considered better than the traditional KNN algorithm. Another novel KNN algorithm was presented by Amal H. et al. [11]. This algorithm was called K-Nearest-Neighbor-Based-OneR (KNNB1R). The work of this algorithm depends on the principles of the One-Attribute-Rule Algorithm to adjust the weights of attributes and heighten the accuracy of the traditional KNN algorithm. After assigning the optimal weights to features of the diabetes dataset used in this study, the authors noted a significant improvement in the performance of the classifier. This algorithm achieved an accuracy of 92.91%. The work of Iqbal H. et al [12], was dedicated to building a diabetes prediction model based on an optimal KNN. Their work aimed to find the optimal value of K that leads to minimizing the errors and enhancing the accuracy. This proposed model was applied to a real dataset obtained from the medical hospital to prove its effectiveness. Gupta and Goel [13], used KNN along with other machine learning methods in a prediction model, to improve the performance of classifiers. They used the PIMA diabetes dataset. One of the goals of their study is to determine the optimal values of the parameter "K" in the KNN classifier. Gupta and Goel observed that the classifier gives its best performance when the number of neighbors (K) is either 33, 40, or 45. These ideal values provide an accuracy of 87.01% and an error rate of 12.99%, which confirms the effectiveness and accuracy of the classifier. Saxena et al. [14], in their study addressed the application of the k-nearest neighbor (KNN) algorithm to the PIMA diabetes dataset. Their work resulted in a significant improvement in accuracy. When using their proposed algorithm, the results showed an increase from 70.1% to 78.58%, representing an improvement of 8.48%. Prasad et al. [15], proposed an adaptable fuzzified K-Nearest Neighborhood (AF-KNN) method for diagnosing diabetes. The authors used the optimal number of k by analyzing the minimum inaccuracy. The two different datasets TLGS and PIDD were used to implement this algorithm. The proposed algorithm gave a consistent performance accuracy of 99.32% for both datasets. Al-Nuwaisir, K. [16], developed an automated method to predict diabetes, the researcher focused on carefully dealing with missing data and improving performance. The proposed approach uses the K-Nearest Neighbor (KNN) imputed features and a Tri-ensemble voting classifier model. The results of this model showed a performance accuracy of 97.49%. Suriya, and Oanish [17], applied the KNN algorithm to various datasets. Different steps of preprocessing have been performed on the datasets including removing the null values, normalization, and feature selection. The authors noticed that the best accuracy was 73% when k=40.

It is noted from the literature works mentioned in this study that were conducted on examining the predictive efficiency of the KNN algorithm for diabetes that there is a gap regarding the systematic examination and improvement of the parameters of the KNN algorithm. Most studies only dealt with investigating the optimal value of k in the KNN algorithm and did not address the best method to measure distance alone or together with the optimal value of k to improve the efficiency of diabetes prediction. We noted the studies [12], [13], and [15] focused on finding the optimal value of k for minimizing the errors. This study simultaneously investigates both the optimal value of k and selects the optimal method for measuring distance to improve the performance of the KNN technique in predicting diabetes.

3. K- Nearest Neighbors Algorithm

K-Nearest Neighbors algorithm is one of the non-parametric supervised learning methods that is used in both classification and regression tasks. The work of this algorithm needs no knowledge in advance about the distribution of the dataset. The ease of use and high accuracy with smaller datasets make this algorithm used in this work. The principle of work of this algorithm is based on the idea of similarity between points of the dataset [18, 19]. In the KNN algorithm, the similarity metrics assign weights to the nearest neighbors K which affect the classification performance. the number of neighboring data points that will be considered in the classification process is decided according to the value of K. The class label can be predicted by majority voting of the selected nearest neighbors. Several methods are available to calculate the similarity between points in the dataset, among these methods are Euclidean, Manhattan, and Minkowski distances. The Euclidean distance method is widely applied to measure similarity. To calculate the Euclidean distance between two points X and Y, we use the formula (1) [20, 21]:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

The Manhattan distance is calculated by finding the sum of the absolute differences between the coordinates of points. For finding the Manhattan distance of two points s and r in space m, formula (2) is used.

$$d(s,r) = \sum_{j=1}^{m} |s_j - r_j| \quad (2)$$

The Euclidean distance and the Manhattan distance are combined in the Minkowski method to calculate the distance between the two points s and r through formula (3):

$$d(s,r) = \left(\sum_{i=1}^{m} |s_i - r_i|^s\right)^{\frac{1}{s}} \quad (3)$$

The procedure of finding the maximum absolute difference between two points is used in Chebyshev distance. Formula (4) is used for finding the Chebyshev distance D in m-dimensional space between two points s and r [22].

$$D(s,r) = max_{i=1}^{m} |s_i - r_i| \quad (4)$$

The characteristics of the dataset are considered the most important factor that can be used to choose the distance measure method.

When talking about the advantages and disadvantages of this algorithm, its advantages are ease and simplicity in implementation, its effectiveness with noisy data, and its efficiency when used with large training data. As for its disadvantages, no measurement guide exists to determine the optimal value of the parameter K, and the high computational cost is due to the necessity of calculating the distances between each test instance and all training samples. Also, when used with multi-dimensional data sets that contain irrelevant features, it gives less accuracy [23].

The procedure for KNN classification involves the following steps:

- Importing the dataset.
- Setting the value of K.
- For each data point in the training dataset:
  - Compute the distance between the test data point with every data point in the training set.
  - Order the distances in ascending mode.
  - Select the first K data points from the arranged list.
  - Obtain the class labels of the selected K data points.
  - For classification problems, return the mode of the K class labels.
  - For regression problems, return the mean of the K class labels.

4. Methodology

An experimental analysis was conducted to evaluate how effectively the KNN algorithm performs on the PIMA Indian Diabetes dataset. This investigation aims to determine the optimal value of K and distance metric that yields the highest performance for the KNN algorithm. The diagram in Fig. 1 illustrates the methodology of the prediction model.

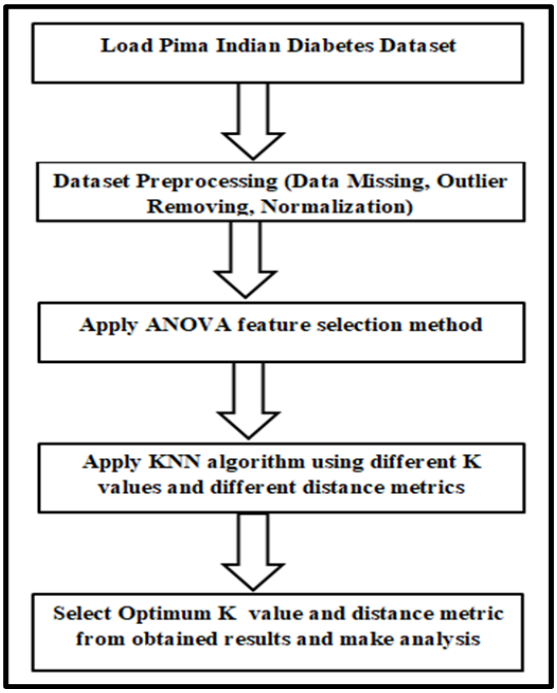

Fig. 1. Work Flow of Proposed Model

The PIMA dataset can be downloaded from the UCI repository. It comprises testing data, totaling 768 rows with 9 features (see Fig. 2). The initial eight features in the dataset include Pregnancy, Blood Pressure, Glucose, Skin Thickness, Insulin, DPF, BMI, and age; which denote the medical information of patients. The ninth feature indicates the outcome. In the dataset, there are 268 diabetic patients and 500 non-diabetic

patients [24]. The dataset might include missing values, errors, or other inaccuracies due to improper data collection methods. These discrepancies could impact the classifier's effectiveness. In this dataset, missing values are replaced with the mean of the respective attributes.

When there is too much variation between the data in the dataset, normalization is used to manage the data on a common scale and make the data easier to compare to one another. The procedure of finding the statistical mean and standard deviation of the attribute values, subtracting the mean of each value, and dividing the output by the standard deviation is known as standardizing a statistical variable. Numerical values with one standard deviation and zero mean are generated. The Python Function Standard Scaler is used to standardize the features of the dataset.

Abbreviation	Factor	Detail
Pr	Pregnancies	Number of times pregnant
Gl	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Bp	Blood Pressure	Diastolic blood pressure
St	Skin Thickness	Triceps skinfold thickness
In	Insulin	2-Hour serum insulin
Bm	BMI	Body mass index
Dpf	Diabetes Pedigree Function	Diabetes pedigree function
Ag	Age	Patient ages

Fig. 2. Pima dataset features

Feature selection is the process of reducing the dimensions in a dataset by assessing their significance in identifying the class label. This process determines and removes irrelevant attributes from the dataset. In this study, the ANOVA method is used for selecting features.

The basis of analysis of variance (ANOVA) is statistical hypothesis testing, which aims to compare means among groups. The key idea behind an ANOVA is to separate the variation in the data as a whole into two parts: the variance within groups and the variance between groups. ANOVA determines if the group means differ substantially from one another by calculating the F-statistic, which is the ratio of between-group variance to within-group variance. Strong discriminating power is exhibited by features with a low p-value and a high F-value, which suggests that they have significant variations in-group means and might be chosen for feature selection. The following formula can be used to get a feature's F-value [25].

$$F = \frac{\frac{SSE1 - SSE2}{m}}{\frac{SSE2}{n-k}} \tag{5}$$

Where SSE represents the sum of the squared differences, 'm' denotes the count of constraints, and 'k' signifies the number of independent variables. Table 1 provides the ANOVA f-values for the eight-dataset features

Table 1. F value of features of the PIMA dataset

Feature	F value
Glucose	213.16
BMI	71.77
Age	46.14
Pregnancies	39.67
Diabetes Pedigree Function	23.87
Insulin	13.28
Skin Thickness	4.3
Blood Pressure	3.25

It is evident from the Table.1 that Skin Thickness and Blood Pressure are less useful features for predicting the result. Thus, the dataset is resized to 768 by 7 and these features are removed.

When varying the number of neighbors and distance measures, KNN displays distinct scores. Choosing the ideal number of neighbors (K) with the optimum distance metric on which it performs best is a bit challenging. It varies depending on the dataset. K is therefore examined ranged from 1 – 50 and four distance measurements used; The Euclidean distance, Manhattan distance measures, the Minkowski distance, and the Chebyshev distance. The optimal K and distance measures are selected according to the best performance of KNN.

5. Result and Discussion

To boost the efficiency of the KNN classifier, the classifier parameters are adjusted using a set of different values of k and using four different distance measures, then comparing their accuracy on the data set.

Figs. 3 - 6 show the accuracy of KNN using various *k* values and different distance measures. The four distance measures included are Euclidean, Manhattan, Chebyshev, and Minkowski distances. As illustrated from these figures the best performance achieved was 80.5% when *k*=35 and the Euclidean distance measure was used.

By observing the results of this study, we can conclude that the performance of the KNN algorithm is affected by the choice of distance measure. It has been observed that the Euclidean distance measure provides the highest accuracy when $k = 35$, while other distance methods such as Manhattan and Chebyshev achieved competitive results, confirming their effectiveness in certain scenarios.

Concerning how *k* value affects the algorithm efficiency, it is noticeable that the accuracy increases with the increasing values *k* until it reaches the greatest when the value is 35, then we notice a gradual decrease in accuracy with the continued increase in the value of *k*. An interpretation of these results suggests that having a moderate number of nearest neighbors is optimal for our data set, with overfitting and underfitting being consistent.

Depending on what has been mentioned, the study emphasizes the importance of choosing a value of *k* and the distance measure depending on the data set used and the nature of the problem.

Other evaluation metrics can be mentioned to describe the performance of the KNN classifier. Tables 2 and 3 show the evaluation metrics of the KNN classifier when k=35 and the distance measurement is Euclidean.

According to the Tables 2 and 3, the model classified correctly the positive class with 76.5%, and the classifier correctly predicted 34.55% of the negative classes. The error rate, which represents the percentage of inaccurate classifications the model made, is around 19.48%.

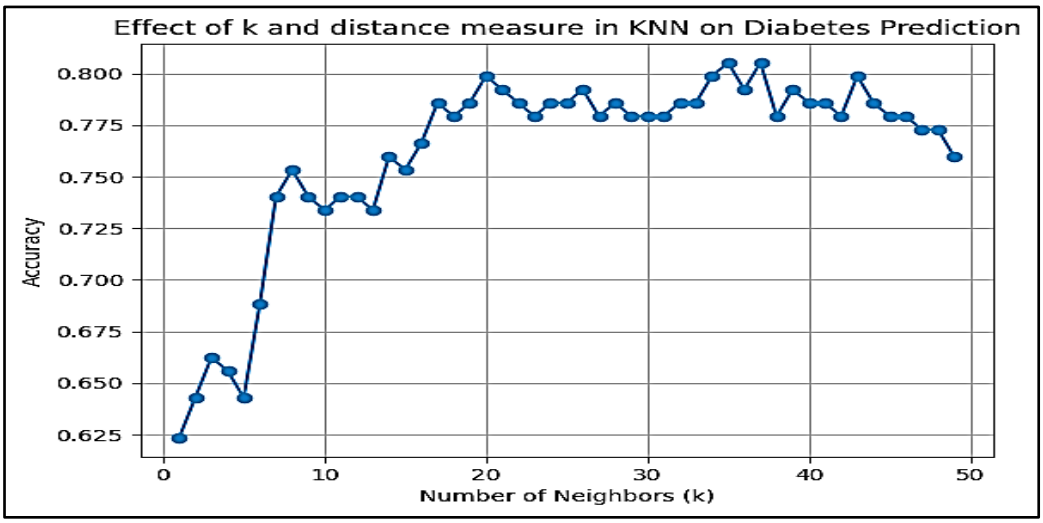

Fig. 3. The KNN performance using different k values and the Euclidean distance method

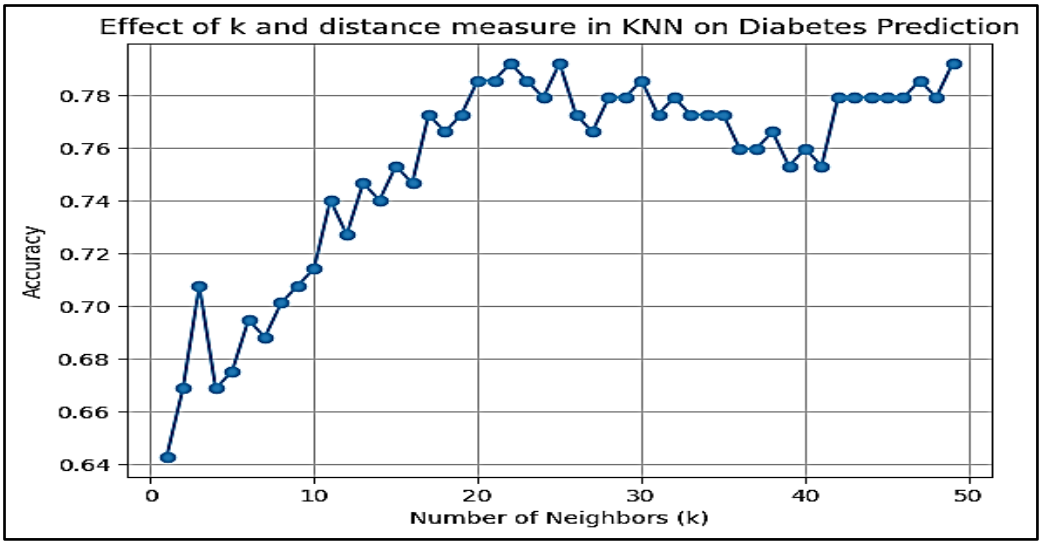

Fig. 4. The KNN performance using different k values and the Manhattan distance method

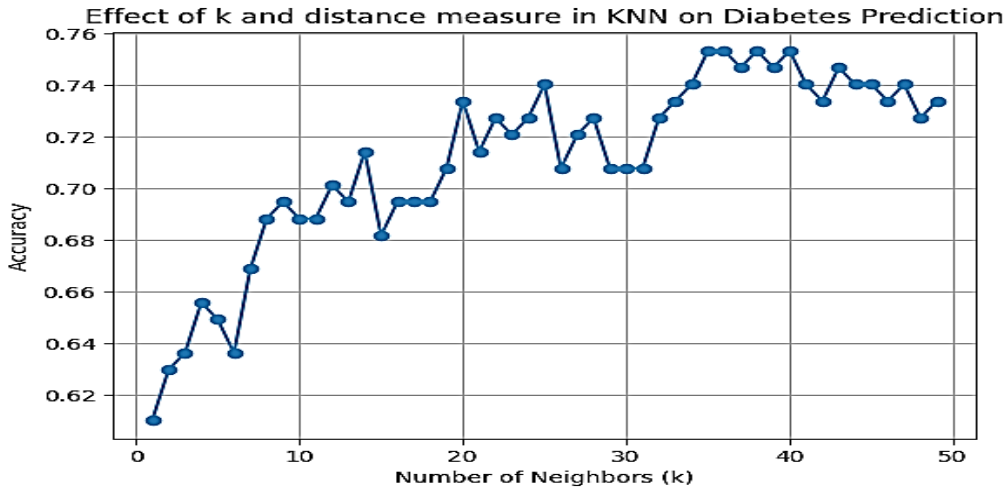

Fig. 5. The KNN performance using different k values and Chebyshev distance method

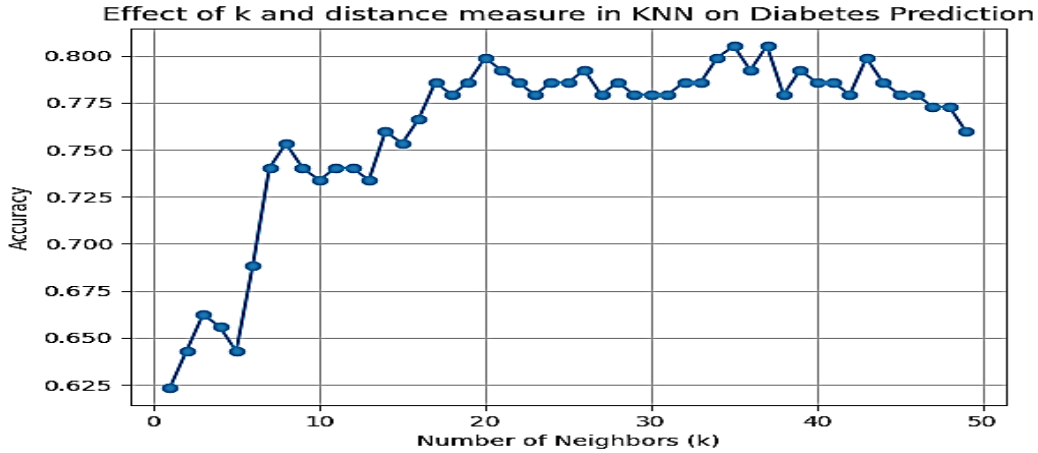

Fig. 6. The KNN performance using different k values and the Minkowski distance method

Table 2. Performance Metrics of KNN when K=35 and distance method is Euclidean

Evaluation metric	value	K_ value	Distance method
Accuracy	80.5	K = 35	Euclidean
Precision	76.5		
Recall	65.5		
F1 Score	70.6		
Specificity	34.5		
Error Rate	19.4		

Table 3. Performance Metrics of two classes of KNN when K=35 and distance method is Euclidean

class	precision	recall	f1-score	support
0	0.82	0.89	0.85	99
1	0.77	0.65	0.71	55

6. Conclusion

Diagnosis of diabetes, a chronic illness associated with abnormally high glucose levels in the blood, at an early stage, is vital for maintaining a healthy life. In this study, the performances of classifiers KNN for early diagnosis of this illness are discussed. The KNN algorithm has two important parameters, the value of K and the distance measure. There is a gap in the systematic analysis and optimizations of the KNN algorithm's parameters, as evidenced by the literature reviews cited in this study, which looked at the prediction effectiveness of the KNN algorithm for diabetes. The majority of research just looked at the ideal value for k in the KNN algorithm; they didn't discuss the best way to measure distance on its own or in conjunction with the ideal value for k to increase the accuracy of diabetes prediction. To enhance the effectiveness of the KNN methodology in predicting diabetes, this study concurrently looks into the ideal value of k and chooses the best approach for measuring distance. This study investigates the effectiveness of these parameters in predicting diabetes. The Pima dataset was used to test the performance of KNN using various *k* values and different distance measures. The four distance measures included are Euclidean, Manhattan, Chebyshev, and Minkowski distances. The results showed that the best performance achieved was 80.5% when *k*=35 and the Euclidean distance measure was used. The study emphasizes the importance of choosing a value of *k* and the distance measure depending on

the data set used and the nature of the problem. Future studies can interest with approaches of selecting the optimal value of k to maximize the accuracy of KNN to predict diabetes.

Acknowledgement

I would like to express my appreciation to the Department of Computer Science, Gujarat University, for their support for the project requirements.

References

[1] A. K. Dewangan and P. Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," Int. J. Eng. Appl. Sci. IJEAS, vol. 2, no. 5, May 2015.
[2] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data-Mining Technologies for Diabetes: A Systematic Review," J. Diabetes Sci. Technol., vol. 5, no. 6, Nov. 2011, https://doi.org/10.1177/193229681100500631.
[3] American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes -2018," Diabetes Care, vol. 41, no. Suppl.1, pp. S13–S27, 2018, https://doi.org/10.2337/dc18-S002.
[4] World Health Organization, Global Report on Diabetes. Geneva: WHO Library, 2016.
[5] S. Bano, M. Naeem, and A. Khan, "A Framework to Improve Diabetes Prediction using k-NN and SVM," Int J Comput Sci Inf Secur IJCSIS, vol. 14, no. 11, 2016.
[6] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J Big Data, vol. 6, 2019, https://doi.org/10.1186/s40537-019-0175-6.
[7] J. Gou, T. Xiong, and Y. Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule," J. Comput., vol. 6, no. 5, May 2011, doi:10.4304/jcp.6.5.833-840.
[8] P. S. Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques," Int J Sci Res Publ, vol. 7, no. 6, pp. 705–709, 2017.
[9] A. Jakka and J. Vakula Rani, "Performance evaluation of machine learning models for diabetes prediction," Int J Innov Technol Explor. Eng, vol. 8, pp. 1976–1980, 2019, https://doi.org/10.1016/j.eswa.2022.116857.
[10] Y. A. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset," IJEAT, vol. 2, no. 3, pp. 396–400, 2013.
[11] A. H. Khaleel, G. A. Al-Suhail, and B. M. Hussan, "A weighted voting of k-nearest neighbor algorithm for diabetes mellitus," Int. J. Comput. Sci. Mob. Comput., vol. 6, no. 1, pp. 43–51, 2017.
[12] I. H. Sarker, M. F. Faruque, H. Alqahtani, and A. Kalim, "K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services", EAI Endorsed Scal Inf Syst, vol. 7, no. 26, p. e4, Jan. 2020, https://doi.org/10.4108/eai.13-7-2018.162737 .
[13] S. C. Gupta and N. Goel, "Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, Aug. 2020, pp. 980–986, https://doi.org/10.1109/ICSSIT48917.2020.9214129.
[14] R. Saxena, "Role of K-nearest neighbour in detection of Diabetes Mellitus," Turk. J. Comput. Math. Educ. TURCOMAT, vol. 12, no. 10, pp. 373–376, 2021.
[15] B. V. V. Prasad, S. Gupta, N. Borah, R. Dineshkumar, H. Lautre, and B. Mouleswararao, "Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach," Prev. Med., vol. 174, p. 107619, Jul. 2023, doi: 10.1016/j.ypmed.2023.107619, https://doi.org/10.1016/j.ypmed.2023.107619.
[16] K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," IEEE Access, vol. 12, pp. 16783–16793, 2024, doi: 10.1109/ACCESS.2024.3359760, https://doi.org/10.1109/ACCESS.2024.3359760.
[17] J. Muthu and s Suriya, "Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm," J. Trends Comput. Sci. Smart Technol., vol. 5, Jun. 2023, doi: 10.36548/jtcsst.2023.2.007.
[18] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," Int. J. Comput. Sci. Trends Technol. IJCST, vol. 2, no. 4, Aug. 2014.
[19] E. M. Hameed and H. Joshi, "Performance comparison of machine learning techniques in prediction of diabetes risk," AIP Conf. Proc., vol. 3051, no. 1, p. 040002, Feb. 2024, https://doi.org/10.1063/5.0191611.
[20] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining. USA: Taylor & Francis Group, 2009.
[21] M. A. M. Khan, "Fast Distance Metric Based Data Mining Techniques Using Ptrees: K-Nearest-Neighbor Classification and k-Clustering," Master's Thesis, North Dakota State University, North Dakota, USA, 2001.
[22] N. Sambasivan and A. Ansari, Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2015.
[23] W. Yu and W. A. Zhengguo, "Fast KNN algorithm for Text Categorization," in Proc. of the 6th International Conference on Machine Learning and Cybernetics, Hong Kong, 2007, https://doi.org/10.1109/ICMLC.2007.4370742.
[24] E. M. Hameed and H. Joshi, "Current Diabetes Classification and Prediction Models Using Intelligent Techniques," in Minar Congress 6, 2022, p. 20, http://dx.doi.org/10.47832/MinarCongress6-2.
[25] S. E. Maxwell, H. D. Delaney, and K. Kelley, Designing Experiments and Analyzing Data: A Model Comparison Perspective. Routledge, 2017.