International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 19s,2025
https://theaspd.com/index.php

An Application Of K-Nearest Neighbor Using Cross-Validation Methods For Prediction Of Diabetes

Meenu Bhagat[1], Brijesh Bakariya[2]
[1,2]Department of Computer Science & Engineering, I.K. Gujral Punjab Technical University, Punjab,
Email ID : Meenu Bhagat , meenubhagat@yahoo.com
Brijesh Bakariya , brijeshmanit@gmail.com

Abstract
Diabetes is a category of long-lasting metabolic diseases that result in person's blood sugar levels remaining consistently high. This disorder can develop either because the patient's body does not create enough insulin or because the cells do not respond to insulin as expected. The three different subtypes of the disease are Type-1, Type-2 and Gestational. Patient's body cannot produce insulin when they have Type-1 diabetes. The body's cells do not utilise insulin as it should be in people with Type-2 diabetes. Pregnancy can lead to Gestational diabetes. Numerous methods are employed to study this illness. For the analysis of diabetes, we employed machine learning on 'Pima diabetes dataset' consisting of 768 records. This study made use of various Cross-validation methods using K-nearest neighbor approach. Stratified K-fold cross-validation outperformed other methods while considering accuracy and other performance parameters. Comparative analysis of results achieved using KNN and Logistic Regression through different cross validation approaches is also studied during research.

Keywords: *KNN, Diabetes, Cross validation, Logistic Regression, Stratified K Fold, Train-Test Split.*

1. INTRODUCTION

Diabetes is a health condition that occurs when the pancreas does not produce enough insulin or when the body does not properly absorb insulin. Insulin is a hormone essential for regulating blood sugar levels. Uncontrolled diabetes can lead to hyperglycemia, which is commonly known as high blood sugar, and this condition can progressively damage numerous body systems. There are 422 million individuals with diabetes living in low- and middle-income countries, and the disease is responsible for approximately 1.5 million deaths annually[1]. The World Health Organization (WHO) indicates that 8.5% of individuals aged 18 and older are affected by diabetes, leading to around 1.6 million deaths globally [2]. While the rate of early mortality due to diabetes decreased in some developing nations between 2000 and 2010, it rose again from 2010 to 2016.

2. LITERATURE SURVEY

Machine learning techniques including Random Forest, K-nearest neighbor, Naïve Bayes and Support vector machines (SVM) were utilised by Sudharsan B et al. [3] and Georga et al. [4] worked on Support vector regression to predict hypoglycemia in Type 2 diabetes patients. Numerous machine-learning techniques were employed by Weifeng Xu et al. [5] in order to predict the occurrence of diabetes. Random Forest turned out to be more accurate in comparison with other data mining methods. In many different fields, including healthcare, cross-validation techniques are widely applied for forecasting disease outcomes, detecting risk factors, and optimising treatment plans. Cross-validation is a useful technique in finance for portfolio optimisation, credit risk assessment, and stock price forecasting. Moreover, cross-validation makes machine translation, sentiment analysis, and text classification easier in natural language processing. Support vector machines surpassed the other algorithms in terms of performance and accuracy, according to Kavakiotis et al. [6], who utilised 10-fold cross-validation as an evaluation method for three machine learning algorithms: Logistic regression, Naïve Bayes and Support vector machines.

In Train-Test Split [7] approach, the original dataset is split into the Train set and Test set. The classifier is trained using the train set, and its accuracy is tested using the test set. This method has the disadvantage of using a big amount of data for testing, which in certain cases may only represent one particular sort of data. For instance, a certain age group, location, economic level, etc. Dataset D is split into k distinct parts in an equal number in a typical (K-fold) cross-validation technique. The classifier is trained using the first k-1 folds of a particular K-fold dataset, and its effectiveness is tested using the one-fold. Cross-validation has an expanded form called Stratified cross-validation [8]. In this type a class is distributed uniformly over n folds in this dataset, meaning that the class distribution in each fold matches that of the original dataset. In K-fold cross-validation, a particular class may be distributed unevenly, with particular folds containing a larger number of instances of that class than others. K-nearest neighbors (KNN) is a supervised learning algorithm that can be utilised to resolve classification and regression issues. While dealing classification challenges, it will locate the k closest neighbors and determine the class based on majority vote of the neighbors. For regression issues, it will identify the k closest neighbors and forecast the value using the mean of the closest neighbors values [9]. KNN showed an accuracy of 70.35% when Bavkar et al. [10] evaluated diabetes prediction using SVM, Decision Tree, Naïve Bayes and KNN on the Pima diabetic dataset.

3. PROPOSED MODEL

Our datasets came from Kaggle [11]. There are 768 samples in the database. 500 out of total samples are negative class instances and 268 samples are positive class instances. The suggested method analyses the use of KNN on diabetic database using Train test Split and other Cross-validation approaches. General steps to be followed are shown below [Fig. 1].

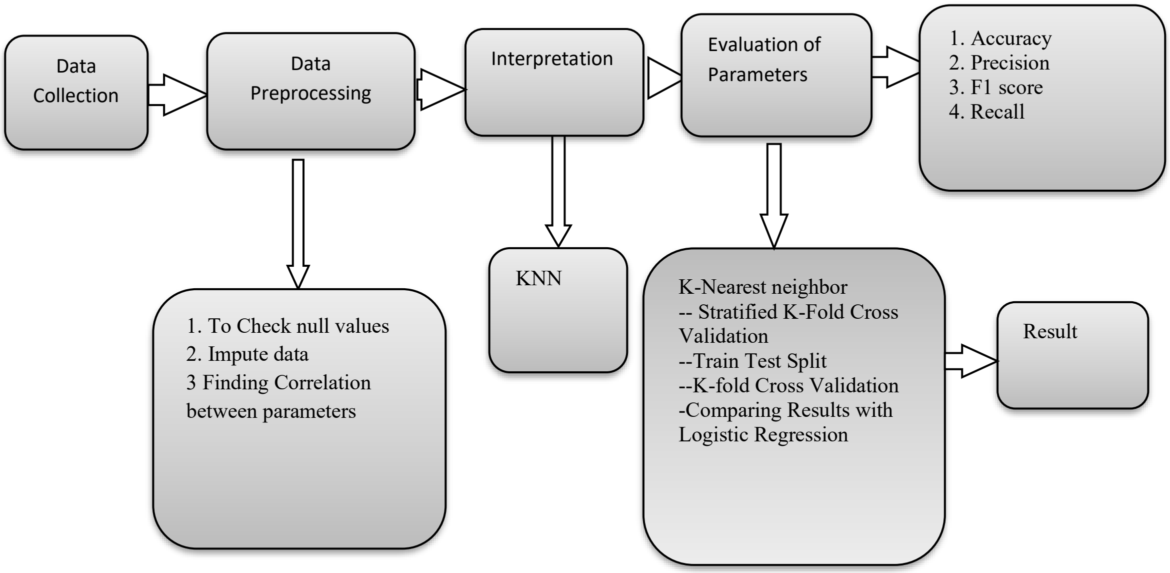

Fig 1: General Process

1. We looked for any null values in the database.
2. The database has been imputed.
3. We used K-Fold, Stratified K-Fold and Train-Test Split to train our model using K-Nearest neighbor algorithm [Table 1].

4. RESULTS

Table 1: Values of Precision, Recall and F1-score and Mean accuracy using K-Fold, Stratified K-Fold and Train-Test Split Method considering all parameters.

Method	Precision		Recall		F1 Score		Support		Accuracy
	0	1	0	1	0	1	0	1	
Train Test Split	0.77	0.61	0.81	0.54	0.79	0.57	151	80	71.86%
Stratified K Fold	0.85	0.82	0.92	0.69	0.88	0.75	50	26	75.54%
K-Fold	0.74	0.91	0.95	0.59	0.83	0.71	21	17	75.31%

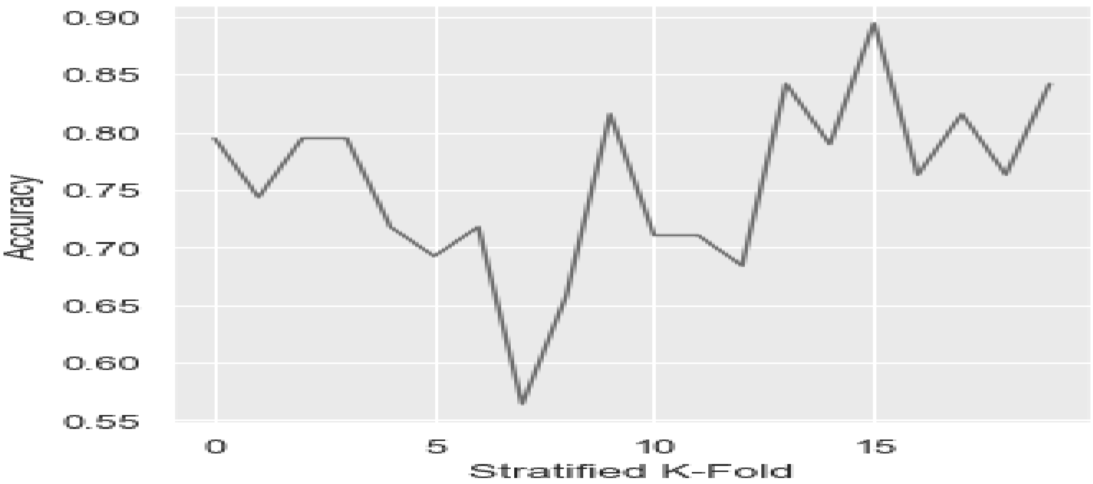

Fig 2: Model accuracy achieved during different folds in Stratified K-Fold method using KNN.

Values of accuracies obtained using the Stratified K Fold technique are depicted in Figure 2. At n-splits=20, The mean accuracy of Stratified K-Fold approach is greater i.e.,75.54% than Train-test Split method (71.86%). The maximum and minimum accuracies in different folds using Stratified K-Fold cross validation has been observed as 89.47% and 56.41% respectively.

The mean accuracy using KNN is highest at K=20 (75.31%). The maximum and minimum accuracies in various folds using K Fold cross validation in KNN has been observed as 92.10% and 56.41% respectively. Using Logistic Regression with Train-test split and Stratified K-Fold approaches, it has been analysed that model accuracies are 75.32% and 76.3% respectively [12].

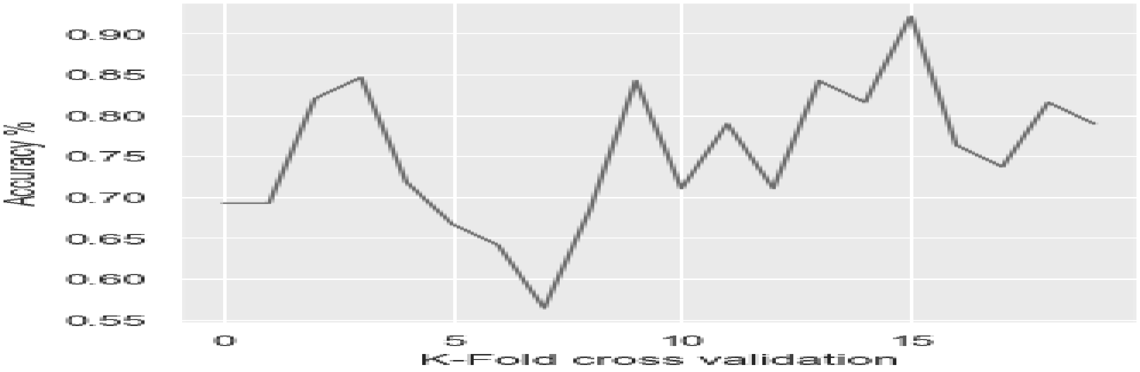

Fig 3: Values of accuracies during K- Fold Cross validation in different folds using KNN.

5. CONCLUSIONS

We have used the KNN algorithm for prediction of diabetes using 'Pima Indian diabetes dataset' for this study and it can be further extended for the prediction of other diseases. The impact of cross-validation strategies like Train-test split, K-fold cross-validation and Stratified K-fold approach using KNN is studied in this research. Stratified K-fold cross validation surpassed in terms of accuracy, Precision, Recall and F1 score than the other methods used. The effect of other cross validation strategies like Repeated K Fold, Leave one out, Leave P out, Shuffle & Split on model accuracy can be included for further research. This research can also be implemented on other datasets while using different machine learning algorithms e.g. Random Forest Classifier, Naive Bayes, Decision Tree, SVM etc.

REFERENCES:

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed: 2023-06-09.

[2] https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed: 2021-04-20.

[3] Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol. 2015 Jan;9(1):86-90. doi: 10.1177/1932296814554260. Epub 2014 Oct 14.

[4] Georga EI, Protopappas VC, Ardigò D, Polyzos D, Fotiadis DI. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. Diabetes Technol Ther. 2013 Aug;15(8):634-43. doi: 10.1089/dia.2012.0285. Epub 2013 Jul 13.

[5] Xu W, Zhang J, Zhang Q, and Wei X "Risk prediction of type II diabetes based on random forest model," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017, pp. 382-386, doi: 10.1109/AEEICB.2017.7972337.

[6] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017 Jan 8; 15:104-116. doi: 10.1016/j.csbj.2016.12.005. PMID: 28138367; PMCID: PMC5257026.

[7] Zeng X, & Martinez T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1), 1– 12. doi: 10.1080/095281300146272.

[8] Breiman L, Friedman J. H., Olshen R. A. and Stone C. J., 1984, Classification and Regression Trees (Wadsworth International Group).

[9] https://towardsdatascience.com/an-introduction-to-k-nearest-neighbours-algorithm-3ddc99883acd. 2023-06-12.

[10] Bavkar VC, Shinde AA (2021) Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement. Indian Journal of Science and Technology 14(10): 869-880. https://doi.org/ 10.17485/IJST/v14i10.2187.

[11] Kaggle.com. 'Pima Indians Diabetes Data Set'. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database [Accessed: 07- June- 2020].

[12] Bhagat, M., Bakariya, B. "Implementation of Logistic Regression on Diabetic Dataset using Train-Test Split, K-Fold and Stratified K-Fold Approach". Natl. Acad. Sci. Lett. 45, 401–404 (2022). https://doi.org/10.1007/s40009-022-01131-9.