Natl. Acad. Sci. Lett. (September–October 2022) 45(5):401–404
https://doi.org/10.1007/s40009-022-01131-9

SHORT COMMUNICATION

Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach

Meenu Bhagat[1] · **Brijesh Bakariya**[1]

Received: 28 October 2021 / Revised: 10 May 2022 / Accepted: 17 May 2022 / Published online: 8 July 2022

Abstract Diabetes is a chronic metabolic disorder causing high blood sugars, that further severely affect body parts like the heart, liver, kidneys, lungs, eyes, nerves, blood vessels etc. There are three types of diabetes- Type-1 Diabetes, Type-2 Diabetes, and Gestational Diabetes. In Type-1, body of the patient fails to produce insulin. In Type-2 diabetes, cells of the body fails to respond to insulin effectively. Gestational diabetes occurs during pregnancy. There are many approaches used to analyse this disease. We have used the Machine learning approach for analysing diabetes. We have used 768 records from "pima diabetes dataset". In this paper, we have used Logistic regression with Train Test Split, K-Fold cross-validation and Stratified K-Fold approach.

Keywords Diabetes · Logistic Regression · Machine Learning · Train-test split · K-Fold · Stratified K-Fold

Significance Statement: In this paper, the proposed approach analyses implementation of Train test Split, K-Fold, and Stratified K-Fold cross-validation techniques while using Logistic Regression on Diabetic Database.

✉ Meenu Bhagat
meenubhagat@yahoo.com

Brijesh Bakariya
dr.brijeshbakariya@ptu.ac.in

[1] Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University, Kapurthala, Punjab, India

Introduction

Diabetes can be majorly categorized into three types: Type 1 diabetes, Type 2 diabetes, and Gestational diabetes. Type 1 diabetes: In this, our immune system destroys all the beta cells in our pancreas. Beta cells are insulin-making cells in our pancreas. Due to lack of insulin glucose from our food is not transferred to our cells leading to many short-term and long-term problems. In Type 2 diabetes, our body becomes insulin resistant resulting starving of cells and excess glucose remains in our bloodstream. Gestational diabetes is a condition experienced during pregnancy. High blood glucose levels can be caused by a combination of hormones and increased insulin content during pregnancy. The chances of developing diabetes in newly born babies is also high [1]. A variety of factors are believed to play a role in the onset and progression of diabetes. Given the clear causal association between obesity and the onset of diabetes [2], obesity is a major risk factor, especially in Type 2 diabetes. Sudharsan B et al. [3] used machine learning methods such as Random Forest, Support vector machines (SVM), K-nearest neighbour, and Naive Bayes to predict Hypoglycaemia among Type 2 diabetes patients, while Georga et al. [4] used Support vector regression for the same purpose. Train-Test Split [5] is a typical strategy where we divide the original dataset into two parts, i.e. Train set and Test set. The Train set is used for training the classifier and the Test set is used to find the accuracy of the classifier. The drawback of this method is that a large amount of dataset is used for testing and in certain situations; the dataset may represent only a specific kind of data. For example a certain age group, a certain city, or a certain income group etc.

Cross-Validation: In a typical (K-Fold) cross-validation method, a dataset D is equally partitioned into k disjoint

subsets. In a particular K-Fold dataset first K-Folds are used for training the classifier and the remaining *k*-1 folds are used for testing. Stratified Cross-Validation is the extended form of cross-validation [6]. In this uniform, distribution of a class is done among n number of folds so the distribution of a class in each fold of dataset is the same as present in the original dataset. Regular cross-validation, on the other hand, arbitrarily partitions S into n folds without taking class distributions into account. K-Fold cross-validation could result in a certain class being distributed unevenly, with some folds containing more cases of the class than others. D.Kohavi [7] has done a comparison of many accuracy estimation techniques and he found that the cross-validation performs better than other techniques and further stratification improve the performance by lowering the bias and variance. Weifeng Xu et al. [8] used a variety of machine learning algorithms to predict diabetes diseases. As a result of these algorithms, Random Forest was found to be more accurate than other data mining techniques. According to Kavakiotis et al. [9] tenfold cross-validation was used as an evaluation method in three different algorithms, i.e. logistic regression, Support vector machines and Naive Bayes and in terms of accuracy and performance, Support vector machines outperformed the other two algorithms. We have taken our datasets (Table 1) from Kaggle [10]. On PIDD, Sisodia et al. [11] discovered that the NB classifier outperforms the SVM, NB, and DT machine learning algorithms, with an accuracy of 76.30 percent. All patient's data were trained and tested using 10 cross-validations with Naive Bayes and decision trees in Amour Diwani et al.'s study [12]. The best algorithm, according to their results was Naive Bayes with a 76.3021% accuracy. Using different classifiers such as Decision Tree, SVM, KNN, RF, and NB, Sneha and Gangil [13] proposed a model for the early detection of diabetes. SVM ranks first among these classifiers with 77.33% accuracy. Aishwarya Jakka and Vakula Rani [14] suggested a performance evaluation approach based on decision-making classifiers. LR, SVM, KNN, RF, and NB are some of the algorithms used. LR has the highest accuracy of 77.60% among these classifiers. The database contains 768 samples. Out of which 500 samples are positive class instances, i.e."1″ and 268 samples were negative class instances, i.e. "0″. Following are the feature of this dataset:

Table 1 Features of PIMA dataset

Sr. No	Features
1)	Number of Pregnancies(NOP)
2)	Plasma Glucose Concentration Within 2 Hours(PGC)
3)	Diastolic Blood Pressure(DBP)
4)	Triceps of Skin Fold Thickness
5)	Serum Insulin Within 2 h
6)	Body Mass Index
7)	Diabetes Pedigree Function
8)	Age
9)	Outcome

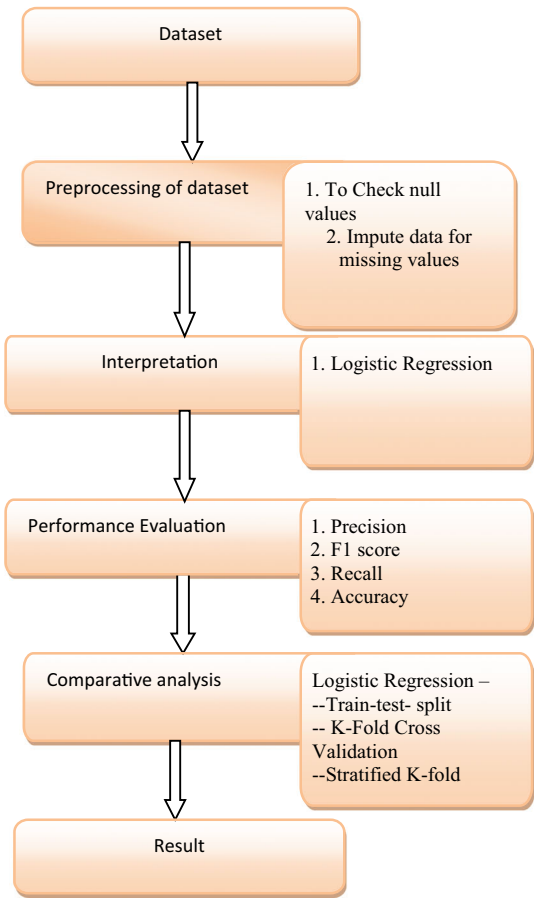

Fig. 1 General Process

Figure 1 is showing the general steps to be followed.

1. We have checked the database for null values.
2. We have imputed the database with the mean or median of the columns that have zero values.
3. For validation, we used Train Test Split, K-Fold cross-validation, and Stratified K-Fold.

We have trained our model with Logistic regression using Train Test Split, K-Fold, and Stratified K-Fold and Table 2 is showing the Precision, Recall and F-score values taking all the eight input parameters using Train-Test split Method.

Figure 2 depicts the accuracy values by using the Stratified K-Fold method. It has been noticed that accuracy in case of Stratified K-Fold at *n*-splits = 10 is more than the Train test Split method. The Model has been tested for

Table 2 Precision, Recall and F1-score, Support and Accuracy values using Train Test split Method and Stratified K-Fold considering all parameters

Method used	Precision		Recall		F1 Score		Support		Accuracy
	0	1	0	1	0	1	0	1	
Train Test Split	79%	66%	84%	82%	62%	59%	151	80	75.32%
Stratified K-Fold	82%	84%	94%	88%	71%	62%	50	26	76.3%

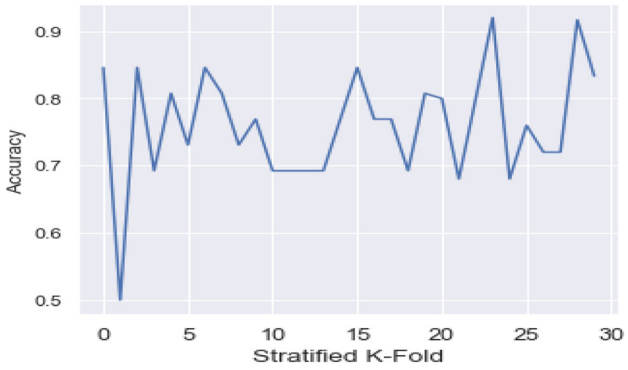

Fig. 2 Accuracy in different folds in Stratified K-Fold method

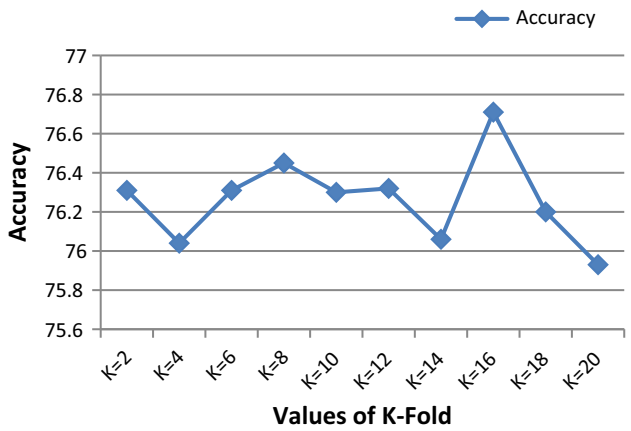

Fig. 3 Accuracy Values using K- Fold Cross-validation for $K = 2$–$K = 20$

different K-Folds (K = 2,4,6,8,10,12,14,16,18,20) (Fig. 3), and it has been observed that mean accuracy was maximum (76.71%) at $K = 16$.

In this paper, we have worked on the diabetes dataset. This work can also be extended for prediction of other diseases also. We have only used Logistic Regression for this study. Other machine learning classifiers like Naïve Bayes, Random Forest Classifier, and KNN can be used for research purposes. Train test split, K-Fold Cross-Validation and Stratified K-Fold methods are used in this research. This work can be extended using different type of datasets with different machine learning algorithms.

References

1. Anna V, van der Ploeg HP, Cheung NW, Huxley RR, Bauman AE (2008) Socio-demographic correlates of the increasing trend in prevalence of gestational diabetes mellitus in a large population of women between 1995 and 2005. Diabetes Care 31(12):2288–2293. https://doi.org/10.2337/dc08-1038
2. Després JP, Lemieux I (2006) Abdominal obesity and metabolic syndrome. Nature 444(7121):881–887. https://doi.org/10.1038/nature05488
3. Sudharsan B, Peeples M, Shomali M (2015) Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol 9(1):86–90. https://doi.org/10.1177/1932296814554260
4. Georga EI, Protopappas VC, Ardigò D, Polyzos D, Fotiadis DI (2013) A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. Diabetes Technol Ther 15(8):634–643. https://doi.org/10.1089/dia.2012.0285
5. Zeng X, Martinez TR (2000) Distribution-balanced stratified cross-validation for accuracy estimation. J Exp Theor Artif Intell 12(1):1–12. https://doi.org/10.1080/095281300146272
6. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984), Classification and regression trees (Wadsworth International Group).
7. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the international joint conference on artificial intelligence (IJCAI), 1137–1143.
8. Xu W, Zhang J, Zhang Q, Wei X (2017) 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). https://doi.org/10.1109/AEEICB.2017.7972337
9. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 8(15):104–116. https://doi.org/10.1016/j.csbj.2016.12.005
10. Kaggle.com. 'Pima Indians diabetes data set' (**Online**). https://www.kaggle.com/uciml/pima-indians-diabetes-database. Accessed 7 June 2020.
11. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585
12. Diwani SA, Sam AE (2014) Diabetes forecasting using supervised learning techniques. Adv Comput Sci: Int J 3:10–18

13. Sneha and Gangil (2019) Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6:13. https://doi.org/10.1186/s40537-019-0175-6
14. Jakka A, Vakula-Rani J (2019) Performance evaluation of machine learning models for diabetes prediction. IJITEE. https://doi.org/10.35940/ijitee.K2155.0981119