ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

Prediction of Diabetes Mellitus through Decision Tree Classifier using K-Fold Cross Validation and Train Test Split

[1] Meenu Bhagat, [2]Brijesh Bakariya

[1][2] Department of Computer Science & Engineering, I.K. Gujral Punjab Technical University, Punjab, India
[1] meenubhagat@yahoo.com, [2] dr.brijeshbakariya@ptu.ac.in

***Abstract*— One of the many chronic diseases that affect the elderly globally is diabetes. This condition develops when the blood sugar or glucose level is extremely high. Serious health issues can be avoided with early diabetes testing and treatment. The current methods for measuring blood glucose are intrusive, which makes patients uncomfortable. In this study, a dataset constructed from samples from the PIMA Indian Diabetes dataset is used to predict diabetes using a machine learning technique. Decision Tree utilising K Fold cross validation and Train Test split are the machine learning techniques utilised in this study. Using Decision Tree and Logistic Regression, comparative study of model accuracies and other performance parameters (Precision, Recall, F1 score) is also investigated.**

***Index Terms*— Diabetes, Decision Tree, Logistic Regression, Machine Learning, K-Fold, Train-test split.**

I. INTRODUCTION

Diabetes is a typical chronic condition that develops when the pancreas is unable to produce enough insulin (Type 1 diabetes) or when the patient's body does not use the insulin properly (Type 2 diabetes). Uncontrolled diabetes frequently results in hyperglycaemia, or elevated blood sugar. Diabetes can seriously harm blood vessels and nerves over time [1]. According to the National Diabetes Statistics Report 2020, there are 34.2 million Americans who have diabetes, representing 10.5% of the country's population. In contrast to the 26.9 million persons with diabetes who have been given a diagnosis, 7.3 million people do not know they have the disease, which represents a 21.4% prevalence [2]. India, which was classified as the second-highest country in the world for the percentage of diabetics, had almost 77 million new cases of the disease in 2019 [3]. Cross validation is a method for evaluating a model's performance and comparing various models side by side. On the same set of data, for instance, we can compare the performance of a support vector machine (SVM) with a K-nearest neighbours (KNN) model. Machine Learning classifiers were used in this study to carry out various cross-validations. The model was validated using the k-fold CV method. To undertake training with test data, the PIDD was divided into 'k' folds. The remaining 'k-1' folds were then joined to create trained data. Original data were divided into 'k' folds (k1, k2,..., ki) at random, and the model was tested 'k' times[4].

II. LITERATURE SURVEY

Type 1 diabetes, Type 2 diabetes, and gestational diabetes can be broadly categorised as the three main kinds of diabetes. Type 1 diabetes: Each and every beta cell in our pancreas is destroyed by this immunological response. Our pancreas' beta cells are the ones that produce insulin. Insufficient insulin prevents the transport of glucose from our food to our cells, which causes a variety of both immediate and long-term issues. Our bodies become less responsive to insulin in Type 2 Diabetes, starving our cells and leaving extra glucose in our bloodstream. An illness that occurs during pregnancy is gestational diabetes. The combination of hormones and more insulin produced during pregnancy can result in high blood sugar levels. Diabetes is also a disease with a significant risk of occurrence in newborns [5].

Razavian et al. (6) developed prediction models based on logistic regression for various onsets of type 2 diabetes prediction in order to deal with the high dimensional datasets. Support vector regression (SVR) was utilised by Georga et al. (7) to predict diabetes, which is a multivariate regression problem, with an emphasis on glucose.

Cross-validation outperforms other strategies, and further stratification improves performance by reducing bias and variance, according to D.Kohavi's [8] assessment of numerous accuracy estimation techniques.

Support Vector Machines beat out Naive Bayes and logistic regression in terms of accuracy and performance, according to Kavakiotis et al. [9], who implemented 10-fold cross validation as an evaluation technique for these three algorithms. To forecast diabetes illnesses, Weifeng Xu et al. [10] used a range of machine learning techniques. These algorithms led to the discovery that Random Forest was more accurate than other data mining methods.

III. DATA PREPROCESSING

3.1 Data Collection

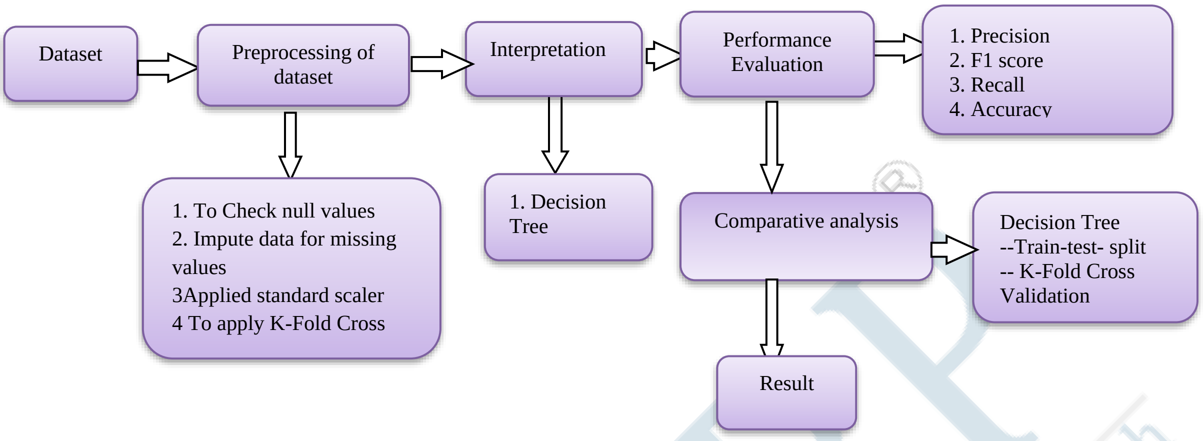

Fig 1: General Process

IV. ALGORITHMS

A fundamental classification and regression technique is the decision tree. The classification of instances based on features can be described using a decision tree model, which has a tree-like structure[11]. It can be viewed as a collection of if-then rules, or as conditional probability distributions that are specified in feature space and class space. Classification is used when the features are grouped, and regression is used when the data is continuous. The entropy of every characteristic is initially calculated by the Decision Tree algorithm. After that, the dataset is divided into groups according to the variables or predictors with the greatest information gain or lowest entropy.

V. RESULTS

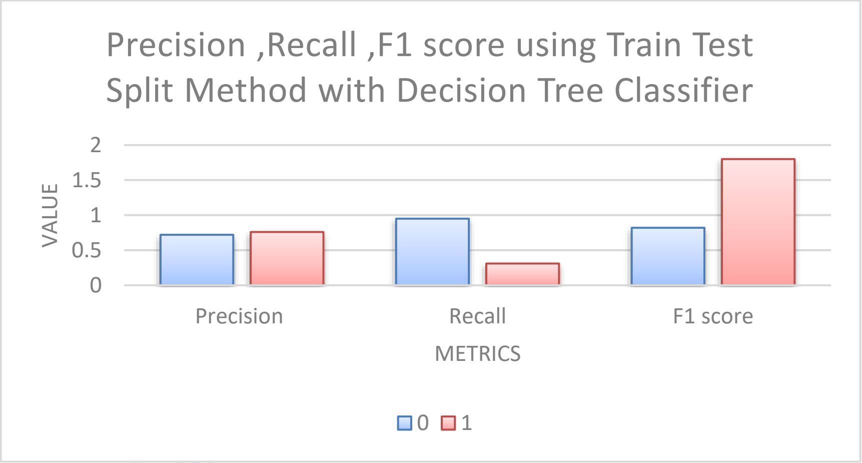

Fig .2. Values of Precision, Recall and F1 score of Model using Decision Tree Classifier with Train Test split.

Figures 2 and 3 display the Precision, Recall, and F1 score values obtained by applying K Fold cross validation and Train Test split in Decision Tree and Logistic Regression models.

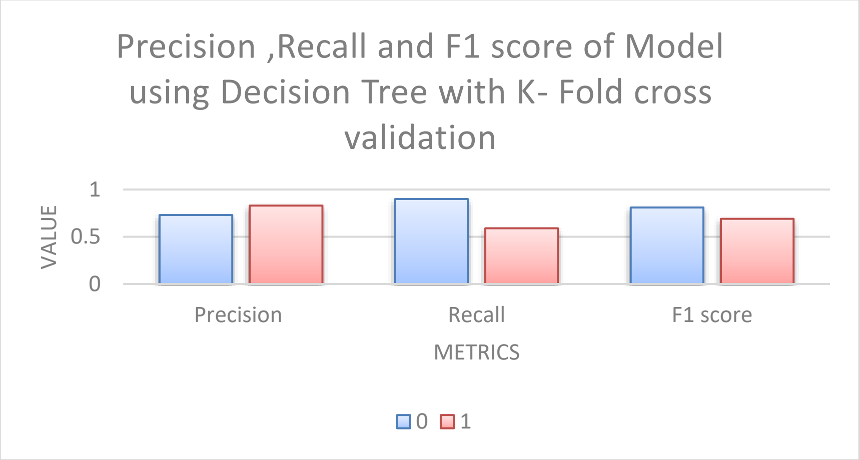

Fig .3. Values of Precision, Recall and F1-score of Model using Decision Tree Classifier with K-Fold Cross Validation.

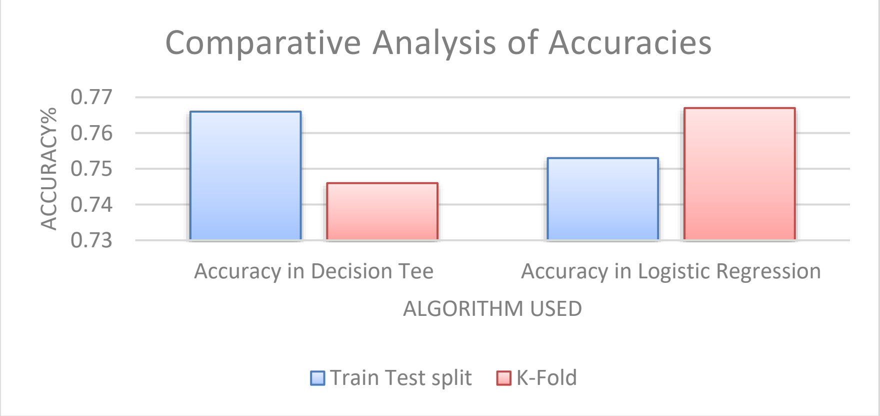

Fig .4. Values of accuracies in Decision Tree Classifier and Logistic Regression

The accuracy values obtained using the K Fold cross-validation and Train Test split using Decision Tree and Logistic Regression are shown in Figure 4. At n-splits=10, it has been observed that the accuracy of the K-Fold approach is greater than that of the Train test Split method in Logistic Regression[g].The accuracy of Model using Decision tree with Train Test split was greater than K Fold cross validation at n-splits=20.

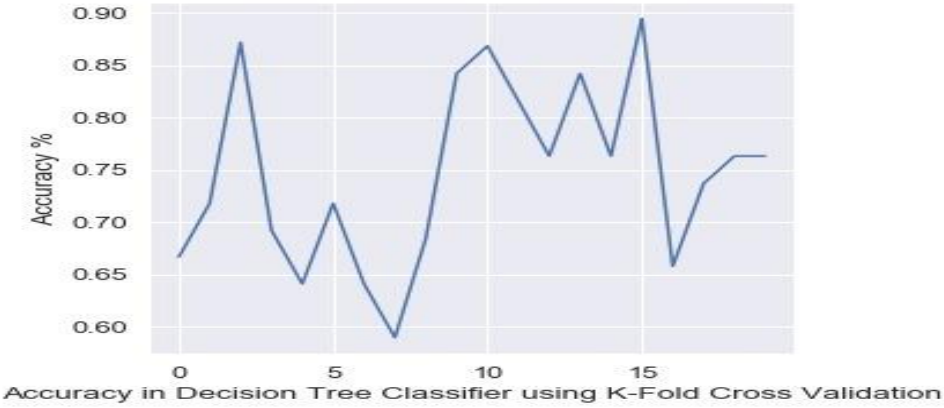

Fig .5. Values of accuracies in Decision Tree Classifier using K Fold Validation

The Decision Tree model was tested for various K-folds (Fig 5), and it was observed that the mean accuracy was 74.65% at n splits=20.The maximum and minimum accuracies achieved using the K-Fold cross validation are 0.8947% and 0.589% respectively.

VI. Conclusion

We used the dataset for diabetes in this paper. This research could be expanded to include disease prediction for other conditions using various other cross-validation approaches. For this research, Decision Tree with Tain test split and K-Fold cross-validation has been used. For

Further study, Naive Bayes, Random Forest Classifier, and KNN are few examples of machine learning classifiers that can be employed. Using various dataset types and machine learning algorithms, this research can be expanded.

REFERENCES

[1] Seshasai, S.R.K.; Kaptoge, S.; Thompson, A.; Di Angelantonio, E.; Gao, P.; Sarwar, N.; Whincup, P.H.; Mukamal, K.J.; Gillum, R.F.; Holme, I.; et al. Diabetes mellitus, fasting glucose, and risk of cause-specific death. N. Engl. J. Med. 2011, 364, 829–841.

[2] C.f.D. Control, Prevention, National Diabetes Statistics Report, 2020, Centers for Disease Control and Prevention, US Department of Health and Human Services, Atlanta, GA, 2020, pp. 12–15.

[3] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A.A. Motala, K. Ogurtsova, Global and regional diabetes preva lence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, Diabetes Res. Clin. Pract. 157 (2019) 107843.

[4] Battineni, G.; Sagaro, G.G.; Nalini, C.; Amenta, F.; Tayebati, S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines 2019, 7, 74. https://doi.org/10.3390/machines7040074.

[5] Anna V, van der Ploeg HP, Cheung NW, Huxley RR, Bauman AE. 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. 2008; 31(12):2288–93. doi: 10.2337/dc08-1038.

[6] Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., and Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 3, 277–287. doi: 10.1089/big.2015. 0020.

[7] Georga, E. I., Protopappas, V. C., Ardigo, D., Marina, M., Zavaroni, I., Polyzos, D., et al. (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J. Biomed. Health Inform. 17, 71–81. doi: 10.1109/TITB.2012.2219876.

[8] 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.

[9] 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.

[10] 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.

[11] Quinlan, J.R. Induction of decision trees. Mach Learn 1, 81–106 (1986). https://doi.org/10.1007/BF00116251.

[12] Meenu Bhagat, and Brijesh Bakariya. "Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach" National Academy science letters 45, no. 5 (2022): 401-404. doi: 10.1007/s40009-022-01131-9.