Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

doi:10.3233/APC210294

Diabetes Risk Forecasting Using Logistic Regression

Metharani N [a,1], Srividya R [a], Rekha G [b] and Ranjith Kumar V [c]
[a] *Department of CSE, Adhiparasakthi Engineering College, Melmaruvathur, India*
[b] *Department of CSE, St Peter's Institute of Higher Education & Research, Chennai, India*
[c] *Department of Mechanical Engineering, Sri Sairam Engineering College, Chennai, India*

Abstract. Diabetes can be a collection of metabolic problems and lots of human beings are affected. Diabetes Mellitus can be caused by a variety of factors including age, stoopedness, lack of activity, inherited diabetes, lifestyle, poor eating habits, hypertension, and so on. Diabetics are more likely to develop diseases like coronary illness, kidney contamination, eye sickness, stroke and other risks. Distributed computing and Internet of Things (IoT) are two instruments that assume a vital part in the present life with respect to numerous angles and purposes including medical care observing of patients and old society. Diabetes Healthcare Monitoring Services are vital these days on the grounds that and that to distant medical care observing in light of the fact that truly going to clinics and remaining in a line is exceptionally ineffectual adaptation of patient checking. Current practice in emergency clinic is to gather required data for diabetes conclusion through different tests and proper treatment is given dependent on analysis. Utilizing enormous data investigation can consider large datasets and discover covered up data, uncertain examples to find information from the data and expect the outcome as demand. Diabetics are caused because of a tremendous uphill in the blood partition containing glucose. There is an advancement conspire accessible using train test split and K overlay cross approval utilizing Scikit learn technique. Various ML algorithms consisting of SVM, RF, KNN, NB, Decision Tree and Logistic Regression are also used.

Keywords. Indian Pima Diabetes Dataset, Decision Tree, K Nearest Neighbor; Random Forest, Logistic Regression, Naïve Bayes

1. Introduction

This paper needs some exacting activities in regards to the control and counteraction of diabetes. Prior the facts previously demonstrated that around one in each ten individuals in US had diabetes. In any case, expectations have been made that by 2045 it might help up to one in each three people. This is a difficult issue we need to manage. The constant sickness of diabetes results right into it when there is an immense expansion in the blood glucose focus. This is a significant reason for different issues and infections, for example, kidney illnesses heart issues. Numerous unfortunate dietary patterns and absence of appropriate body practices likewise causes the diabetic

[1] Metharani N, Department of CSE, Adhiparasakthi Engineering College, Melmaruvathur, TN, India; E-mail: metharani53@gmail.com.

pre conduct. It has been expressed by the WHO that the absolute check of individuals experiencing diabetes has inconceivably expanded in the course of recent years.

Managing numerous diabetic datasets is obligatory to improve the current pace of patients experiencing diabetes and to take it to an unadulterated insignificant level by zeroing in on to lessening it by huge scale. Type 1, Type 2, and gestational diabetes are the most well-known forms of diabetes.

Type 1 diabetes: It is indeed a type of diabetes for which the body your body does not produce insulin if you have type 1 diabetes. It is very often evaluated in kids and teens, spite of the fact that it can show at whatever age.

Type 2 diabetes: You can create type 2 diabetes at whatever stage in life, in any event, during adolescence.

Gestational diabetes: Gestational diabetes creates in certain ladies when they are pregnant. More often than not, this sort of diabetes disappears after the child is conceived. Notwithstanding, on the off chance that you've had gestational diabetes, you have a more prominent possibility of creating type 2 diabetes sometime down the road. The patient risk level is classified using data mining techniques such as K nearest neighbor, Decision tree, Random forest, Logistic Regression and Naïve Bayes.

2. Literature Review

Defusal Faruque and Asaduzzaman, Iqbal H.Sarker mentioned that polygenic disease is one among the foremost common disorder of the material body it's caused due the metabolic disorder. They used various and important ML algorithms that are Support Vector machine, NB, KNN and DT to predict diabetes [1]. Sidong Wei, Xuejiao Zhao and Chunyan Miao presented that diabetes is mostly called as disorder in which glucose level in the body is high. In this paper they use popular methods such as SVM and a deep neural network for identify the disease and data processing [2]. Jian-xunChen, Shih-LiSu and Che-Ha Chang discussed about Ontology that generate a primary care planning to the medical professionals. The result of the research paper shows the model can be provided personalize diabetes mellitus care planning efficiently [3]. MM Alotaib, RSH.Istepanian, and A.Sungoor they are present a clever based mobile polygenic disease control system & tutoring model for the patients with diabetes. In this, a system is able to store the clinical information about the diabetes system, such an often blood sugar level and BP is measured and hypo glycaemia event [4]. Berina Alic and Lejila Gurbea,Almir Badnjevic they presented the overview of techniques in the machine learning in the diabetes classification and cardiovascular diseases using BNs and ANN [5]. ElliotB.Sloane, Nilmini Wickramasingle and Steve Goldberg they presented Wireless diabetes monitoring which is a cloud-based diabetes, it’s a coaching platform for diabetes management and low cost, innovative, cloud-based diabetes support system . Minyechil Alehegn and Rahul Joshi had present about the ML technology that help to identify a dataset at the elementary so that rescue the life.By implementing NB and K-nn algorithms . Umatejaswi and P.Suresh Kumar had discussed about algorithms such as SVM, NB, DT for identify the mellitus make use of technique like data mining .

3. Methodology

In this paper, we have utilized our dataset for applying different machine learning algorithms for recognizing if an individual has coronary illness or not is shown in Figure 1. At that point, we will deal with the missing qualities in the dataset, visualize the dataset and notice the precision acquired by various AI calculations. AI calculations utilized are characterized underneath.

4. Result

4.1 Correlation Matrix

From this Figure 1. diagram, we can see that a few features are exceptionally related and some are definitely not.

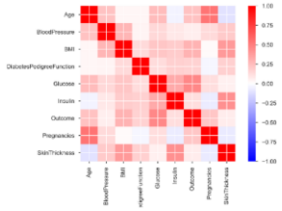

Figure 1. Correlation matrix

4.2 Histogram

A histogram is a statistical tool for the portrayal of the appropriation of the data set. It is an overall assessment of the likelihood conveyance of a persistent arrangement of variable information is shown in Figure 2. It is really a plot that answers all the inquiries with the fundamental recurrence appropriation of a bunch of nonstop and plausible information, it gives a feeling of the thickness of information.

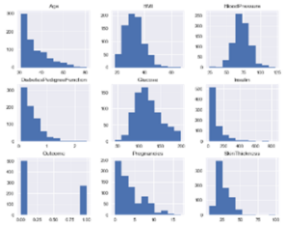

Figure 2. Histogram.

Outcome" is that the features in Figure 3.we have a tendency to be attending to predict, zero suggests that No diabetes, one means diabetes. Of these 768 knowledge points, 500 are tagged as zero and 268 as 1:

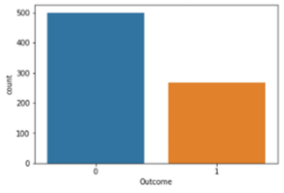

Figure 3. Count of Outcome variable

Classifying target variable between male and female and visualize the result is shown in Figure 4.

Box plot for target class with different features

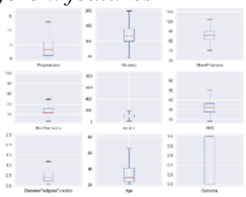

Figure 4. Target Class Box plot

Machine Learning Algorithms

4.3. Logistic Regression

The logistic regression, like all regression analyses, is a predictive analysis. To describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is used. In this paper, we achieved the accuracy of 73% by using this model.

4.4. Naive Bayes Classifier:

In short, the naive Bayes classifier expects that the presence of a specific capacity in one class has nothing to do with the presence of another capacity. Naive Bayes version is straightforward to construct and especially beneficial for extremely massive information sets. In this paper, we achieved the accuracy of 73% by using this model.

4.5. KNN Classifier:

The most straightforward AI calculation envisioned so far till now is the K-NN calculation because of its high exactness. The littlest reversible substances that make up the essential little units to work together and join to address a strong construction is this information that is thoroughly machine autonomous in all fields. The real factors and measurements those are altogether together consolidated with the end goal of reference and investigation is crude information objects. It is the fundamental methods for estimation and framework thinking for standardization purposes. The starting technique comes as the assortment of the information required and further articulating the given issue proclamation which characterizes the necessary information for framework examination. We should control the wonder of social event or probably joining the information and data from broadly accessible various sources. Presently our principle task comes as information preprocessing where three primary undertakings are played out those are organizing, cleaning and inspecting. In this paper, we achieved the accuracy of 78% by using this classifier.

4.6. Decision Tree Classifier

It learns to partition based on the value of an attribute. Recursive partitioning is a process of partitioning the tree in a recursive manner. This image is a framework. In this paper, we achieved the accuracy of 76% by using this model.

4.7. Random Forest Classifier

It is simple and easy to implement. A forest is comprised of trees. This classifier creates selection timber on randomly decided on records samples, receives prediction from every tree and selects the first-class answer with the aid of voting. The random wooded area composed of multiple selection timber. It creates a forest of trees. In this paper, we achieved the accuracy of 71% by using this model.

Table 1. Accuracy Values

Algorithms	Accuracy
Logistic Regression	73%
Naïve Bayes Classifier	73%
Decision Tree Classifier	76%
Random Forest Classifier	71%
KNN Classifier	78%

In Table 1, shows that KNN Classifier gives the best accuracy with 78% in assessment with the alternative system gaining knowledge of algorithms used in this paper.

5. Conclusion and Future Work

This document involves predicting the diabetes disease dataset with proper computing and implementation of machine learning algorithms. In this article, we will use five machine learning algorithms to make predictions.

Calculating Accuracy
Among all the machine learning algorithms used in this paper, the highest accuracy is achieved by K Nearest Neighbors Classifier with 87%. This article shows that the machine learning algorithms is accustomed predict the center sickness simply with totally different parameters and models. Machine learning is very useful in prediction, solving problems and other areas. Machine learning is an effective way to solve the problems in different areas too.

6. Future Enhancements

This may permit new calculation improvement to be performed off-site utilizing distributed computing programming, and afterward got back to the clinical setting as applications program interfaces (APIs) for PCs, cell phones and tablets.

References

[1] Wei, S., Zhao, X., & Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), 291–295. https://doi.org/10.1109/WF-IoT.2018.8355130

[2] Rotheram-Borus, M. J., Tomlinson, M., Gwegwe, M., Comulada, W. S., Kaufman, N., & Keim, M. (n.d.). Diabetes buddies: peer support through a mobile phone buddy system. The Diabetes Educator, 38(3), 357–365. https://doi.org/10.1177/0145721712444617

[3] Alotaibi, M. M., Istepanian, R. S. H., Sungoor, A., & Philip, N. (2014). An intelligent mobile diabetes management and educational system for Saudi Arabia: System architecture. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 29–32. https://doi.org/10.1109/BHI.2014.6864296

[4] Alic, B., Gurbeta, L., & Badnjevic, A. (2017). Machine learning techniques for classification of diabetes and cardiovascular diseases. 2017 6th Mediterranean Conference on Embedded Computing (MECO), 1–4. https://doi.org/10.1109/MECO.2017.7977152

[5] Hsu, W. C., Lau, K. H. K., Huang, R., Ghiloni, S., Le, H., Gilroy, S., Abrahamson, M., & Moore, J. (2016). Utilization of a Cloud-Based Diabetes Management Program for Insulin Initiation and Titration Enables Collaborative Decision Making Between Healthcare Providers and Patients. Diabetes Technology & Therapeutics, 18(2), 59–67. https://doi.org/10.1089/dia.2015.0160