e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

DETECTION AND ANALYSIS OF DIABETES BY USING LOGISTIC REGRESSION (LR)

Mohammed Guhdar Mohammed[*1]

[*1]Department Of Computer Science, Faculty Of Science, University Of Zakho,
Zakho, Kurdistan Region.

DOI : https://www.doi.org/10.56726/IRJMETS32941

ABSTRACT

Logistic regression is a type of statistical model that can be used to predict the probability of an outcome occurring, given a set of input features. In the case of diabetes, logistic regression could be used to predict the probability that an individual has diabetes, based on a set of risk factors such as age, family history, and body mass index. Logistic regression is a popular method for predicting binary outcomes, like the presence or absence of a disease, and can be a useful tool for identifying individuals at high risk of developing diabetes. However, Logistic Regression is a simple method and its predictions are based on a linear combination of input features, it may not work very well when the relationship between the inputs and the outcome is non-linear or when there is a high degree of interaction between the input features, In this research, we are going to apply logistic regression on a Kaggle dataset to predict the probability of an individual having diabetes, based on a set of risk factors such as age, family history, and body mass index. The model will be trained on the data available in the Kaggle dataset and will be used to make predictions on new, unseen data.

I. INTRODUCTION

Diabetes is a chronic medical condition characterized by high levels of sugar (glucose) in the blood. The body's ability to produce or respond to insulin, a hormone that regulates blood sugar, is impaired in diabetes. This can result in a variety of symptoms and complications, including fatigue, increased thirst and urination, blurred vision, slow healing of wounds, and an increased risk of heart disease and stroke (American Diabetes Association, 2020). There are two main types of diabetes: Type 1 diabetes and Type 2 diabetes. Type 1 diabetes is an autoimmune disorder, where the body's immune system attacks and destroys the cells that produce insulin. This results in little or no insulin production, which leads to the need of exogenous insulin for survival. This type of diabetes is typically diagnosed in childhood or early adulthood, and accounts for about 5-10% of all diabetes cases (American Diabetes Association, 2020). Type 2 diabetes is a metabolic disorder, resulting from a combination of genetic and environmental factors. The body produces insulin, but the cells are resistant to its action or the pancreas is not able to produce enough insulin to meet the body's needs. This type of diabetes is often associated with obesity, sedentary lifestyle and aging, and accounts for about 90-95% of all diabetes cases (American Diabetes Association, 2020). If not managed properly, diabetes can lead to a variety of serious health problems, including cardiovascular disease, neuropathy, nephropathy, retinopathy and amputation (American Diabetes Association, 2020). Uncontrolled diabetes can cause damage to blood vessels and nerves, leading to heart disease, stroke, kidney failure, blindness, and amputations. It also increases the risk of developing serious infections (Moradian, A.D., 1988). To manage diabetes, individuals need to monitor their blood sugar levels and take steps to keep them in a healthy range. This can be done through a combination of lifestyle changes such as healthy diet, regular physical activity, maintaining healthy weight and not smoking, as well as medication such as insulin or oral diabetes drugs (American Diabetes Association, 2020). Furthermore, regular check-ups, monitoring and management of diabetes are important to prevent or delay the onset of serious complications (American Diabetes Association, 2020). Overall, diabetes is a serious and complex disease that requires ongoing management and monitoring to prevent serious health problems. Early diagnosis and proper management can help individuals with diabetes lead healthy and fulfilling lives (American Diabetes Association, 2020).

II. RELATED WORK

Diabetes is a chronic disease that can lead to severe health issues, as highlighted in a report from the World Health Organization (Sneha, N., & Gangil, T. 2019). This report states that diabetes is responsible for over 22 lakh deaths worldwide, emphasizing the need for effective methods of analyzing diabetes data. In previous

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

studies, various classification algorithms have been used to analyze diabetes datasets, such as in (Hina, S., Shaikh, A., & Sattar, S. A. (2017), where the Pima Indian Diabetes dataset was analyzed using techniques like Naive Bayes (NB), Zero R, J48, Random Forest (RF), and Logistic Regression (LR). In this study, a data mining tool called WEKA was used, and it was found that the Multilayer perceptron (MLP) technique had the highest accuracy and performance. Another study in (Nai-Arun, N., & Moungmai, R. (2015) applied machine learning classifiers like Decision Tree, Artificial Neural Networks (ANN), Logistic Regression, and Naive Bayes to characterize the risk of diabetes mellitus. Bagging and Boosting methods were utilized to improve the performance of the model, and it was found that the Random Forest classifier had the best results among all the classifiers, with high accuracy. Another study in (Alfian, Ganjar, et al, 2020) proposed a forecast model using Artificial Neural Network (ANN) and Fasting Blood Sugar (FBS) to predict chronic diabetes. Decision Tree (DT) has also been used to distinguish the symptoms of diabetes on patients' health, as shown in (Han, J., Rodriguez, J. C., & Beheshti, M. (2009).

III. METHODOLOGY

The proposed algorithms are carried kaggle diabetes dataset downloaded from www.Kaggle.com

The proposed model is evaluated based on accuracy measures obtained from Logistic regression (LR).

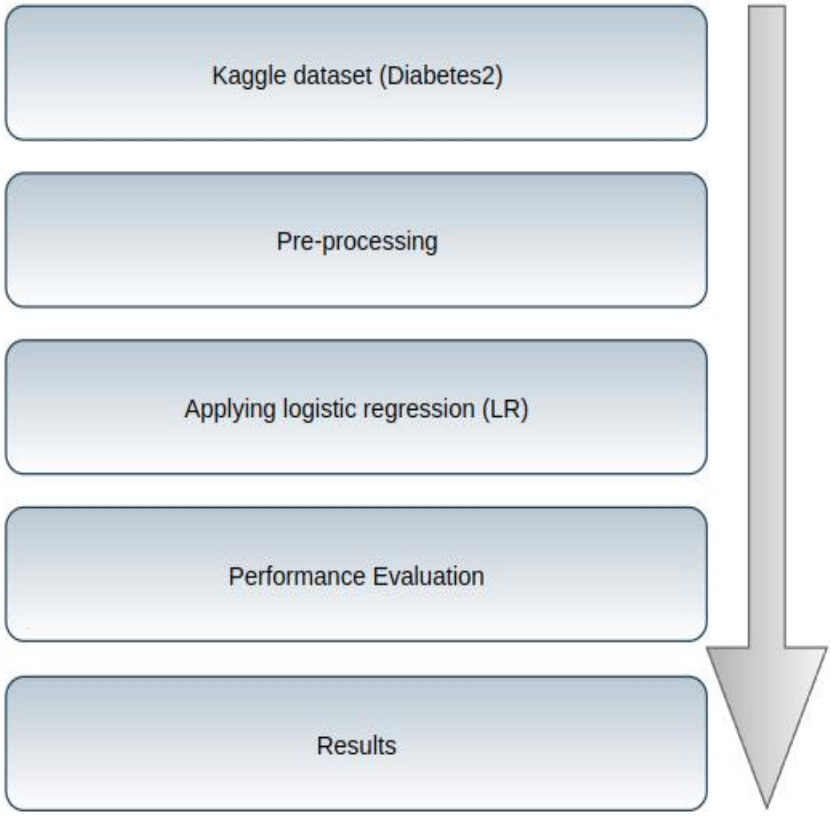

Figure 1: Proposed model

Regression analysis is a type of predictive modeling that is used to make predictions. There are three main types of regression: Linear Regression, Logistic Regression, and Polynomial Regression. Logistic Regression [23] is particularly useful in natural sciences, sociology, and has a connection to neural networks. It is used to solve classification problems and gives a binary output that predicts the outcome of a discrete dependent variable. In logistic regression, the algorithm uses a logistic function or sigmoid function, which maps real values to a value between 0 and 1. The word 'logistic' comes from the logit function, which is the main function of the algorithm. Logistic Regression is a supervised learning algorithm that deals with two possible outcomes, such as yes or no, true or false, 0 or 1, high or low. It is based on the likelihood of the outcome and if the probability is less than 0.5 it is rounded to 0, otherwise it is considered 1. The sigmoid function, which is in the shape of an "S", is used to take real values and plot them in the range of [0,1], as shown in Fig 2.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

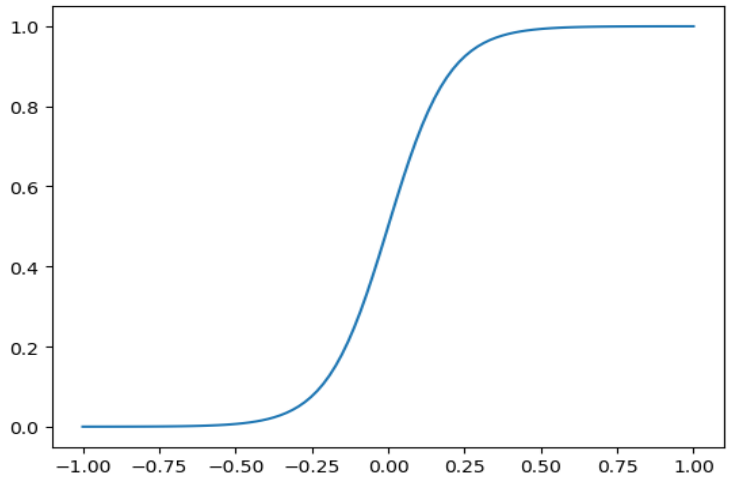

Figure 2: sigmoid function range 0/1

Fig. 3 shows that an analysis of the diabetes dataset has demonstrated a correlation among various attributes. By examining the data and evaluating the statistical relationship among the different attributes, it becomes clear that certain factors, such as age, glucose levels, insulin levels, and skin thickness, have a strong positive correlation with the outcome. This suggests that these factors play a significant role in the results and should be given special attention in further research. Additionally, this information can aid in the development of models and predictions, as well as in identifying potential risk factors for diabetes. It is important to note that correlation does not necessarily imply causation. Further research and analysis is required to establish causality and understand the underlying mechanisms of how these factors affect the outcome. However, the demonstration of correlation among these attributes can be used as a starting point for further studies and can aid in the development of hypotheses to be tested. Overall, the correlation analysis provides a deeper understanding of the complex relationships among the attributes and can guide future research and inform treatment decisions.

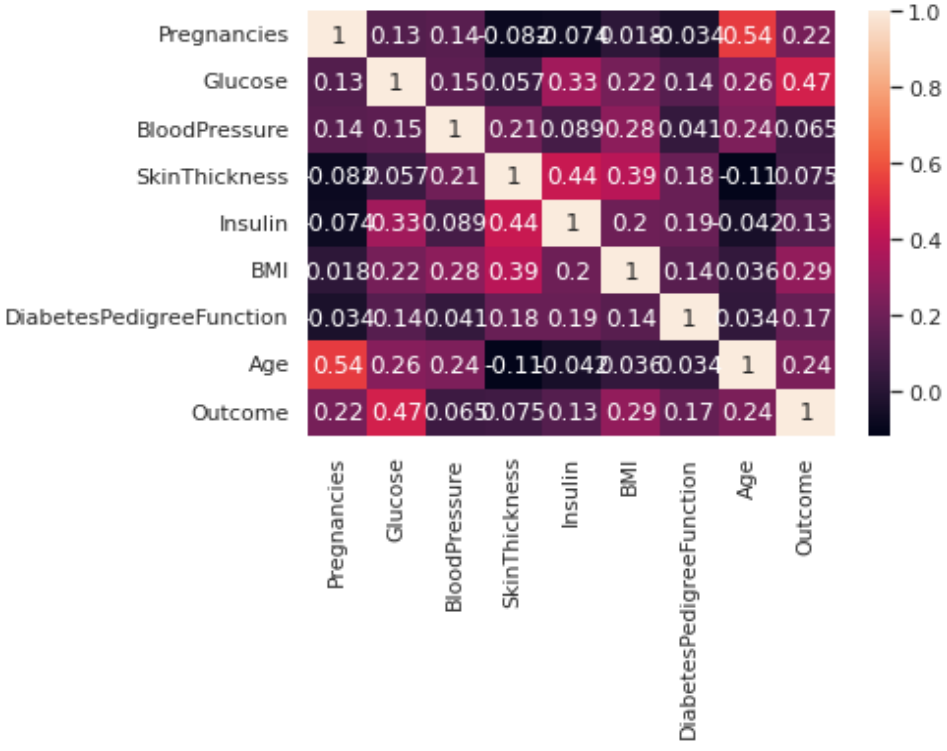

Figure 3: Correlation among different attributes

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:01/January-2023 Impact Factor- 6.752 www.irjmets.com

IV. RESULTS AND DISCUSSION

Evaluating the performance of a model is an essential step in the process of analyzing a dataset. One common method of evaluating the performance of a classification model is through the use of the F1 score. In this work, a logistic regression model was applied to a Kaggle dataset of diabetic patients and the resulting F1 score was calculated to be 0.607. It is important to note that an F1 score is a measure of a model's performance that takes both precision and recall into account. A high F1 score indicates that the model has both a high precision and a high recall, which are desirable characteristics for a model that aims to classify and predict accurately. An F1 score of 0.607 is considered to be an acceptable level of performance for logistic regression on this particular diabetic Kaggle dataset. Additionally, it's important to note that model's performance is dependent on several factors including dataset quality, model's complexity and number of samples, etc. For comparison to be possible, the scoring should be done on an independent validation dataset or by using cross-validation to measure the generalization performance. Furthermore, in medical domain, it is important to ensure that the model is not only accurate but also fair and unbiased. It's important to evaluate the model performance across different sub-populations in the dataset and make sure that there's no significant differences in performance between these groups. In conclusion, the logistic regression model applied to the diabetic Kaggle dataset has an acceptable F1 score of 0.607, demonstrating its potential to accurately classify and predict cases of diabetes. However, further examination and refinement of the model is necessary to ensure the robustness and generalizability of its performance.

```
Accuracy on Train set 0.7635009310986964
Accuracy on Test set 0.7705627705627706
Fl-score on Test set: 0.6074074074074073
              precision    recall  fl-score   support

           0       0.82      0.86      0.84       160
           1       0.64      0.58      0.61        71

    accuracy                           0.77       231
   macro avg       0.73      0.72      0.72       231
weighted avg       0.77      0.77      0.77       231
```

Note: test results directly taken from a python notebook.

V. CONCLUSION

Diabetes is a prevalent and serious real-world problem that requires early prediction and diagnosis. In order to address this issue, a system has been designed that utilizes the logistic regression algorithm to analyze the disease. By conducting simulations on a Kaggle dataset using various classifiers, it has been shown that logistic regression offers strong performance. This simulation model has the potential to be adapted and applied to other diseases, potentially leading to the automation of the diagnosis and management of chronic diseases in the future. Additionally, this work is further advanced with the integration of deep learning techniques, including the use of pre-trained networks such as Recurrent Neural Networks, Visual Geometry Group Nets and AlexNet. The application of logistic regression algorithm in the diagnosis of diabetes is a significant step towards developing a more efficient and accurate system for disease diagnosis and management. The use of deep learning techniques and pre-trained networks further enhances the performance of the model and open more door for the research. The goal of this work is to contribute to the advancement of medical technology and improve patient outcomes.

VI. REFERENCES

[1] Association, A. D. (2020). 1. Improving care and promoting health in populations: Standards of Medical Care in Diabetes—2020. Diabetes care, 43(Supplement_1), S7-S13.

[2] Mooradian, A. D. (1988). Diabetic complications of the central nervous system. Endocrine reviews, 9(3), 346-356.

[3] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 1-19.

[4] Hina, S., Shaikh, A., & Sattar, S. A. (2017). Analyzing diabetes datasets using data mining. Journal of Basic and Applied Sciences, 13, 466-471.

[5] Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142

[6] Alfian, G., Syafrudin, M., Fitriyani, N. L., Anshari, M., Stasa, P., Svub, J., & Rhee, J. (2020). Deep neural network for predicting diabetic retinopathy from risk factors. Mathematics, 8(9), 1620.

[7] Han, J., Rodriguez, J. C., & Beheshti, M. (2009). Discovering decision tree based diabetes prediction model. In International conference on advanced software engineering and Its applications (pp. 99-109). Springer, Berlin, Heidelberg.