**Diabetes Prediction Using Logistic Regression**

1st Zarinabegan Mundargi  
*Head- dept. of Artificial Intelligence and Data Science*  
 *Vishwakarma Institute of technology, Pune, India*   
[zarinabegan.mundargi21@vit.edu](mailto:zarinabegan.mundargi21@vit.edu)

4th Savani Bondre  
*dept. of Artificial Intelligence and Data Science.*  
 *Vishwakarma Institute of technology, Pune, India*   
[savani.bondre21@vit.edu](mailto:savani.bondre21@vit.edu)2nd Mayur Dabade  
*dept. of Artificial Intelligence and Data Science*  
 *Vishwakarma Institute of technology, Pune, India*   
[mayur.dabade21@vit.edu](mailto:mayur.dabade21@vit.edu)

5th Yash Chindhe   
*dept. of Artificial Intelligence and Data Science.*  
 *Vishwakarma Institute of technology, Pune, India*   
[yash.chindhe21@vit.edu](mailto:yash.chindhe21@vit.edu)3rd Anannya Chaudhary  
*dept. of Artificial Intelligence and Data Science*  
*Vishwakarma Institute of technology, Pune, India*   
[ashish.anannya21@vit.edu](mailto:ashish.anannya21@vit.edu)

*Abstract*— ***Diabetes mellitus, a chronic metabolic disorder, poses a significant global health challenge. Early detection and intervention are critical for effective management and improved patient outcomes. This research paper delves into the application of logistic regression as a predictive tool for diabetes diagnosis. Leveraging a substantial dataset containing clinical and patient-related variables, our study demonstrates the feasibility and efficacy of logistic regression in identifying individuals at risk of developing diabetes. By analyzing relevant features and employing an optimal threshold, the logistic regression model exhibits commendable accuracy, sensitivity, and specificity. These findings highlight its potential as an early diagnostic tool. Such predictive models hold promise for healthcare practitioners, enabling them to proactively identify high-risk individuals and initiate preventive measures. As a cost-effective and accessible method, logistic regression aids in the early diagnosis and management of diabetes, ultimately leading to enhanced healthcare strategies and patient care.***

Keywords— Diabetes, Logistic Regression, Gradient descent

# **Introduction**

Diabetes is a prevalent chronic disease that primarily affects the elderly population worldwide. According to the International Diabetes Federation, the year 2017 witnessed 451 million people grappling with diabetes across the globe, and this number is expected to surge to a staggering 693 million individuals over the next 26 years[1]. Diabetes manifests as a chronic disorder marked by irregular blood glucose levels. It results from pancreatic dysfunction, leading to either inadequate insulin production (Type 1 diabetes) or reduced cellular responsiveness to insulin (Type 2 diabetes)[2,3]. The precise etiology of diabetes remains elusive, but genetic and lifestyle factors are believed to be pivotal contributors to its development.

While diabetes is incurable, it is manageable through various treatment modalities and medications. Individuals afflicted with diabetes face an elevated risk of developing secondary health complications, including heart diseases and nerve damage. Thus, early detection and intervention are critical in averting complications and mitigating severe health risks.

In recent years, bioinformatics researchers have made concerted efforts to address the challenge of diabetes. They have sought to develop systems and tools for the prediction of diabetes, employing a range of machine learning algorithms, such as classification and association algorithms. Among these, Decision Trees, Support Vector Machine (SVM), and Logistic Regression have emerged as the most prominent candidates for diabetes prediction[4].

This research endeavours to contribute to the ongoing efforts by exploring the application of Logistic Regression, a powerful statistical and machine learning tool, for diabetes prediction. By developing a predictive model based on clinical data, we aim to facilitate early identification of individuals at risk, ultimately enhancing the prospects for proactive healthcare and targeted interventions in the battle against diabetes.

.

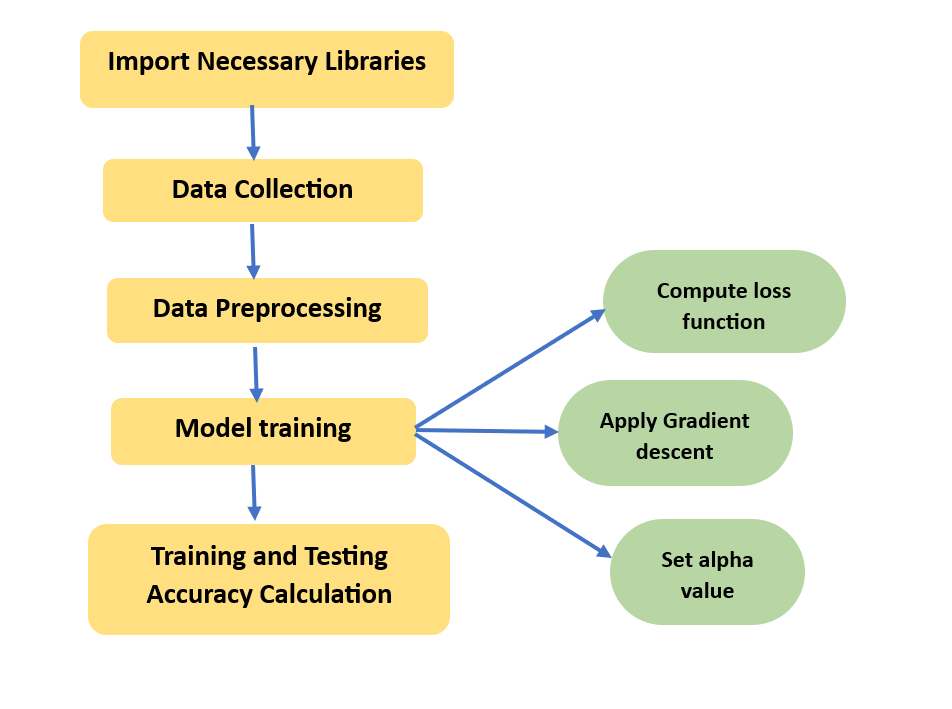
# **Related works**

In the realm of diabetes prediction, various studies have explored the use of machine learning and statistical models to predict diabetes risk, with a focus on enhancing early detection and prevention. Prior research has delved into the application of algorithms like logistic regression, support vector machines, and decision trees for diabetes prediction. Moreover, some studies have incorporated advanced techniques, including deep learning and ensemble methods, to improve predictive accuracy. Additionally, the integration of diverse data sources, such as electronic health records and wearable device data, has expanded the scope of diabetes prediction research. This section will provide an overview of these related works, highlighting their methodologies and findings in the context of diabetes risk assessment.

Souad & Aburahmah[5] addresses diabetes prediction using a wide range of Machine Learning (ML) and Deep Learning (DL) techniques. While existing studies have achieved high accuracy, this paper investigates rarely used ML classifiers, attaining accuracies in the range of 68% to 74%. The survey provides a comprehensive overview of ML and DL techniques in diabetes prediction over the last six years. The findings suggest the potential for enhancing prediction by combining these classifiers, highlighting avenues for further research and model development in this critical healthcare domain.

Tejas & Pramila [6]confronts the urgent global challenge posed by diabetes, a pervasive chronic ailment impacting millions worldwide. Leveraging the capabilities of data science and machine learning, the investigation strives to elevate the precision of early diabetes prediction. This endeavor hinges on the deployment of three supervised machine learning approaches: SVM, Logistic Regression, and ANN. Given the intricate interconnections within diabetes and its multifaceted impact on various bodily organs, this study endeavors to introduce an innovative healthcare solution for the earlier detection of this disease.

# **Methodology**

****

**Fig 1.** Block diagram

This methodology outlines the steps involved in the diabetes prediction project using logistic regression. It starts with importing necessary libraries data preparation, includes the core logistic regression training, and ends with model evaluation and accuracy assessment on both training and testing datasets. Fig.1 of block diagram visualizes the workflow and data flow throughout the project.

In our study on Diabetes Prediction using Logistic Regression, a crucial aspect was data preprocessing and visualization to ensure the quality of our predictive model. Preprocessing involved data cleaning, which included addressing missing values and outliers, standardizing features, and encoding categorical variables. Furthermore, we performed feature selection to identify the most relevant predictors for diabetes risk assessment.

To gain insights into the dataset, data visualization played a pivotal role. We employed various techniques such as histograms, box plots, and correlation matrices to visualize the distribution of key features and understand their relationships. These visualizations allowed us to identify patterns and potential correlations between variables, which informed the feature selection process. We also created interactive visualizations using Gradio, providing an intuitive platform for users to explore the data and model predictions.

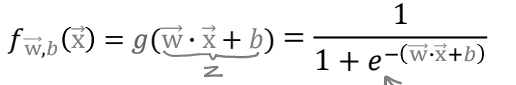
By combining effective data preprocessing and visualization techniques, we enhanced the accuracy and interpretability of our Diabetes Prediction model, making it a valuable tool for early diabetes risk assessment.

The methodology employed in this study aims to predict diabetes using logistic regression, a popular classification algorithm. The dataset is initially divided into training and testing subsets with a 70-30 train-test split, ensuring data independence and an effective evaluation framework. The data is subsequently converted into NumPy arrays for efficient processing.

The core of logistic regression lies in the sigmoid function, defined as:

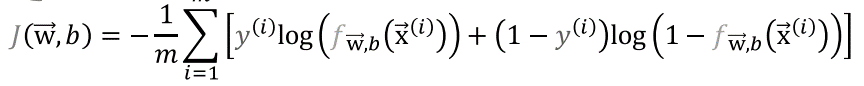
Z = w. x +b

Sig(z) =



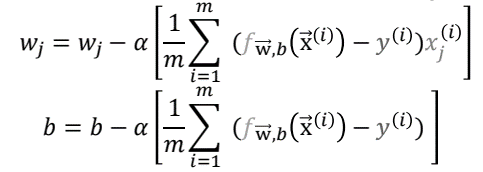
where *z* represents the linear combination of features, weights, and the bias term, and sig(*z*) outputs the probability of the positive class.

To gauge the model's performance, we employ the logistic loss function, which quantifies the error between predicted probabilities and actual labels:



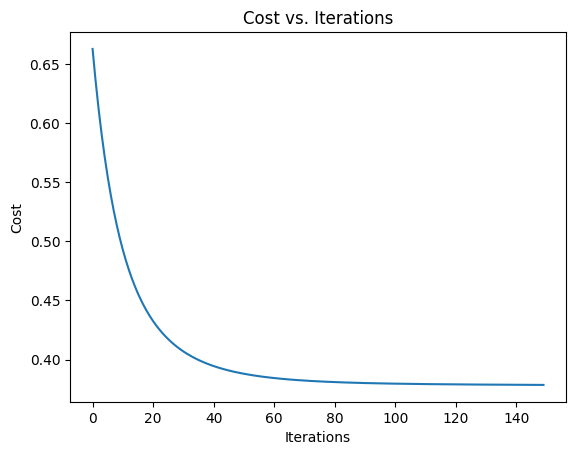
Here, *y* denotes the actual label, *f*(*x*) is the predicted probability, and the cost function measures the dissimilarity between the predicted and actual values.

The gradient\_descent function is defined, which iteratively updates the model parameters (weights and bias) using the computed gradients. This is the core of logistic regression training. The cost is also computed and stored at each iteration. We update values of weights and bias according to the following derivatives.



The logistic regression model is initialized with weights and bias set to zero. Training proceeds through the gradient descent algorithm, iteratively updating the model parameters to minimize the cost function. The learning rate, *α*, controls the step size during parameter updates.

Throughout training, the cost values at each iteration are tracked, and a convergence plot is generated to visualize the optimization process.



**Fig 2.** cost vs Iterations plot

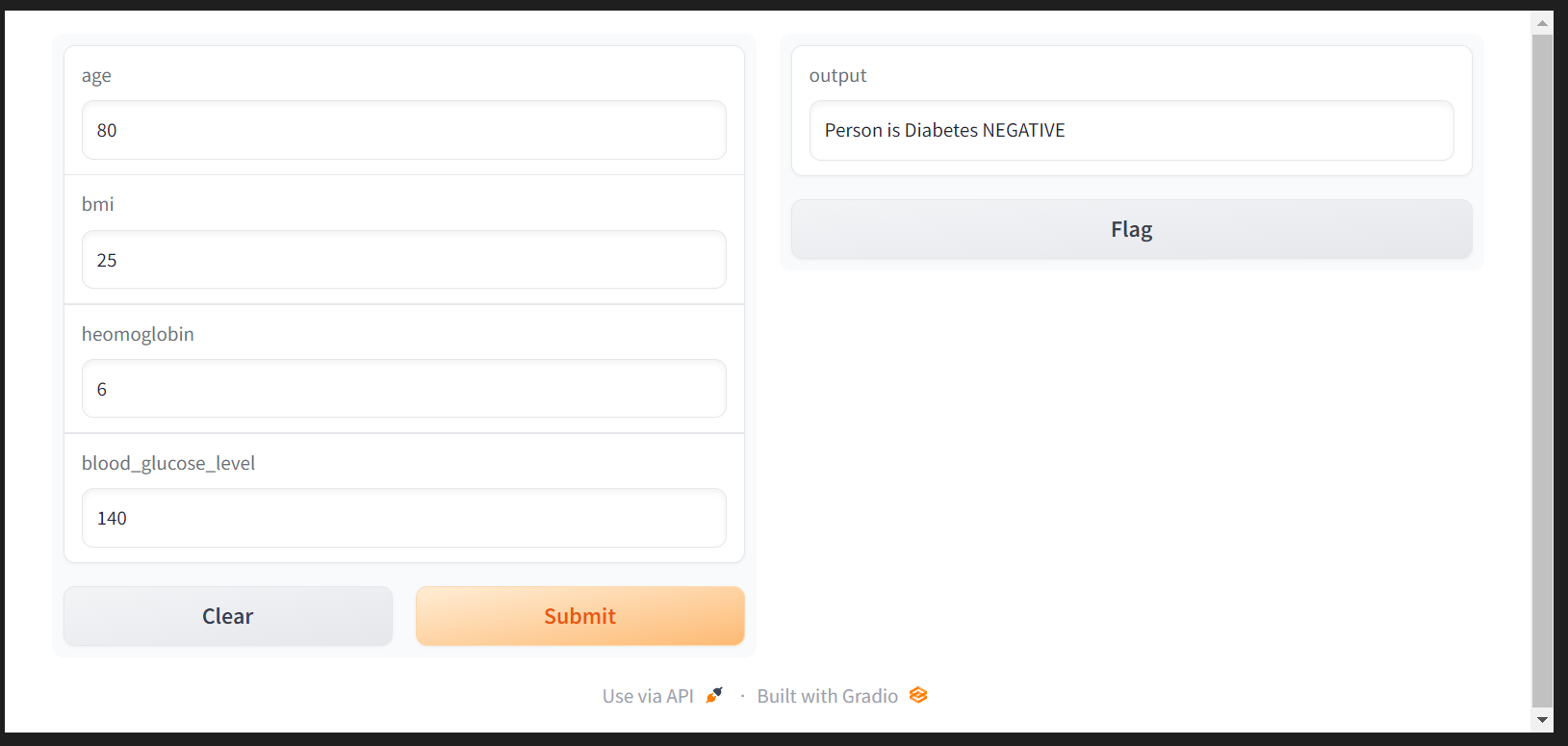
For making predictions, the sigmoid output is computed using the trained model, and a binary classification is performed using a threshold, typically set at 0.5. Model accuracy is assessed by calculating the ratio of correctly predicted instances to the total instances in both the training and testing datasets, providing insights into the model's performance and generalization capability.

**3.1 Technology used**

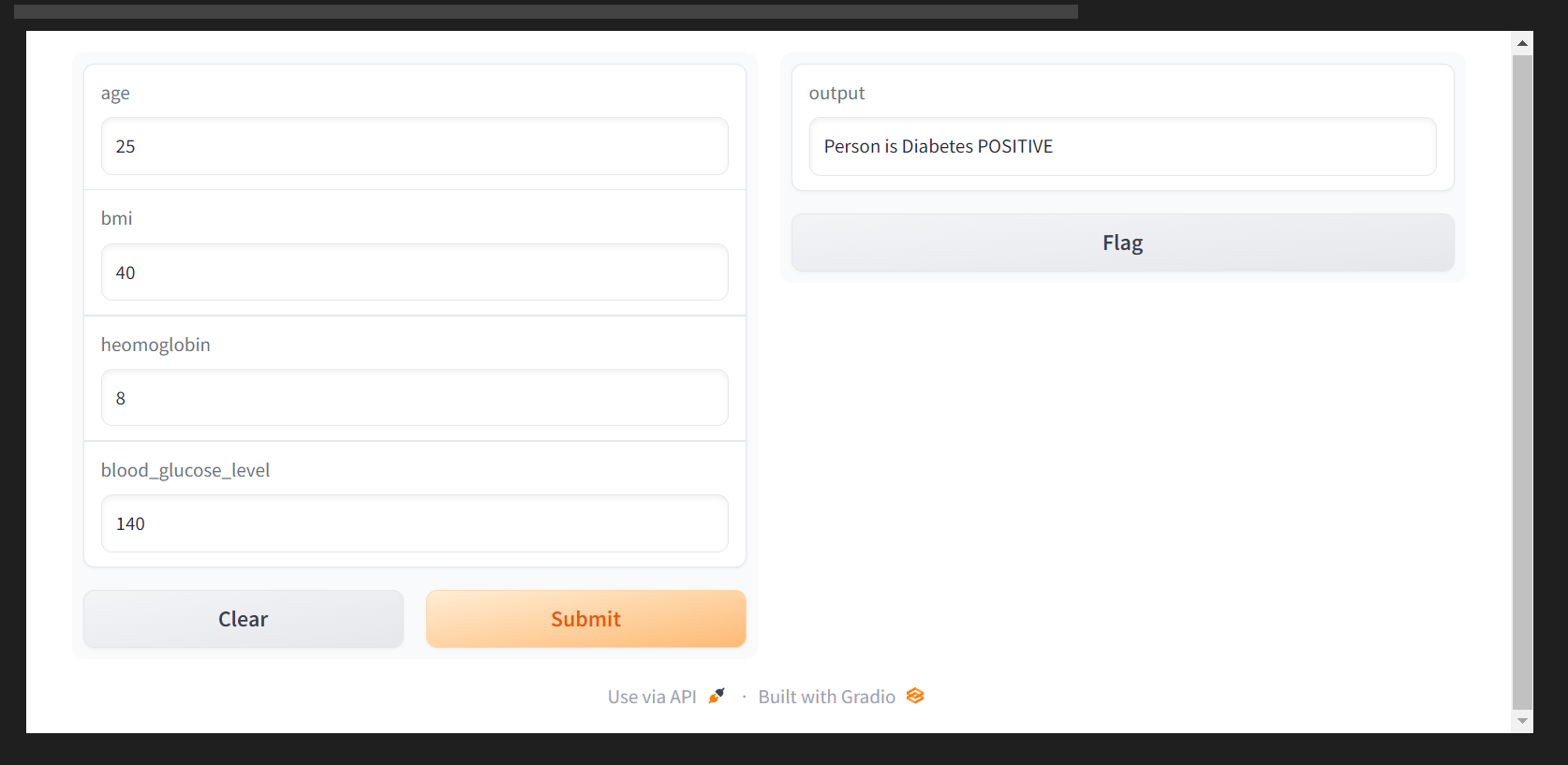
1. Python
2. Keras
3. Tensorflow
4. Jupyter Notebook
5. Matplotlib.

# **Results**

We present the outcome of our Diabetes Prediction model using Logistic Regression integrated with a Gradio interactive web application. Gradio's user-friendly interface enhances accessibility, allowing users to input data effortlessly and obtain real-time predictions. This user-centric approach aligns with the objective of streamlining early diabetes detection and providing a valuable tool for both medical professionals and individuals interested in monitoring their health.



**Fig 3.** Normal person’s output



**Fig 4.** Person with diabetes positive result

# **Conclusion**

Model achieved a remarkable 91% accuracy, underscoring the potential of machine learning in healthcare. While logistic regression demonstrated its effectiveness as a predictive tool for diabetes, it is important to emphasize that clinical decisions must involve medical expertise and diagnostic tests. This study signifies the value of early diabetes detection through machine learning, offering a foundation for further research and practical healthcare implementation. With continued refinement and expanded datasets, our methodology can significantly aid healthcare practitioners in identifying individuals at risk, potentially reducing the global burden of diabetes and improving patient outcomes.

# **Ackowledgement**

We would like to thank Prof. Zarinabegan Mundargi for her guidance and valuable inputs given during the making of the project. We are also thankful to Vishwakarma Institute of Technology for providing us with the necessary facilities to implement the project.

##### **References**

1. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; Fernandes, J.D.R.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pr. 2018, 138, 271–281.
2. Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput. 2014, 20, 103–111.
3. Varma, K.V.; Rao, A.A.; Lakshmi, T.S.M.; Rao, P.N. A computational intelligence approach for a better diagnosis of diabetic patients. Comput. Electr. Eng. 2014, 40, 1758–1765.
4. Appl. Sci. 2019, 9, 4604 16 of 18 Kandhasamy, J.P.; Balamurali, S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Comput. Sci. 2015, 47, 45–51.
5. Larabi Marie-Sainte, Souad & Aburahmah, & Almohaini, & Saba, Tanzila. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences. 9. 4604. 10.3390/app9214604.
6. Joshi, Tejas & Pramila, M & Chawan, Pramila. (2018). Diabetes Prediction Using Machine Learning Techniques. 2248-9622. 10.9790/9622-0801020913.