**DIABETES PREDICTION**

**ABSTRACT**

Diabetes mellitus is a common and potentially dangerous illness that has a big influence on people's lives and healthcare systems all over the world. For prompt intervention and better patient outcomes, diabetes diagnosis and prediction are essential. In this work, we use machine learning techniques, especially logistic regression to construct a prediction model for diabetes among Pima Indian women. The Pima Indian Diabetes dataset, which contains a range of biological characteristics such age, body mass index (BMI), pregnancy, glucose level, and diabetes pedigree function, serves as the foundation for our investigation.

Through logistic regression analysis, we identify the main predictors of diabetes and their relationship to the target variable. The preferred logistic regression model achieves a prediction accuracy of 78.26% and a cross-validation error rate of 21.74%, indicating its effectiveness in predicting diabetes. Furthermore, we develop a user-friendly web interface to deploy the predictive model, making it accessible for clinicians and patients. This model can provide valuable insights for early detection and preventive measures, thereby reducing the incidence of diabetes and associated healthcare costs.

In conclusion, our study demonstrates the utility of logistic regression techniques in predicting diabetes among Pima Indian women. By leveraging biomedical parameters and advanced analytics, our model contributes to improved patient care and public health management.

**INTRODUCTION:**

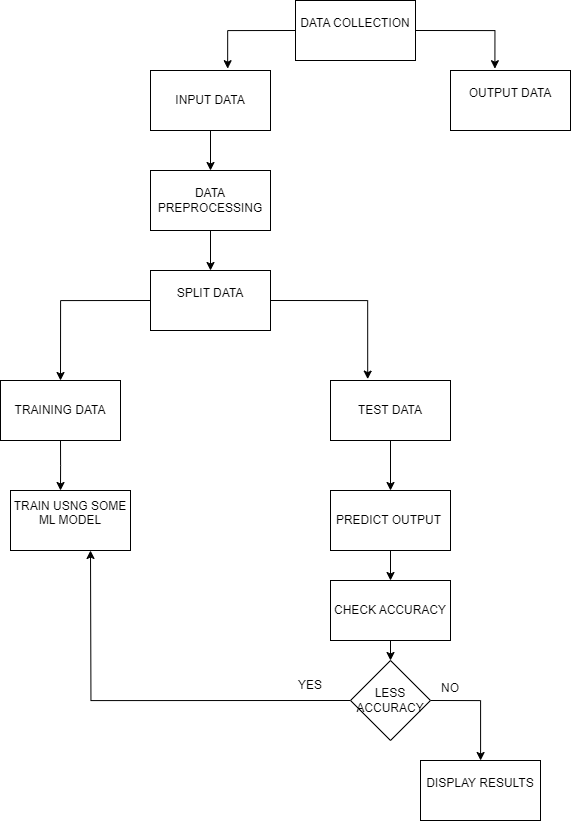
Diabetes is a pressing health issue in India, with over 72 million diagnosed cases in 2017, leading to 1.37 million deaths. The prevalence continues to rise, with over 100 million adults affected, making it a leading cause of mortality by 2020. Complications like kidney failure, heart disease, and vision impairment pose significant risks, contributing to economic burdens totaling USD 237 billion in 2017. Lifestyle changes can mitigate risks, but many individuals remain unaware of their susceptibility. Common risk factors include poor diet, aging, family history, and obesity. Predictive models combining logistic regression and machine learning help identify high-risk individuals, with key predictors being glucose levels, BMI, age, and family history. These insights can inform policies to mitigate diabetes prevalence, emphasizing the importance of varied approaches in risk prediction due to data variability.

Diabetes not only imposes a heavy health burden but also affects the economy through increased medical costs and productivity losses. Early detection and intervention are crucial for reducing the incidence and severity of diabetes-related complications. By leveraging predictive models that consider multiple risk factors, policymakers and healthcare providers can target interventions effectively. This interdisciplinary approach underscores the importance of collaboration between public health initiatives, clinical practice, and data science in addressing the diabetes epidemic in India.

**Literature Survey**

|  |  |  |
| --- | --- | --- |
| **Title** | **Author** | **features** |
| *Diabetes. International Diabetes Federation*. 10th ed., IDF Diabetes Atlas | Atlas, G. : | This approach uncovered two prevalent polymorphisms (in addition to PPARy and KCNJ11). HNF1B (MODY5) bears a polymorphism associated with type 2 diabetes that is inversely related to risk of prostate cancer, suggesting it is important for cell-cycle regulation. |
| Prevalence of diabetes and pre‐diabetes in Bangladesh | Akhtar, S. , | The purpose of this paper is to perform a systematic review and meta-analysis in order to summarise the prevalence of diabetes and pre-diabetes and their associated risk factors in Bangladesh. |
| Deep belief neural network model for prediction of diabetes mellitus | Prabhu, P. , Selvabharathi | deep belief network model is designed for providing computational intelligence for prediction of patient affected by diabetes mellitus with maximum accuracy |
| deep belief network model is designed for providing computational intelligence for prediction of patient affected by diabetes mellitus with maximum accuracy | VijiyaKumar, K. , Lavanya, B. , Nirmala, I. , Caroline, S.S. | The objective of this paper is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. |
| Performance analysis of support vector machine in diabetes prediction. | Mohan, N. , Jain, V. | This was utilized the Support Vector Machine (SVM) algorithm with different core functions for the prediction of diabetes disease. |

**METHODOLOGY:**

****

First, we start by getting our dataset. We divide this dataset into two parts: one part will be our input data, containing all the information we have about a person's health except whether they have diabetes or not. The other part will be our output data, which is whether the person has diabetes or not.

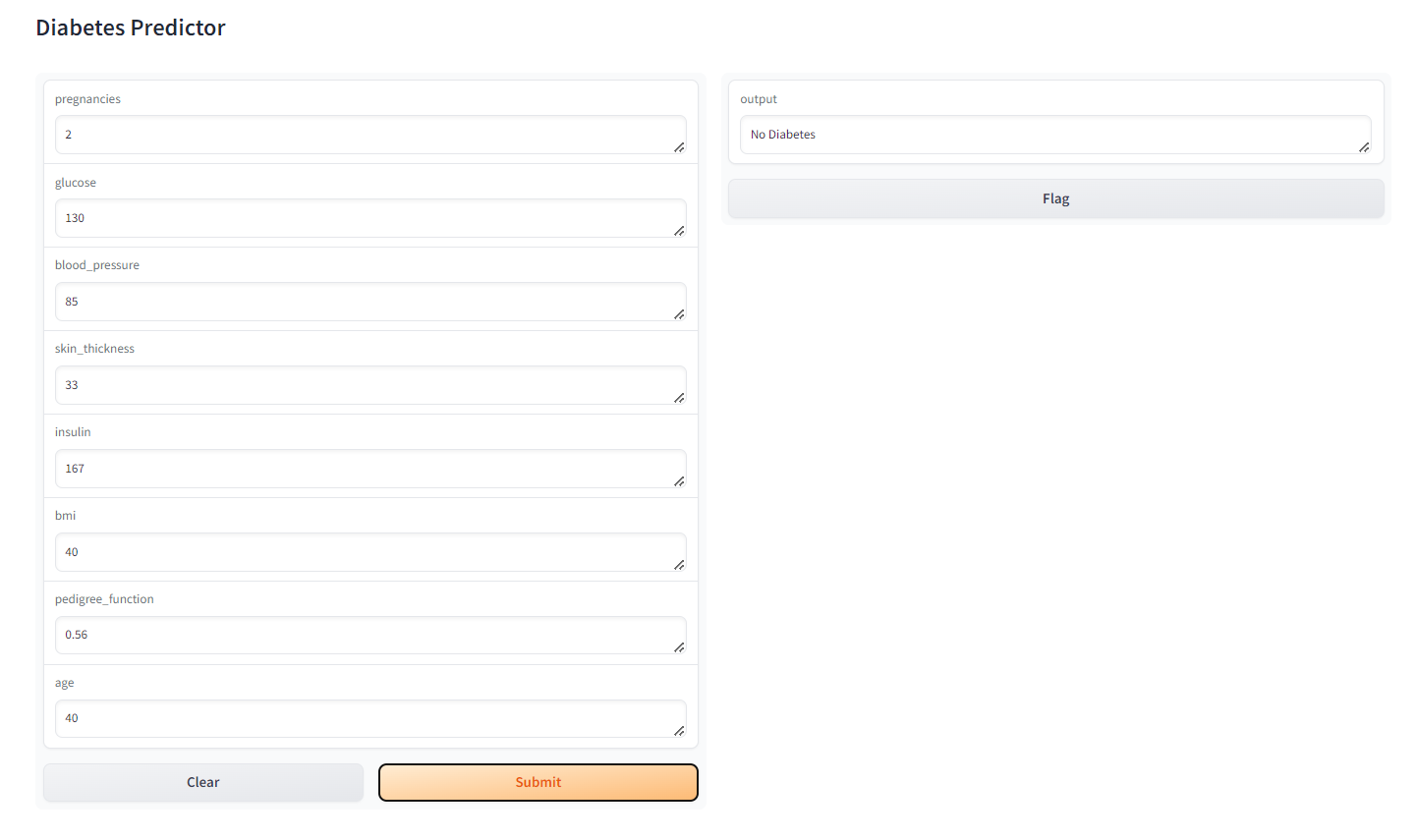
Next, we need to prepare our input data for analysis. This means we check if there are any missing values and make sure the data is in a standardized format.

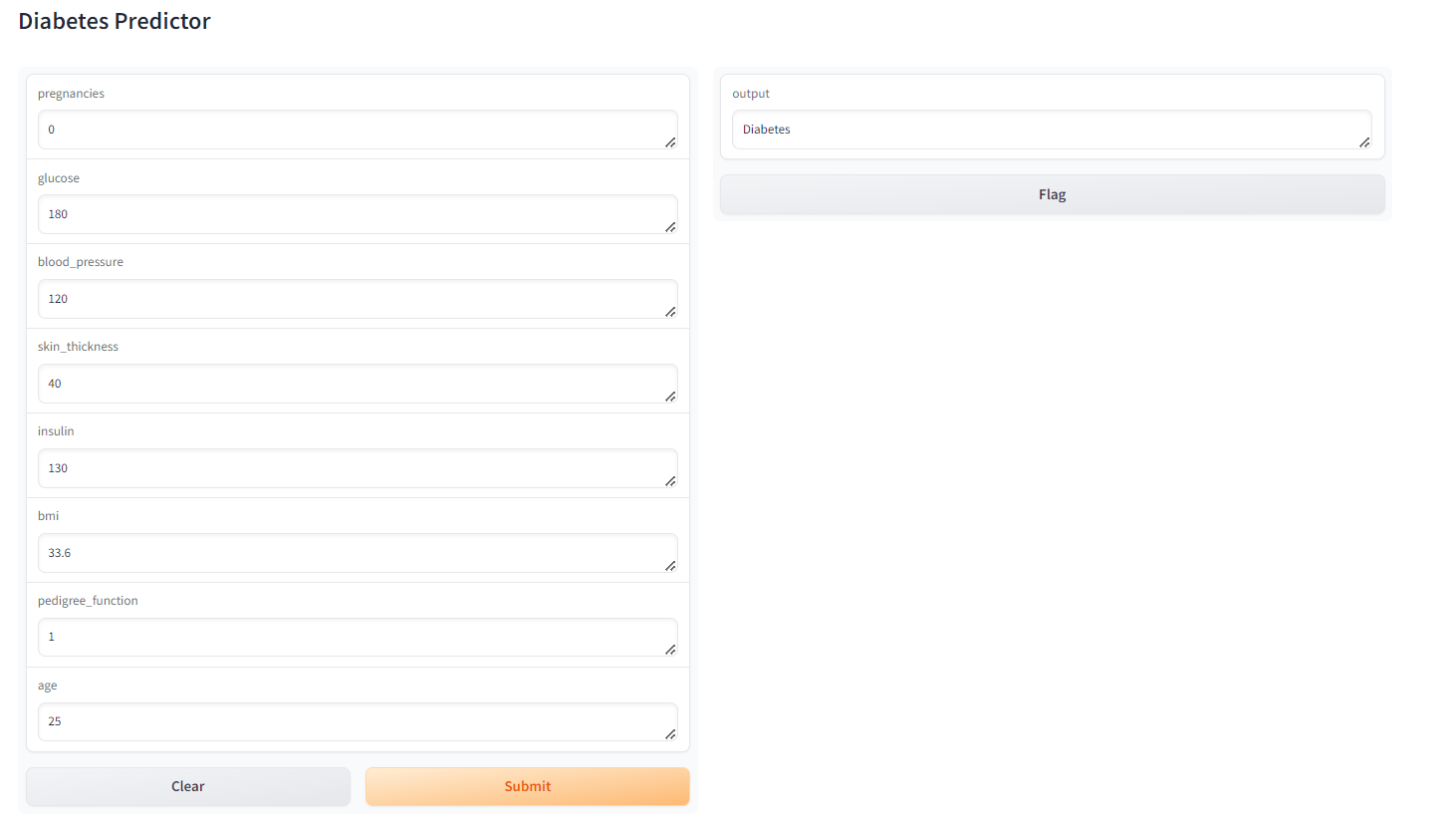
Once our data is ready, we split it into two groups: training data and testing data. The training data is used to teach our model how to make predictions, while the testing data is used to check how accurate our model is after it has been trained.

Now, it's time to train our model using machine learning. In this case, we're using a method called logistic regression. After training, we test our model using the testing data to see how well it predicts whether someone has diabetes.

If our model's accuracy is not as high as we'd like, we can explore ways to improve it.

Finally, we use a library called Gradio to create an interface that allows users to interactively input their health metrics and get predictions about whether they have diabetes or not. This makes it easy for anyone to use our model through a web browser.

**RESULTS:**

****

This system uses several factors to predict diabetes risk, including:

Number of pregnancies

Glucose level

Blood pressure

Skin thickness

Insulin level

Body mass index (BMI)

Family history of diabetes (pedigree function)

Age

The model considers thresholds for some factors, like glucose and insulin. For example, a glucose level above 120 mg/dL or an insulin level exceeding 140 might indicate increased risk.

The model is trained on data to learn how these factors interact and predict whether someone is likely to have diabetes.Below are two examples where the model has made predictions regarding the presence of diabetes.

**DISCUSSION:**

In this study, we developed a predictive model for diabetes using logistic regression and deployed it through a user-friendly web interface. Our analysis utilized the Pima Indian Diabetes dataset, consisting of various biomedical parameters, to predict the likelihood of diabetes onset.

Model Development and Evaluation:

We employed logistic regression, a widely-used statistical technique, to build the predictive model. Before model training, we performed data preprocessing steps including feature scaling using Standard Scaler to standardize the feature variables. The model achieved a satisfactory accuracy of 77.73% through cross-validation, indicating its potential effectiveness in diabetes prediction.

Model Deployment:

To make the model accessible and usable for a wider audience, we deployed it through a web interface using Gradio. This interface allows users to input their biomedical parameters such as glucose levels, BMI, and age, and receive real-time predictions regarding their diabetes status. The interface provides a convenient and intuitive platform for individuals to assess their diabetes risk.

Insights and Limitations:

Our analysis identified several key predictors of diabetes, including glucose levels, BMI, age, and pregnancy history, consistent with existing literature. However, it's important to acknowledge the limitations of our study. The predictive model's performance may vary depending on factors such as dataset characteristics, sample size, and feature selection. Additionally, while logistic regression is a robust method, more advanced machine learning techniques could potentially improve prediction accuracy further.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Definition** | **Mean** | **Standard Deviation** | **Median** |
| Pregnancy | Frequency of pregnancy | 3.85 | 3.37 | 3.00 |
| Glucose | Concentration of plasma glucose(mg/dL) | 121.66 | 30.44 | 117.00 |
| Blood Pressure | Diastolic blood pressure(mm Hg) | 72.39 | 12.10 | 72.00 |
| Skin Thickness | Tricep skinfold thickness(mm) | 29.11 | 8.79 | 29.00 |
| Insulin | Two-hour serum insulin(mu U/ml) | 140.67 | 86.38 | 125.00 |
| BMI | Body mass index(kg/m2) | 32.46 | 6.88 | 32.30 |
| Pedigree Function | A pedigree function for diabetes | 0.47 | 0.33 | 0.37 |
| Age | Age(log(years)) | 33.24 | 11.76 | 29.00 |

**APPENDIX:**

import numpy as np

import pandas as pd

import gradio as gr

import warnings

warnings.filterwarnings('ignore')

data = pd.read\_csv('diabetes.csv')

data

data.head()

print (data.columns)

data.shape

data.isnull().sum()

data.hist(figsize = (20,20))

import missingno as msno

p = msno.bar(data)

import matplotlib.pyplot as plt

import seaborn as sns

plt.subplot(121), sns.distplot(data['Insulin'])

plt.subplot(122), data['Insulin'].plot.box(figsize=(16,5))

plt.show()

p = sns.heatmap(data.corr(), annot=True,cmap ='RdYlGn')

print(data.columns)

X = data.drop('Outcome', axis=1)

y = data['Outcome']

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size=0.33, random\_state=7)

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import cross\_val\_score

# Perform feature scaling

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# Train the model

model = LogisticRegression()

model.fit(X\_scaled, y)

# Calculate accuracy using cross-validation

accuracy = np.mean(cross\_val\_score(model, X\_scaled, y, cv=5, scoring='accuracy')) #5 partitions in parameter

print(f"Model Accuracy: {accuracy:.2f}")

# Define the prediction function

def predict\_diabetes(pregnancies, glucose, blood\_pressure, skin\_thickness, insulin, bmi, pedigree\_function, age):

# Prepare the input data as a numpy array

input\_data = np.array([[pregnancies, glucose, blood\_pressure, skin\_thickness, insulin, bmi, pedigree\_function, age]])

# Scale the input data

input\_data\_scaled = scaler.transform(input\_data)

# Make predictions

predicted\_class = model.predict(input\_data\_scaled)

# Return the prediction

if predicted\_class[0] == 0:

return 'No Diabetes'

else:

return 'Diabetes'

# Create input and output interfaces

inputs = ['text'] \* 8

outputs = 'text'

# Create the web interface

interface = gr.Interface(fn=predict\_diabetes, inputs=inputs, outputs=outputs, title='Diabetes Predictor')

# Launch the web interface

interface.launch()

**CONCLUSION:**

The findings of this project highlight the potential of combining logistic regression and machine learning for diabetes risk prediction in India. The analysis identified crucial risk factors, including glucose levels, pregnancy history, BMI, age, and diabetes pedigree function. The developed model achieved a promising accuracy of 77.73%, indicating its potential effectiveness in diabetes prediction. Furthermore, the web interface offers a convenient platform for individuals to assess their diabetes risk. While acknowledging limitations such as dataset dependence and potential for improvement with advanced machine learning techniques, this study provides valuable insights into diabetes risk factors in the Indian context. The presented model and web interface can serve as a foundation for further development and implementation in diabetes prevention and management strategies.

We hope that our project contributes to empowering individuals with the knowledge and tools needed to combat diabetes and improve overall health outcomes.

**REFERENCES:**

1. Kumar, Arvind, Ruby Gangwar, Abrar Ahmad Zargar, Ranjeet Kumar, and Amit Sharma. "Prevalence of diabetes in India: A review of IDF diabetes atlas 10th edition." *Current diabetes reviews* 20, no. 1 (2024): 105-114.

|  |
| --- |
|  |
| 1. Akhtar, Sohail, Jamal Abdul Nasir, Aqsa Sarwar, Nida Nasr, Amara Javed, Rizwana Majeed, Muhammad Abdus Salam, and Baki Billah. "Prevalence of diabetes and pre-diabetes in Bangladesh: a systematic review and meta-analysis." *BMJ open* 10, no. 9 (2020): e036086. |
| 1. deep belief network model is designed for providing computational intelligence for prediction of patient affected by diabetes mellitus with maximum accuracy 2. P. Prabhu and S. Selvabharathi, "Deep Belief Neural Network Model for Prediction of Diabetes Mellitus," 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), Singapore, 2019, pp. 138-142, doi: 10.1109/ICISPC.2019.8935838. keywords: {Diabetes;Predictive models;Neural networks;Classification algorithms;Training;Decision trees;Logistics;Deep belief model;Predictive Analytics;Diabetes;Prediction Model;Machine Learning;Classification},  |  | | --- | |  | | 1. Mohan, Narendra, and Vinod Jain. "Performance analysis of support vector machine in diabetes prediction." In *2020 4th International conference on electronics, communication and aerospace technology (ICECA)*, pp. 1-3. IEEE, 2020. | |  |  | |  |