Predicting Diabetes using Logistic Regression

Contents

[1. Introduction 2](#_Toc176716405)

[2. Problem Statement 2](#_Toc176716406)

[3. Importance of the Study 2](#_Toc176716407)

[4. Methodology 3](#_Toc176716408)

[4.1 Data Overview 3](#_Toc176716409)

[4.2 Data Preprocessing 3](#_Toc176716410)

[4.3 Model Selection: Logistic Regression 3](#_Toc176716411)

[4.4 Hyperparameter Tuning 4](#_Toc176716412)

[4.5 Model Evaluation 4](#_Toc176716413)

[5. Results 5](#_Toc176716414)

[5.1 Correlation Analysis 5](#_Toc176716415)

[5.2 Model Performance Metrics 5](#_Toc176716416)

[6. Findings and Discussion 5](#_Toc176716417)

[7. Conclusion 5](#_Toc176716418)

[8. References 6](#_Toc176716419)

## 1. Introduction

Diabetes mellitus, commonly referred to as diabetes, is a chronic disease that occurs when the body cannot effectively regulate blood sugar (glucose) levels. High glucose levels over time lead to serious health complications such as heart disease, kidney damage, and nerve damage. Early diagnosis and intervention are critical to managing diabetes effectively, reducing healthcare costs, and improving patient outcomes.

**Machine learning** models, particularly **Logistic Regression**, offer an efficient method to predict the onset of diabetes based on key medical and demographic data. This research focuses on developing and fine-tuning a Logistic Regression model to predict the likelihood of diabetes in individuals using the Pima Indians Diabetes Dataset.

## 2. Problem Statement

Despite numerous advancements in medical diagnostics, early detection of diabetes continues to be a challenge, especially in resource-limited settings. Traditional diagnostic methods, such as the oral glucose tolerance test and fasting blood sugar tests, are time-consuming and require laboratory facilities, which might not be available in all regions. This research aims to:

1. Develop a machine learning model to predict the onset of diabetes using readily available medical data.
2. Identify the key factors that influence diabetes prediction.
3. Achieve an accurate and interpretable model that can aid healthcare professionals in making informed decisions regarding early intervention and treatment.

The goal is to determine whether Logistic Regression, a widely used classification algorithm, can provide reliable predictions based on key health indicators like BMI, insulin levels, and glucose levels.

## 3. Importance of the Study

Diabetes, if left untreated, leads to severe health complications and increased mortality rates. The ability to predict diabetes early can have numerous benefits:

* **Reduced healthcare burden**: Early diagnosis helps reduce the complications of diabetes, decreasing the demand for more advanced, expensive treatments.
* **Targeted interventions**: Identifying high-risk individuals allows for targeted interventions, such as lifestyle changes and medications, that could prevent or delay the onset of diabetes.
* **Resource efficiency**: In low-resource settings, predicting diabetes using machine learning could allow healthcare providers to make better use of available resources without requiring costly diagnostic tests for every individual.

By building a predictive model using Logistic Regression, we aim to demonstrate how data-driven approaches can augment existing healthcare diagnostics and improve patient outcomes.

## 4. Methodology

### 4.1 Data Overview

We used the **Pima Indians Diabetes Dataset** for this study, which is a well-known dataset from the UCI Machine Learning Repository. The dataset contains several medical features, including:

* **Pregnancies**: Number of times the individual has been pregnant.
* **Glucose**: Plasma glucose concentration (measured after a 2-hour oral glucose tolerance test).
* **Blood Pressure**: Diastolic blood pressure (mm Hg).
* **Skin Thickness**: Triceps skinfold thickness (mm).
* **Insulin**: (mu U/ml).
* **BMI**: Body mass index (weight in kg/(height in m²)).
* **Diabetes Pedigree Function**: A function representing a genetic predisposition to diabetes.
* **Age**: Age of the individual in years.
* **Outcome**: Binary class label (1 for diabetic, 0 for non-diabetic).

### 4.2 Data Preprocessing

Data preprocessing is essential to ensure the model performs optimally. In this step, we addressed the following:

1. **Feature Scaling**: Since the dataset contains features with varying scales (e.g., Age, Glucose, Insulin), standardization was performed using StandardScaler. This step is crucial for Logistic Regression, as it assumes the input features are on the same scale.

**Code example:**

from sklearn.preprocessing import StandardScaler

scalar = StandardScaler()

X\_scalled= scalar.fit\_transform(X)

print(X\_scalled)

### 4.3 Model Selection: Logistic Regression

Logistic Regression is a linear classification model commonly used for binary classification problems like this one. The logistic function maps predicted values to probabilities, providing a way to interpret the output as the probability of the outcome (diabetes in this case).

The model was trained using the following pipeline:

1. **Train-Test Split**: The dataset was split into 70% training and 30% testing sets to evaluate the model’s performance.

**Code example:**

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

X\_scalled\_train,X\_scalled\_test,y\_train,y\_test = train\_test\_split(X\_scalled, y, test\_size=0.3, random\_state=42)

1. **Logistic Regression Training**: The model was trained using the training set.

**Code example:**

from sklearn.linear\_model import LogisticRegression

lr = LogisticRegression()

from sklearn.model\_selection import GridSearchCV

parameter = {'penalty':['l1','l2','elasticnet'],'C' : [1,2,3,4,5,6,7,8,9,10,15,19,20,30], 'max\_iter' :[100,150,200,250,300]}

classifier=GridSearchCV(lr, param\_grid=parameter, scoring='accuracy', cv=10))

classifier.fit(X\_scalled\_train, y\_train)

### 4.4 Hyperparameter Tuning

To further optimize the model, **GridSearchCV** was used to search for the best combination of regularization methods (L1, L2) and the regularization strength (C).

The best parameters were identified as:

* **C**: 1
* **max\_iter**: 100
* **penalty**: 'L2' (Ridge regularization)

These parameters provided the best balance between the model's complexity and its ability to generalize to unseen data.

### 4.5 Model Evaluation

To evaluate the model's performance, the following metrics were considered:

* **Accuracy**: The overall correctness of the model.
* **Precision**: The proportion of true positive predictions for diabetes.
* **Recall**: The model’s ability to identify individuals with diabetes.
* **F1-Score**: The harmonic mean of precision and recall, providing a single score to evaluate the model's performance.

The results were obtained as follows:

**Code example:**

Diabetes\_predict = classifier.predict(X\_scalled\_test)

Diabetes\_predict

from sklearn.metrics import accuracy\_score, classification\_report

print('Accuracy Score', accuracy\_score(Diabetes\_predict, y\_test))

print(classification\_report(Diabetes\_predict, y\_test))

## 5. Results

### 5.1 Correlation Analysis

A heatmap and correlation matrix revealed that **Glucose**, **BMI**, and **Age** had the highest correlations with diabetes onset, while variables like **Blood Pressure** and **Skin Thickness** showed weaker correlations.

### 5.2 Model Performance Metrics

* **Accuracy**: 0.74 (The model correctly predicted the outcome 74% of the time).
* **Precision**: 0.59 for class 1 (diabetic), indicating that 59% of individuals predicted as diabetic actually had diabetes.
* **Recall**: 0.63 for class 1, indicating that the model identified 63% of the actual diabetic cases.
* **F1-Score**: 0.61 for class 1, reflecting a balance between precision and recall.

## 6. Findings and Discussion

The findings from this research demonstrate that Logistic Regression is a viable tool for predicting diabetes using readily available medical data. The model’s accuracy of 74% is relatively strong. **Glucose levels** and **BMI** were identified as the most significant predictors of diabetes, which aligns with medical research emphasizing their critical role in diabetes onset.

## 7. Conclusion

This study successfully applied Logistic Regression to predict the onset of diabetes, achieving an accuracy of 74%. The model provides valuable insights into the key factors contributing to diabetes, with **Glucose** and **BMI** being the most important predictors. Logistic Regression is a simple, interpretable model that can be implemented in real-time clinical settings to assist healthcare professionals in identifying high-risk patients for early intervention.

## 8. References

* Pima Indians Diabetes Dataset (UCI Machine Learning Repository)
* Scikit-learn: Pedregosa et al., (2011)
* International Diabetes Federation: Global Statistics Report