Diabetes Risk Prediction Using Machine Learning

Submitted for:

CSET211 - Statistical Machine Learning

Submitted by:

(E23CSEU2075) Uttkarsh Thakur

(E23CSEU2095) Dheer Pratap Singh

(E23CSEU2080) Khush Parashar

Submitted to:

(Prashant Kapil)

July-Dec 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY

THE TIMES GROUP

Index

Abstract

This project explores the application of machine learning techniques for predicting diabetes using the **PIMA Indian Diabetes Dataset**. Diabetes is a chronic condition that affects millions worldwide, and its early detection plays a vital role in preventing severe complications. The primary aim of this project is to design a predictive model that can identify individuals at risk of diabetes based on their physiological and demographic data.

The project follows a structured methodology starting with a detailed analysis of the dataset to identify and address data inconsistencies, including erroneous zero values in critical features like glucose levels, BMI, and insulin. Rigorous preprocessing steps were implemented, such as replacing invalid entries with mean values and standardizing features to ensure consistency. Logistic Regression was selected as the core algorithm due to its balance between simplicity, computational efficiency, and interpretability.

The model underwent hyperparameter tuning using Grid Search CV to optimize its performance, yielding a training accuracy of **77.59%** and a testing accuracy of **72.91%**. Evaluation metrics such as precision (80.49%), recall (77.95%), and F1-score (79.20%) highlight the model's ability to deliver reliable predictions. A confusion matrix analysis further demonstrated the model's effectiveness in balancing true positives and negatives.

This project provides valuable insights into the role of proper data handling and feature engineering in building robust machine learning models. The findings emphasize the potential of Logistic Regression as a predictive tool in healthcare. Future enhancements include addressing class imbalances, experimenting with advanced algorithms, and deploying the model for real-world applications.

All project materials, including the dataset, implementation code, presentations, and detailed documentation, are available in the GitHub Repository. This work serves as a foundation for further exploration and improvement in healthcare analytics, focusing on early intervention for diabetes management.

Introduction

Diabetes is a chronic illness affecting millions globally. It poses severe health risks, including cardiovascular diseases and organ failure. Early detection is crucial to mitigating these risks and improving patient outcomes. Machine learning techniques offer a powerful means of identifying patterns and making accurate predictions based on patient data.

This project focuses on predicting diabetes using Logistic Regression, a robust and interpretable classification algorithm. The goal was to analyze the relationship between various health indicators (e.g., glucose levels, BMI, age) and diabetes occurrence. Contributions include:

- Rigorous data preprocessing to address gaps and errors in the dataset.
- Implementation of feature scaling for enhanced model performance.
- Use of Grid Search CV for hyperparameter tuning, ensuring an optimal model configuration.

The project aims to bridge the gap between theoretical machine learning concepts and practical healthcare applications, providing a valuable tool for early diabetes detection.

Related Survey

Numerous studies have explored machine learning techniques for diabetes prediction. Algorithms such as Support Vector Machines (SVM), Decision Trees, and Neural Networks are frequently applied due to their high accuracy. However, these methods often lack interpretability and require extensive computational resources.

Logistic Regression, although simpler, provides several advantages:

- It is computationally efficient, making it suitable for real-time applications.
- It offers straightforward interpretability, allowing healthcare professionals to understand the relationship between features and predictions.
- It is robust when combined with proper preprocessing techniques and parameter optimization.

Existing research has highlighted the importance of handling data inconsistencies and outliers to improve prediction accuracy. This project builds on these insights by emphasizing rigorous data preprocessing and employing Logistic Regression to balance accuracy, efficiency, and interpretability.

Datasets

The PIMA Indian Diabetes Dataset, sourced from the UCI repository, was utilized for this project. It comprises 768 instances and 9 attributes, including physiological and demographic features such as pregnancies, glucose levels, blood pressure, BMI, and age. The target variable (Outcome) is binary, indicating whether an individual is diabetic (1) or non-diabetic (0).

3.1 Data Preprocessing

Data preprocessing is a critical step to ensure the quality and reliability of the model's predictions. The following steps were undertaken:

1. Handling Missing Values: No null values were found; however, certain features, such as *Glucose* and *Insulin*, contained zero entries, which are biologically implausible. These were replaced with their respective means to maintain data integrity.
2. Outlier Detection and Handling: Outliers were identified, particularly in the *Insulin* feature. These were addressed by replacing extreme values with the median to reduce their influence on the model.
3. Feature Scaling: Standardization was applied to ensure all features had a mean of zero and a standard deviation of one, improving model convergence and performance.

By addressing these issues, the dataset was prepared for effective training and testing.

Methodology

The project methodology involved a systematic approach to data analysis and model implementation:

1. Data Splitting:
   The dataset was divided into training (75%) and testing (25%) sets using stratified sampling to maintain balanced class distribution.
2. Model Selection:
   Logistic Regression was chosen due to its linear decision boundary, computational efficiency, and interpretability. Its suitability for small to medium-sized datasets made it an ideal choice for this project.
3. Hyperparameter Tuning:
   To optimize the model, Grid Search CV was employed. The parameters tuned included:
   - C (inverse of regularization strength),
   - penalty (L1 or L2 regularization),
   - solver (optimization algorithm).
     This ensured the model achieved a balance between bias and variance, enhancing its generalizability.

Hardware and Software Requirements

- Hardware:
  - ASUS TUF A15 Laptop with AMD Ryzen processor.
  - Minimum requirements: 16 GB RAM, octa-core CPU, 512 GB SSD.
- Software:
  - Programming Language: Python 3.12+
  - Libraries: pandas, NumPy, scikit-learn, matplotlib, seaborn.
  - Tools: Jupyter Notebook for implementation.

4.2 Performance Metrics

The model was evaluated using the following metrics:

- Accuracy: The proportion of correctly predicted instances.
- Precision: The ratio of true positives to all predicted positives, highlighting the reliability of positive predictions.
- Recall: The ratio of true positives to all actual positives, indicating the model's ability to identify positive cases.
- F1-Score: The harmonic means of precision and recall, balancing both metrics.
- Confusion Matrix: A tabular representation of true positives, true negatives, false positives, and false negatives.

Results and Analysis

The results of the Logistic Regression model are summarized below:

1. Preprocessed Dataset:
    - Zero entries in features such as *Glucose* and *BMI* were replaced with their means.
    - Outliers were addressed using median replacement.
    - Standardization ensured uniform feature scaling.
2. Model Training and Evaluation:
    - Best parameters obtained from Grid Search CV:

Arduino

{'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}

    - Training Accuracy: 77.59%
    - Testing Accuracy: 72.91%
3. Confusion Matrix:

    [[99, 24],

     [28, 41]]

The model exhibited a good balance between true positives and true negatives.

4. Performance Metrics:
    - Precision: 80.49%
    - Recall: 77.95%
    - F1-Score: 79.20%

5. Insights:
    - Proper preprocessing significantly enhanced model performance.
    - The slightly lower recall indicates a need for techniques to handle class imbalance, such as SMOTE.

Conclusions and Future Works

The Logistic Regression model demonstrated its potential as a predictive tool for diabetes, balancing simplicity with efficiency. The project highlights the importance of preprocessing and hyperparameter tuning in achieving optimal performance.

Future Directions:

1. Advanced Algorithms: Experimenting with ensemble methods like Random Forest or Gradient Boosting.
2. Class Imbalance: Applying techniques such as SMOTE or oversampling to improve recall.
3. Feature Expansion: Incorporating external datasets or additional features to enhance predictive power.
4. Model Deployment: Developing a web-based mobile application for easy accessibility and real-world usage.

The project sets a strong foundation for future research and practical applications in healthcare analytics.