Diabetes Prediction Using Machine Learning

Zagros Baban

May 16, 2025

Abstract

Diabetes is a common disease that affects many people all over the world. If we find diabetes early and treat it quickly, we can stop many health problems and save money for both patients and hospitals. The Pima Indians Diabetes dataset is famous for testing machine learning models. It contains information about patients and their health measurements. In this project, we use machine learning to predict who might have diabetes. We start by studying the data, then build and test a model to make predictions.

1 Introduction

Diabetes is a prevalent chronic condition with significant global impact. Early detection and timely intervention are crucial for mitigating severe health complications and reducing healthcare costs for both individuals and healthcare systems. The Pima Indians Diabetes dataset is a widely recognized benchmark for evaluating machine learning models in this domain, containing various patient health metrics.

This project leverages machine learning techniques to predict the likelihood of diabetes. The workflow encompasses data exploration, model development, and performance evaluation. Team contributions were as follows:

- **Data Scientist:** Responsible for data cleaning, exploratory data analysis (EDA) including chart generation, model building, and results validation.
- **ML Engineer:** Focused on writing, optimizing, and refining the machine learning code.
- **Documentation Specialist:** Tasked with authoring this report and creating all accompanying visualizations.

2 Related Work

Numerous researchers have applied machine learning to predict diabetes. The Pima dataset is frequently employed with a range of models, from simpler approaches like logistic regression and decision trees to more complex ones such as support vector machines (SVM), random forests, and neural networks.

- **Smith et al. (1988)** [1]: Utilized logistic regression for diabetes prediction with this dataset.
- **Sisodia & Sisodia (2018)** [2]: Compared SVM, Decision Trees, and Naive Bayes, finding SVM to yield the best performance.
- **Kavakiotis et al. (2017)** [3]: Conducted a review of multiple studies, highlighting Random Forest as an often effective model for this type of data.

Common metrics for evaluating model quality include accuracy, precision, recall, F1-score, and the confusion matrix. Most literature emphasizes the importance of preliminary data exploration through visualizations before model construction.

3 Proposed Method

3.1 Problem Definition

The objective is to predict whether an individual has diabetes based on their health data. This is formulated as a binary classification problem, where the outcome is either diabetic or non-diabetic.

3.2 Dataset Overview

- **Source:** UCI Machine Learning Repository [4].
- **Number of Instances:** 768, exclusively women of Pima Indian heritage.
- **Data Columns:** The dataset includes the features described in Table 1.

Table 1: Dataset Features and Descriptions

Feature	**Description**
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body Mass Index (weight in kg / (height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age	Age in years
Outcome	Class variable (0 = No diabetes, 1 = Diabetes)

3.3 Data Preprocessing

- **Missing Values:** Certain physiological measurements (e.g., Glucose, BloodPressure, SkinThickness, Insulin, BMI) cannot realistically be zero. Such zero values were treated as missing and imputed using the median value for each respective column.
- **Feature Scaling:** All numerical features were scaled to a range between 0 and 1 (Min-Max scaling) to ensure that all features contribute equally to the model training process and to improve model convergence.
- **Splitting Data:** The dataset was divided into a training set (80% of the data) and a testing set (20% of the data) to evaluate the model's performance on unseen data.

3.4 Exploratory Data Analysis (EDA)

Key visualizations generated to understand the data include:

1. **Histograms:** To show the distribution of each feature (e.g., BMI, Age).
2. **Correlation Heatmap:** To visualize the relationships between features and their correlation with the diabetes outcome.
3. **Age Distribution by Outcome:** To observe if age is a differentiating factor for diabetes. This typically shows that older individuals are more prone to diabetes.

4. **BMI vs Glucose Scatter Plot:** To identify patterns; often, individuals with higher BMI and glucose levels are more likely to have diabetes.
5. **Outcome Countplot:** To display the class distribution (number of diabetic vs. non-diabetic individuals).
6. **Boxplots by Outcome:** To compare the distribution of health values (e.g., Glucose, BMI) between diabetic and non-diabetic groups.

These charts aid in understanding data characteristics and guiding feature selection for the model.

3.5 Machine Learning Pipeline

Why Random Forest? We selected Random Forest as our primary model due to its numerous advantages for this type of dataset:

- **Reduced Overfitting:** By aggregating predictions from multiple decision trees, it is less prone to overfitting compared to a single decision tree.
- **Handles Complex Data:** Random Forest can capture non-linear relationships and complex interactions between features.
- **Feature Importance:** It provides a measure of feature importance, indicating which health attributes are most influential in predicting diabetes.
- **Effective with Small Datasets:** It generally performs well even when the dataset size is not excessively large.
- **Robust to Preprocessing Variations:** It can handle missing values to some extent and is less sensitive to feature scaling compared to other algorithms.

Furthermore, existing research, as noted in Section 2, supports Random Forest's efficacy for medical datasets.

- **Model Used:** Random Forest Classifier.
- **Evaluation Metrics:** Accuracy, precision, recall, F1-score, and confusion matrix.
- **Feature Importance:** Extracted from the trained Random Forest model.

4 Results

4.1 Dataset Statistics

Descriptive statistics for the features in the dataset (prior to zero-value imputation for relevant columns and scaling) are presented in Table 2.

Table 2: Descriptive Statistics of Dataset Features

Feature	Mean	Std	Min	Max
Pregnancies	3.85	3.37	0	17
Glucose	120.9	31.9	44	199
BloodPressure	69.1	19.4	24	122
SkinThickness	20.5	16.0	7	99
Insulin	79.8	115.2	14	846
BMI	31.9	7.9	18.2	67.1
DiabetesPedigreeFunction	0.47	0.33	0.078	2.42
Age	33.2	11.8	21	81

Note: Statistics for Glucose, BloodPressure, SkinThickness, Insulin, and BMI are shown before imputation of zero values, if applicable, or should be clarified if they are post-imputation.

4.2 EDA Insights

- Analysis indicated that older individuals and those with higher glucose levels have a greater likelihood of diabetes.
- The dataset exhibits a class imbalance, with approximately 65% non-diabetic individuals and 35% diabetic individuals.
- Glucose and BMI were identified as features most strongly correlated with the diabetes outcome.

4.3 Model Performance

The Random Forest model achieved the performance metrics shown in Table 3.

Table 3: Model Performance Metrics

Metric	**Score**
Accuracy	74.0%
Precision (No Diabetes)	0.80
Precision (Diabetes)	0.63
Recall (No Diabetes)	0.79
Recall (Diabetes)	0.65
F1-Score (No Diabetes)	0.80
F1-Score (Diabetes)	0.64

Confusion Matrix

The confusion matrix for the test set is presented in Table 4.

Table 4: Confusion Matrix

		Predicted	
		0 (No Diabetes)	**1 (Diabetes)**
Actual	**0 (No Diabetes)**	79	20
	1 (Diabetes)	19	36

Most Important Features: The Random Forest model identified Glucose, BMI, and Age as the features with the highest impact on predictions.

4.4 Visualizations

(This section would typically include the actual chart images. For this text-based LaTeX, we list them as described by the user.)

- **Histograms:** Illustrating the distribution of health features.
- **Correlation Heatmap:** Showing inter-feature correlations and their relation to diabetes.
- **Boxplots by Outcome:** Comparing health measurement distributions for diabetic and non-diabetic individuals.
- **Outcome Countplot:** Visualizing the class imbalance in the dataset.

- **Scatter Plot (BMI vs Glucose):** Highlighting potential clusters for individuals with and without diabetes based on these two key features.

Example placeholder for an image:

5 Conclusions

This project demonstrates the utility of machine learning and thorough data analysis in identifying individuals at risk of diabetes. The Random Forest model yielded an accuracy of 74.0% on the test set. The most influential features for prediction—Glucose, BMI, and Age—align with established clinical knowledge.

Future work could explore the following avenues:

- **Addressing Class Imbalance:** Employ techniques such as resampling (oversampling the minority class, undersampling the majority class) or synthetic data generation (e.g., SMOTE) to balance the dataset.
- **Exploring Alternative Models:** Evaluate other algorithms like Support Vector Machines (SVM), Gradient Boosting, or Neural Networks.
- **Data Augmentation:** Incorporate additional data from other sources, if feasible, to enhance model robustness and generalizability.
- **Deployment:** Develop a simple web application or interface to allow clinicians to utilize the model as a decision support tool.

It is imperative to remember that such predictive tools are intended to assist medical professionals, not to replace their clinical judgment.

6 References

References

[1] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

[2] Sisodia, D., Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[4] Dua, D. Graff, C. (2019). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Dataset available at: https://archive.ics.uci.edu/dataset/38/pima+indians+diabetes