Diabetes Prediction Using ML

Avula Thanu sree

July 1, 2024

Abstract

This report outlines the development of a machine learning model to predict diabetes using the PIMA diabetes dataset. The model employs a Support Vector Machine (SVM) classifier, demonstrating significant accuracy in distinguishing between diabetic and non-diabetic patients based on various health metrics.

1 Introduction

Diabetes is a chronic medical condition characterized by high levels of glucose in the blood. Early detection and management are crucial in preventing severe complications. This project aims to create a predictive model that can classify individuals as diabetic or non-diabetic using the PIMA diabetes dataset, which contains several health-related attributes.

2 Dataset Description

The PIMA diabetes dataset includes 768 instances and 9 attributes. These attributes are:

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- **Blood Pressure**: Diastolic blood pressure (mm Hg)
- **Skin Thickness**: Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)**Diabetes Pedigree Function** : $Diabetespedigreefunction$
- **Age**: Age (years)
- **Outcome**: Class variable (0 or 1) indicating whether the patient is diabetic

3 Data Loading and Exploration

The dataset is loaded using the `pandas` library. Initial exploration involves displaying statistical measures (mean, standard deviation, etc.) to understand the data distribution. The first few rows are inspected to get a sense of the data structure.

3.1 Data Dimensions and Outcome Distribution

The dataset consists of 768 rows and 9 columns. The outcome distribution reveals the balance between diabetic (268 instances) and non-diabetic (500 instances) cases. This information is critical for understanding the class balance and potential biases in the model.

4 Data Analysis

Mean values of the attributes are calculated separately for diabetic and non-diabetic patients. This analysis helps in identifying patterns and differences between the two groups. For instance, diabetic patients might have higher average glucose levels compared to non-diabetic patients.

5 Data Preprocessing

5.1 Feature and Label Separation

The features and labels are separated, with the features (X) consisting of all columns except the outcome, and the labels (y) being the outcome column.

5.2 Data Standardization

Standardization is applied to the features to ensure that each attribute has a mean of 0 and a standard deviation of 1. This step is crucial for improving the performance of the SVM classifier, which is sensitive to the scale of input data.

6 Model Training and Evaluation

6.1 Data Splitting

The dataset is split into training and testing sets using an 80-20 split. Stratified sampling ensures that the training and testing sets have a similar distribution of the outcome variable.

6.2 Support Vector Machine Classifier

A Support Vector Machine (SVM) with a linear kernel is used for classification. The model is trained on the training set and evaluated on both training and testing sets.

6.3 Accuracy Metrics

The accuracy of the model on the training set is approximately 79%, while the accuracy on the testing set is around 77%. These metrics indicate the model's effectiveness in predicting diabetes.

7 Model Prediction

To demonstrate the model's prediction capability, a sample input (e.g., an individual with specific health metrics) is standardized and passed to the trained model. The model then predicts whether the individual is diabetic or non-diabetic based on the input data.

8 Results and Discussion

The SVM classifier shows promising results with good accuracy on both training and testing datasets. The performance can be attributed to effective data standardization and appropriate feature selection. However, further improvements could be achieved by exploring other models, tuning hyperparameters, or addressing class imbalances through techniques like SMOTE (Synthetic Minority Over-sampling Technique).

9 Conclusion

This project successfully demonstrates the application of an SVM classifier in predicting diabetes using the PIMA diabetes dataset. The model's accuracy underscores its potential as a reliable tool for early diabetes detection. Future work could focus on enhancing the model's performance and exploring its integration into real-world healthcare systems.

10 Future Work

- **Hyperparameter Tuning**: Experimenting with different SVM kernels and tuning hyperparameters to improve model performance.
- **Feature Engineering**: Creating new features or selecting the most important features to enhance predictive accuracy.

- **Addressing Class Imbalance**: Implementing techniques such as SMOTE to balance the classes and improve model robustness.
- **Model Comparison**: Evaluating other machine learning models like Random Forest, Gradient Boosting, or Neural Networks for potential performance gains.

11 References

- PIMA Indians Diabetes Database on Kaggle
- Scikit-learn documentation: https://scikit-learn.org/stable/