

Diabetes Prediction Model

Abstract

This project focuses on developing a machine learning model to predict diabetes based on health parameters.

Using a dataset containing 768 entries with nine features, I implemented a Support Vector Machine (SVM) classifier to achieve an effective predictive system. The model demonstrated high accuracy during training and testing and can be used to classify individuals as diabetic or non-diabetic based on their health metrics.

Introduction

Diabetes is a chronic condition that affects millions worldwide. Early detection is crucial to managing and reducing health risks. Machine learning models can assist in predicting diabetes by analyzing various health parameters.

This project aims to design a predictive system using SVM, leveraging a publicly available dataset to train and evaluate the model.

Dataset Details

The dataset used for this project has the following characteristics:

- Source: Pima Indians Diabetes Dataset
- Size: 768 rows and 9 columns
- Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age
- Target Variable: Outcome (0: Non-Diabetic, 1: Diabetic)

Methodology

1. Data Preprocessing:

- Loaded the dataset using pandas.
- Applied the StandardScaler to normalize the features.
- Split the data into training (80%) and testing (20%) sets using train_test_split with stratification.

2. Model Training:

- Used a Support Vector Machine with a linear kernel for training.

3. Model Evaluation:

- Training Accuracy: 78.92%
- Testing Accuracy: 77.92%

4. Predictive System:

- Accepts user inputs, standardizes them, and predicts outcomes.

5. Model Saving:

- The trained model was saved as trained_model.sav using pickle.

Results

- Training Accuracy: 78.92%
- Testing Accuracy: 77.92%
- The model demonstrates reliable performance for early-stage diabetes prediction.

Predictive System

The predictive system allows users to input health metrics and returns a classification.

1. Input Example:

- Pregnancies: 2, Glucose: 112, Blood Pressure: 66, Skin Thickness: 22, Insulin: 0, BMI: 25,

Diabetes Pedigree Function: 0.307, Age: 24

2. Prediction: The person is diabetic/non-diabetic.

Additionally, a Streamlit-based web application was developed to:

- Allow users to input health metrics interactively.

- Display predictive results instantly upon submission.

Conclusion

This diabetes prediction model demonstrates the application of machine learning in healthcare. The SVM classifier achieved good accuracy and can be enhanced with a larger dataset and advanced feature engineering. The Streamlit interface ensures accessibility and usability for non-technical users.

References

1. Python Libraries: pandas, numpy, scikit-learn, matplotlib, pickle, streamlit
2. Dataset Source: Pima Indians Diabetes Dataset

Prepared by

Saran B (2205102)

3rd B.Sc. Computer Science

Government Arts College (Autonomous), Karur, India