README: Diabetes Prediction Using SVM

Avula Thanu Sree

July 1, 2024

1 Project Overview

This project involves building a machine learning model to predict diabetes using the PIMA diabetes dataset. The model utilizes a Support Vector Machine (SVM) classifier to achieve high accuracy in identifying diabetic patients based on various health metrics.

2 Dataset

The dataset used in this project is the PIMA Indians Diabetes Database, which is available on Kaggle. It consists of 768 samples with 9 features:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration after 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)$DiabetesPedigreeFunction :$ $Afunctionthatscoreslikelihoodofdiabetesbasedonfamilyhistory$

- Age: Age of the patient (years)
- Outcome: Class variable (0 or 1) indicating if the patient is diabetic

3 Installation

To run this project, you need to have Python and the following libraries installed:

- pandas

- numpy
- seaborn
- matplotlib
- scikit-learn

You can install the required libraries using pip:

```
pip install pandas numpy seaborn matplotlib scikit-learn
```

4 Usage

1. Clone the repository:

```
git clone https://github.com/yourusername/diabetes-prediction.git
cd diabetes-prediction
```

2. Ensure the dataset ‘diabetes.csv‘ is in the project directory.
3. Run the script:

```
python diabetes_prediction.py
```

The script will load the dataset, train an SVM classifier, evaluate its accuracy, and make a prediction for a sample input.

5 Results

The model achieves the following accuracy:

- Training Data Accuracy: 79%
- Testing Data Accuracy: 77%

6 Prediction Example

To predict the outcome for a new individual, modify the sample input in the script with the person’s health metrics. The model will output whether the individual is diabetic or non-diabetic based on the input data.

7 Contributing

Contributions are welcome! Please create a pull request or open an issue to discuss your changes.

8 License

This project is licensed under the MIT License.

9 References

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