Diabetes Prediction Model

Overview

This project aims to predict the likelihood of diabetes in patients using the PIMA Diabetes Dataset. The dataset contains various medical predictor variables and one target variable, Outcome, which indicates whether the patient has diabetes (1) or not (0). The project involves data preprocessing, model training using a Support Vector Machine (SVM), and evaluation of the model's performance.

Dataset

The dataset used in this project is the **PIMA Diabetes Dataset**, which contains 768 rows and 9 columns. The columns include:

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skinfold 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
- **Age**: Age (years)
- **Outcome**: Class variable (0 or 1)

Project Steps

1. Importing Libraries

The necessary libraries for data processing, model training, and evaluation are imported:

- **NumPy**: For numerical operations.
- **Pandas**: For data manipulation and analysis.
- **StandardScaler**: For standardizing the dataset.
- **train_test_split**: For splitting the dataset into training and testing sets.
- **SVM**: For training the Support Vector Machine model.
- **accuracy_score**: For evaluating the model's accuracy.

2. Data Collection and Analysis

The dataset is loaded into a Pandas DataFrame, and basic statistical measures are analyzed to understand the data distribution.

3. Data Preprocessing

The dataset is split into features (X) and labels (Y). The features are standardized to ensure that all variables contribute equally to the model.

4. Splitting the Data

The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.

5. Model Training

A Support Vector Machine (SVM) classifier with a linear kernel is trained on the training data.

6. Model Evaluation

The model's accuracy is evaluated on both the training and testing datasets to ensure it generalizes well to unseen data.

7. Making Predictions

The trained model is used to make predictions on new data. The input data is standardized before making the prediction.

Results

- **Training Data Accuracy**: Approximately 78.66%
- **Test Data Accuracy**: Approximately 77.27%

The model performs well on both the training and test datasets, indicating that it generalizes well to unseen data.

Conclusion

This project demonstrates the use of a Support Vector Machine (SVM) for predicting diabetes based on medical data. The model achieves a reasonable accuracy and can be further improved by tuning hyperparameters or using more advanced techniques.

How to Run the Code

1. Clone the repository.
2. Install the required libraries using pip install -r requirements.txt.
3. Run the Jupyter notebook Diabetes_Prediction.ipynb.

Requirements

- Python 3.x
- NumPy
- Pandas
- Scikit-learn

Acknowledgments

- The PIMA Diabetes Dataset is sourced from Kaggle.
