Title: Diabetes Prediction Using Support Vector Machine Classifier

Abstract

Diabetes is a chronic disease affecting millions worldwide. Early detection can prevent severe complications and improve patient outcomes. This study proposes a machine learning model using Support Vector Machine (SVM) to predict diabetes based on patient data. Using the PIMA Indian Diabetes dataset, the model demonstrates an accuracy of approximately 78% on test data, which indicates promising results for this methodology in preliminary diabetes screening.

1. Introduction

Diabetes is a prevalent condition that impairs the body's ability to regulate blood sugar levels. Undiagnosed or uncontrolled diabetes can lead to serious health issues, including cardiovascular disease, kidney damage, and neuropathy. Early and accurate prediction of diabetes can significantly reduce these risks by prompting timely medical intervention. This study explores a supervised machine learning approach using the Support Vector Machine (SVM) classifier to predict diabetes. We aim to determine the model's predictive power and accuracy when analyzing patient health data features.

2. Dataset and Preprocessing

The dataset used is the **PIMA Indian Diabetes dataset**, which includes various health indicators for 768 individuals, including those related to pregnancy history, glucose levels, blood pressure, BMI, age, and diabetes pedigree function. Each row in the dataset corresponds to a single patient, and the target variable, labeled "Outcome," denotes whether a patient is diabetic (1) or non-diabetic (0).

2.1 Data Cleaning and Exploration

The dataset was loaded and examined for shape, statistical summary, and distribution of the outcome variable. An initial inspection revealed the presence of values that may need further exploration for outliers or missing values (e.g., 0 values in columns where such values are biologically implausible).

2.2 Data Standardization

To improve model performance, we standardized the feature variables using StandardScaler. Standardization scales the data to have zero mean and unit variance, which is essential for SVM as it ensures that all features contribute equally to the distance calculations in the SVM algorithm.

```
# Code for Standardization

Scalar = StandardScaler()

standar_data_without_outcome =
Scalar.fit_transform(new_colom_without_outcome)
```

3. Methodology

Support Vector Machine (SVM) is chosen as the primary classification algorithm due to its effectiveness in handling high-dimensional data. The SVM algorithm works by finding a hyperplane that best separates the classes. For this study, a linear kernel was selected, which is suitable for linearly separable data.

3.1 Model Training

The dataset was split into training and testing sets in an 80-20 ratio with stratified sampling to ensure an equal proportion of diabetic and non-diabetic cases in each set. After partitioning, the model was trained on the standardized data using the SVM classifier with a linear kernel.

```
# Code for Training the SVM Model

classifier = svm.SVC(kernel='linear')
```

```
classifier.fit(x_train, y_train)
```

4. Results and Evaluation

Model evaluation was conducted using accuracy scores on both training and test sets. Accuracy measures how well the model correctly predicts outcomes for each set. After training, predictions were made for both the training and test datasets.

```
# Code for Prediction and Accuracy Evaluation

x_train_prediction = classifier.predict(x_train)

training_data_accurancy = accuracy_score(x_train_prediction, y_train)

x_test_prediction = classifier.predict(x_test)

test_data_accurancy = accuracy_score(x_test_prediction, y_test)
```

Output

- **Training Accuracy**: 0.78
- **Testing Accuracy**: 0.77

The results demonstrate that the model performs well on both the training and test data, indicating no significant overfitting.

5. Predictive System

A predictive system was built to assess individual cases. Given a set of patient data, the model predicts whether the person is diabetic.

```
# Code for Predictive System

input_data = (8, 183, 64, 0, 0, 23.3, 0.672, 32)

input_data_as_numpy_array = np.asarray(input_data).reshape(1, -1)

std_data = Scalar.transform(input_data_as_numpy_array)
```

```
prediction = classifier.predict(std_data)
```

This predictive system allows for practical application in healthcare by potentially aiding in early diagnosis.

6. Discussion

The model achieved a relatively high accuracy, indicating its effectiveness in predicting diabetes based on clinical features. However, the model could be improved by addressing the limitations in data quality and considering alternative machine learning algorithms for comparison.

7. Conclusion

This study presents a machine learning approach to predict diabetes using an SVM classifier. The results suggest that the model can serve as a reliable preliminary diagnostic tool, with further refinement needed for clinical applications. Future research may explore additional feature engineering, hyperparameter tuning, and integration with other algorithms to enhance predictive accuracy.

References

(1) Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

(2) American Diabetes Association. (2021). Statistics About Diabetes. Retrieved from https://www.diabetes.org/