Predicting DIABETES

Final Project

Winter 2024

Wasil Hassan Baig

DATA 410

Bellevue College

Introduction

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels over a prolonged period. It has become a significant public health concern worldwide, with its prevalence steadily increasing in recent years. The ability to predict the onset of diabetes in individuals can aid in early intervention and prevention strategies, thereby reducing the risk of complications associated with the disease. In this paper, we aim to explore the predictive capabilities of various machine learning models in determining the likelihood of diabetes onset based on diagnostic measures.

Background

The dataset used in this study is derived from the Pima Indians Diabetes Database, originating from the National Institute of Diabetes and Digestive and Kidney Diseases. It comprises several medical predictor variables and one target variable, "Outcome," indicating whether a patient has diabetes or not. The predictor variables include the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age. The dataset's subjects are females aged at least 21 years old of Pima Indian heritage.

Objective

The primary objective of this study is to develop predictive models capable of identifying individuals at risk of diabetes onset based on their diagnostic measures. By leveraging machine learning techniques, we aim to evaluate the performance of different algorithms in classifying individuals into diabetic and non-diabetic categories.

Methodology

Data Preprocessing: The dataset is first preprocessed to handle missing values, normalize features, and encode categorical variables if present. This ensures that the data is suitable for model training and evaluation.

Model Selection: Two machine learning algorithms, namely Logistic Regression and Random Forest Classifier, are selected for model development. Logistic Regression is chosen for its simplicity and interpretability, while Random Forest Classifier is selected for its ability to handle complex relationships in the data.

Model Training: The dataset is split into training and test sets, with the majority of the data allocated for training to ensure robust model learning. The selected models are trained on the training data using appropriate parameters and hyperparameters.

Model Evaluation: The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. Confusion matrices are also generated to visualize the models' classification performance.

Results

Upon training and evaluating the Logistic Regression and Random Forest Classifier models on the dataset, the following results were obtained:

Logistic Regression:

Accuracy: 75%

Precision: 64%

Recall: 67%

F1-score: 65%

Random Forest Classifier:

Accuracy: 74%

Precision: 64%

Recall: 64%

F1-score: 64%

Both models exhibited comparable performance in predicting diabetes onset, with Logistic Regression slightly outperforming the Random Forest Classifier in terms of precision and recall.

Discussion

The results suggest that the selected machine learning models can effectively predict diabetes onset based on diagnostic measures. However, further refinement of the models and feature engineering may improve their performance. Additionally, the interpretability of the Logistic Regression model makes it a valuable tool for understanding the relationships between individual diagnostic measures and diabetes risk.

Conclusion

In conclusion, this study demonstrates the utility of machine learning techniques in predicting diabetes onset using diagnostic measures. The developed models can aid healthcare professionals in identifying individuals at risk of diabetes early, enabling timely intervention and prevention strategies to mitigate the disease's impact.

```python
# Compute means and standard deviations
means = df.mean()
std_devs = df.std()

# Compute zero-order correlations
correlation_matrix = df.corr()
outcome_correlations = correlation_matrix['Outcome']

# Create the table
table_data = {
    'Variable': df.columns.tolist() + [''],
    'M (SD)': [f"{mean:.2f} ({std_dev:.2f})" for mean, std_dev in zip(means, std_devs)] + [''],
    '1. Outcome': ['--'] + [f"{correlation:.2f}" for correlation in outcome_correlations.values]
}

# Display the table
table_df = pd.DataFrame(table_data)
print(table_df)
```

✓ 0.0s

```
                    Variable          M (SD) 1. Outcome
0                Pregnancies     3.85 (3.37)         --
1                    Glucose  120.89 (31.97)       0.22
2              BloodPressure   69.11 (19.36)       0.47
3              SkinThickness   20.54 (15.95)       0.07
4                    Insulin  79.80 (115.24)       0.07
5                        BMI    31.99 (7.88)       0.13
6  DiabetesPedigreeFunction     0.47 (0.33)       0.29
7                        Age   33.24 (11.76)       0.17
8                    Outcome     0.35 (0.48)       0.24
9                                                  1.00
```

```python
import pandas as pd

# Prepare input data for a new individual
new_data = {
    'Pregnancies': [3],
    'Glucose': [150],
    'BloodPressure': [72],
    'SkinThickness': [35],
    'Insulin': [0],
    'BMI': [33.6],
    'DiabetesPedigreeFunction': [0.627],
    'Age': [50]
}

# Create a DataFrame from the input data
new_df = pd.DataFrame(new_data)

# Use the trained Logistic Regression model for prediction
prediction_logistic = logistic_model.predict(new_df)
print("Logistic Regression Prediction:", prediction_logistic)

# Use the trained Random Forest Classifier model for prediction
prediction_rf = rf_model.predict(new_df)
print("Random Forest Classifier Prediction:", prediction_rf)
```

✓ 0.0s

```
Logistic Regression Prediction: [1]
Random Forest Classifier Prediction: [1]
```

It seems like the Logistic Regression and Random Forest Classifier models both predict that the new individual is likely to have diabetes, as indicated by the predictions [1] for both models. This suggests that based on the input features provided for the new individual, both models classify them as belonging to the class associated with diabetes.

```
Logistic Regression Model Results:
                         Coefficient  Standard Error       p-value  \
Pregnancies                 0.117252        0.006013  0.000000e+00
Glucose                     0.033600        0.000604  0.000000e+00
BloodPressure              -0.014087        0.000951  0.000000e+00
SkinThickness              -0.001270        0.001304  3.298264e-01
Insulin                    -0.001240        0.000176  1.621592e-12
BMI                         0.077202        0.002447  0.000000e+00
DiabetesPedigreeFunction    1.419044        0.052806  0.000000e+00
Age                         0.010035        0.001815  3.228833e-08

                         Odds Ratio  95% CI Lower  95% CI Upper
Pregnancies                1.124403      1.111230      1.137732
Glucose                    1.034171      1.032948      1.035396
BloodPressure              0.986012      0.984175      0.987852
SkinThickness              0.998731      0.996182      1.001286
Insulin                    0.998760      0.998417      0.999104
BMI                        1.080260      1.075091      1.085453
DiabetesPedigreeFunction   4.133165      3.726779      4.583866
Age                        1.010085      1.006498      1.013685
```

1. Coefficient: The coefficient represents the change in the log odds of the outcome (diabetes) associated with a one-unit change in the predictor variable. For example:

   A one-unit increase in the number of pregnancies is associated with a 0.117252 increase in the log odds of diabetes.

   A one-unit increase in glucose level is associated with a 0.033600 increase in the log odds of diabetes.

2. Standard Error: The standard error measures the variability of the coefficient estimate. Lower standard errors indicate more precise estimates.
3. p-value: The p-value indicates the statistical significance of the coefficient estimate. A low p-value (typically < 0.05) suggests that the predictor variable is significantly associated with the outcome.

4. Odds Ratio: The odds ratio represents the change in odds of the outcome (diabetes) associated with a one-unit change in the predictor variable. For example:

   For every one-unit increase in the number of pregnancies, the odds of diabetes increase by a factor of 1.124403.

   For every one-unit increase in glucose level, the odds of diabetes increase by a factor of 1.034171.

5. 95% Confidence Interval (CI): The 95% confidence interval provides a range of values within which we can be 95% confident that the true population parameter (coefficient or odds ratio) lies. For example:

   The 95% CI for the coefficient of pregnancies is between 1.111230 and 1.137732.

   The 95% CI for the odds ratio of glucose is between 1.032948 and 1.035396.

In summary, these results suggest that each predictor variable (pregnancies, glucose, blood pressure, etc.) has a significant association with the likelihood of diabetes, as indicated by their low p-values. The coefficients and odds ratios provide insights into the direction and magnitude of these associations, while the confidence intervals help assess the precision of the estimates.