Implementation Report

DiabPredict - Predictive Modeling for Diabetes Onset

Team Members: Nagendra Madi Reddy & Vishwajeeth Balaji

In the provided code, the analysis focuses on a diabetes dataset using Python with libraries such as NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn. Let's break down the code into sections:

1) Dataset Upload and Exploration:

* The code begins by loading a diabetes dataset from a CSV file using pd.read_csv and displaying the first few rows using head().

* Column names are checked using diabetes_df.columns, and dataset information is obtained with diabetes_df.info().

2) Handling Missing Values:

* Initial exploration for missing values using isnull() is followed by a realization that 0 values represent missing data in certain columns.

* A copy of the dataset (diabetes_df_copy) is created to preserve the original data.

* The code replaces 0 values with NaN in specified columns ('Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI').

* The count of NaNs in each column is displayed using diabetes_df_copy.isnull().sum().

3) Data Distribution Plots Before and After Imputation:

* Histograms of the data distribution before and after handling missing values are plotted using hist().

4) Feature Correlation:

* A heatmap of the correlation matrix is generated to visualize the relationships between features.

5) Data Scaling:

The original dataset is displayed, and then feature scaling is performed using StandardScaler on selected columns.

6) Data Splitting:

* The dataset is split into features (X) and the target variable (y), followed by a train-test split using train_test_split.

* The features (X) comprise all columns excluding the target variable (Outcome), the target variable (y) is the variable to be predicted (Outcome), the test set constitutes 33% of the

dataset, and the random_state ensures reproducibility by providing a seed for the random number generator during the data splitting process.

7) Model Building - Random Forest:

* A Random Forest Classifier is trained and evaluated on both the training and test sets.

* The accuracy scores, confusion matrix, and classification report are displayed.

8) Model Building - Decision Tree:

* A Decision Tree Classifier is trained and evaluated similarly, with accuracy scores, confusion matrix, and classification report.

9) Model Building - Support Vector Machines (SVM):

* A Support Vector Machine Classifier is trained and evaluated, showing accuracy score, confusion matrix, and classification report.

10) Models Comparison:

* Random Forest, Decision Tree, and SVM models are trained and their accuracy scores are compared using a heatmap.

Each section of the code focuses on a specific aspect of the analysis, including data exploration, preprocessing, model building, and comparison.

Conclusion:

The DiabPredict predictive modeling analysis has been conducted using three different machine learning algorithms: Decision Tree, Random Forest, and Support Vector Machines (SVM). The primary evaluation metric, accuracy, is utilized to assess the performance of each model on predicting diabetes onset. The obtained accuracy scores are as follows:

* The Random Forest model exhibits the highest accuracy among the three models, reaching 78%. This suggests that the Random Forest algorithm, which leverages an ensemble of decision trees, demonstrates a robust ability to predict diabetes onset based on the provided features.

* The Decision Tree model follows with a respectable accuracy of 70%, indicating its effectiveness in capturing patterns within the dataset. However, it appears to be slightly outperformed by the Random Forest ensemble approach.

* The Support Vector Machines (SVM) model achieves an accuracy of 75%, positioning it between the Decision Tree and Random Forest models. SVM is known for its capability to handle complex decision boundaries, and its performance in this context aligns with expectations.

In conclusion, the DiabPredict predictive modeling analysis provides a promising foundation for predicting diabetes onset, with the Random Forest model standing out as the most accurate and reliable choice among the evaluated algorithms.

GITHUB LINK: https://github.com/Vishwajeeth2000/DiabPredict---Predictive-Modeling-for-Diabetes-Onset