Diabetes Prediction Using Machine Learning

Introduction

The objective of this project is to develop a machine learning model to predict whether an individual has diabetes based on various health-related attributes. The dataset used contains 520 instances with 16 features, including age, gender, and various symptoms related to diabetes. The goal is to preprocess the data, apply feature engineering techniques, and train multiple machine learning models to determine which performs best in predicting diabetes.

Data Preprocessing

Loading and Initial Exploration:

- The dataset contains 17 columns and 520 rows. The columns are: Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity, and class.
- Initial exploration shows no missing values, and the dataset is well-structured.

Data Cleaning and Encoding:

- Categorical variables such as Gender, Polyuria, Polydipsia, etc., were encoded using `LabelEncoder` to convert them into numerical values suitable for machine learning models.
- The target variable `class` was also encoded into a new column `encoded_class` and the original `class` column was dropped.

Feature Engineering

Scaling:

- Features were scaled using `MinMaxScaler` to normalize the data, ensuring all features contribute equally to the model training. This scaling process ensures that features with larger ranges do not dominate those with smaller ranges.

Correlation Analysis:

- A correlation matrix was generated to analyze the relationships between different features. The heatmap visualization helped in identifying any strong correlations among the features which might be redundant.

Data Visualization

- Seaborn heatmap was used to visualize missing values (none found).
- Another heatmap was created to visualize the correlation matrix of the dataset after encoding and scaling.

Model Selection and Training

Splitting the Dataset:

- The data was split into training and testing sets with an 80-20 ratio. Stratified sampling was used to maintain the proportion of the target variable in both sets.

Training Models: Various machine learning models were trained including:

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors (KNN)

Model Evaluation:

- The performance of each model was evaluated using accuracy and classification reports.

Results and Analysis

Training Accuracy:

- Decision Tree: 95.91%
- Random Forest: 97.84%
- SVM: 98.32%
- Logistic Regression: 92.55%
- Naive Bayes: 88.7%
- K-Nearest Neighbors: 95.67%

Testing Accuracy:

- Decision Tree: 94.23%
- Random Forest: 96.15%
- SVM: 98.08%
- Logistic Regression: 96.15%
- Naive Bayes: 93.27%
- K-Nearest Neighbors: 92.31%

Classification Reports:

- The classification reports for each model on the test set provided detailed insights into precision, recall, and f1-score for both classes (0 and 1). For instance, the SVM model achieved an accuracy of 98.08% with high precision and recall for both classes.

Discussion

Best Model: The SVM model outperformed others with the highest accuracy on both training (98.32%) and testing (98.08%) datasets.

Model Comparison: While SVM and Random Forest showed high accuracy, Logistic Regression also performed well with 96.15% accuracy on the test set. Naive Bayes had the lowest accuracy, highlighting its potential limitations for this dataset.

Overfitting: Despite high training accuracy, some models like the Decision Tree and Random Forest maintained strong performance on the test set, indicating minimal overfitting.

Conclusion

The project successfully applied various machine learning models to predict diabetes based on health-related attributes. SVM emerged as the best-performing model with the highest accuracy. Proper data preprocessing, feature engineering, and model evaluation were crucial in achieving high predictive performance. Future work could explore advanced feature selection techniques and hyperparameter tuning to further improve model accuracy and robustness.