Diabetes Prediction:

Abstract

This project investigates the application of machine learning techniques to predict diabetes based on clinical and lifestyle factors using data from the National Health and Nutrition Examination Survey (NHANES). The study explores models like K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machines (SVM), focusing on effective preprocessing, feature selection, and rigorous evaluation to achieve reliable predictions.

Introduction

Diabetes is a chronic condition affecting millions globally, with significant public health implications. It is characterized by high blood sugar levels resulting from either inadequate insulin production (Type 1 diabetes) or ineffective insulin use (Type 2 diabetes). If left untreated, diabetes can lead to severe complications such as heart disease, kidney failure, nerve damage, and vision impairment. Early detection is critical to managing and preventing complications.

Type 2 diabetes, which accounts for the majority of cases, is strongly linked to lifestyle factors such as diet, physical inactivity, and obesity. Preventative measures, including a healthy diet, regular exercise, and routine health screenings, play a vital role in reducing diabetes prevalence.

NHANES is a nationally representative program conducted by the Centers for Disease Control and Prevention (CDC), designed to assess the health and nutritional status of adults and children in the United States. The survey integrates interviews, laboratory tests, and physical examinations, providing valuable insights into the prevalence of various diseases, including diabetes. NHANES collects data on demographics, dietary habits, environmental exposures, and medical conditions, making it a comprehensive resource for public health research. This project seeks to utilize this rich dataset to address gaps in early diagnosis and prevention strategies.

Materials and Methods

Dataset Description

The dataset integrates multiple cycles of NHANES, containing diverse features:

- Demographics: Age, gender, race.
- Clinical Measurements: Fasting glucose, glycohemoglobin (HbA1c), HDL-cholesterol, BMI, and blood pressure.
- Lifestyle Factors: Dietary sugar intake and family history of diabetes.
- Biomarkers: Albumin-to-creatinine ratio, indicating kidney function.

NHANES also includes self-reported data on diabetes diagnosis and medication use, enabling the classification of individuals into diabetic and non-diabetic categories. This detailed information is crucial for training robust machine learning models.

Data Preprocessing

To ensure data quality and model performance, the following steps were taken:

1. Handling Missing Values: Imputed or removed missing data based on feature importance.
2. Scaling and Encoding: Standardized numerical features and applied one-hot encoding to categorical variables.
3. Train-Test Split: Divided the dataset into training (80%) and testing (20%) subsets.

Model Development

Three machine learning models were explored:

- K-Nearest Neighbors (KNN): Tuned hyperparameters such as the number of neighbors, distance metrics, and weighting schemes.
- Naive Bayes: Adjusted smoothing parameters to handle data variability.
- Support Vector Machines (SVM): Optimized kernel types, regularization parameters (C), and gamma values.

Evaluation Metrics

Performance was assessed using:

- Accuracy, precision, recall, and F1-score.
- ROC-AUC for evaluating the trade-off between sensitivity and specificity.
- Cross-validation to ensure generalization and robustness.

Results and Discussion

- KNN Model: Achieved an optimized accuracy, demonstrating good performance on balanced datasets.
- Naive Bayes: Provided faster predictions but slightly lower accuracy due to feature independence assumptions.
- SVM: Delivered the best overall performance, with high accuracy and an ROC-AUC score, highlighting its effectiveness for complex decision boundaries.

Key Findings

- Glycohemoglobin (HbA1c), fasting glucose, and BMI emerged as the most predictive features.
- Standardization and hyperparameter tuning significantly improved model outcomes.
- NHANES data revealed a higher prevalence of diabetes among individuals with elevated HbA1c levels and obesity, aligning with global trends.

Conclusion

This study underscores the potential of machine learning in diabetes prediction, leveraging clinical and lifestyle data for accurate risk assessment. The purpose of this project is to bridge the gap between data analysis and real-world applications, enabling healthcare providers to identify high-risk individuals early and improve intervention strategies. Future work will explore deep learning techniques and real-time applications to enhance usability in clinical settings.

References

1. Centers for Disease Control and Prevention (CDC). "National Health and Nutrition Examination Survey." Available at: https://www.cdc.gov/nchs/nhanes/
2. National Diabetes Statistics Report. (2022). Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/diabetes/data/statistics-report/index.html