Final Project

Group 10

Srinidhi Aduri
Sucharith Prasanna Krishna
Sameera Mandalika
Mandar Jadhav
Dat Tran

Aduri.s@northeastern.edu
Prasannakrishna.s@northeastern.edu
Mandalika.s@northeastern.edu
Jadhav.man@northeastern.edu
Tran.dat@northeastern.edu

Diabetic Health Status Classification

Goal

Diabetics is a prevalent condition in the United States due to unfavorable lifestyle and eating habits. This can lead to reduced life expectancy and complications like heart disease and vision loss.

The goal of the project is to build a model that can accurately classify patients as diabetic/pre-diabetic or healthy based on many factors that fall under healthcare, demographics and lifestyle information recorded through surveys. By predicting the health status of individuals, we can detect and prevent diabetics early.

Project Setting

The project utilizes the CDC Diabetics health indicators dataset containing **21** features and **~200,000** records that includes crucial health information like Blood Pressure, Cholesterol, BMI, and lifestyle factors like type of food consumption and mental health status. The target variable for classification is a binary indicator of whether the patient is diabetic. We will implement a **classification algorithm** on the dataset.

Data Dictionary*

Variable Name	Role	Type	Description	Missing Values
ID	ID	Integer	Patient ID	no
Diabetes_binary	Target	Binary	0 = no diabetes 1 = prediabetes or diabetes	no
HighBP	Feature	Binary	0 = no high BP 1 = high BP	no
HighChol	Feature	Binary	0 = no high cholesterol 1 = high cholesterol	no
CholCheck	Feature	Binary	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years	no
BMI	Feature	Integer	Body Mass Index	no
Smoker	Feature	Binary	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes	no
Stroke	Feature	Binary	(Ever told) you had a stroke. 0 = no 1 = yes	no
HeartDiseaseorAttack	Feature	Binary	coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	no
PhysActivity	Feature	Binary	physical activity in past 30 days - not including job 0 = no 1 = yes	no

*Above is the snippet of the 21 features that we have (showcasing 10 of 21)

Solution

Data Exploration and Pre-processing:
We will explore the health dataset by checking for missing values, summarizing statistics, and visualizing feature distributions, then pre-process it by handling categorical variables, scaling features, addressing outliers, and performing necessary imputations for missing values.

1. **Feature Engineering**:
   After exploration & pre-processing we will perform feature engineering by creating meaningful features through transformations, considering interactions between variables, and leveraging domain-specific knowledge to improve the model's predictive capabilities.

2. **Model Selection and Training**
   In this phase of the project, we will select a suitable classification model from options like Logistic Regression, Random Forest, or Gradient Boosting based on the dataset. Implement a pipeline for model training, incorporating cross-validation, and optimize performance by tuning hyperparameters through grid search or randomized search.

3. **Model Evaluation**:
   - Choose evaluation metrics that are suitable for the classification task, such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC).
   - Use techniques such as cross-validation to obtain reliable estimates of model performance and avoid overfitting.

4. **Insights**:
   - Analyze the trained model to gain insights into the most important features contributing to the prediction of diabetes.
   - Understand the underlying factors associated with the disease and provide actionable insights for healthcare professionals.

Future Considerations and Scope

In addition to the core steps mentioned above, there are several future considerations to enhance the project:

- **Hyperparameter Tuning:** Fine-tune the selected models by optimizing their hyperparameters using techniques such as grid search or Bayesian optimization. This can further improve the model's performance.
- **Ensemble Methods:** Explore ensemble methods, such as bagging or boosting, to combine multiple models and improve the overall prediction accuracy.