CS-577 - Machine Learning
Diabetes Prediction Using Machine Learning Techniques

Alexandros Angelakis csdp1363, angelakis@csd.uoc.gr

Problem Description and Motivation

Diabetes is a chronic disease that affects millions of people worldwide. Early detection is crucial to prevent severe complications. In this project, I aim to build a predictive model that can identify individuals at risk of developing diabetes based on clinical features. The motivation for this project lies in the potential to aid healthcare professionals in diagnosing diabetes more efficiently and accurately, thereby improving patient outcomes.

The Data

We will use the diabetes dataset provided, which contains relevant clinical features such as glucose levels, BMI, age, blood pressure, and other diagnostic metrics. This dataset appears to be structured for classification tasks, where the goal is to predict whether a patient is diabetic (1) or not (0). The dataset will be preprocessed to handle missing values, normalize features, and prepare it for machine learning algorithms. You can find the dataset here: dataset

Code Implementation

We will use Python libraries such as scikit-learn, pandas, numpy, and matplotlib for model development, data preprocessing, and visualization. Specific steps include:

1. **Data Cleaning and Preprocessing:** Handling missing values, feature scaling, and exploratory data analysis. I also might perform SMOTE or hybrid sampling because the dataset is highly imbalanced, (class 0: 91500, class 1: 8500).
2. **Feature Selection**: Given the low number of features in the dataset, I will apply the Best Subset Selection algorithm to identify the most relevant features for prediction. Additionally, I can also apply the causal-based feature selection method that we've been taught in the course.
3. **Model Implementation:** Implementing multiple machine learning models such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM) and Naïve Bayes Classifier.
4. **Model Evaluation:** Using metrics like accuracy, precision, recall, F1-score, and ROC-AUC etc. to evaluate performance.
5. **Hyperparameter Tuning:** Using cross-validation and GridSearch to optimize models.

Required Reading and Resources

1. **Machine Learning Techniques:** Basics of classification algorithms, model evaluation metrics, hyperparameter tuning, and feature selection.
2. **Research Papers and Tutorials:** Articles on diabetes prediction, best subset selection, SMOTE, and causality in machine learning.
3. **Documentation:** Libraries such as scikit-learn and resources on data preprocessing and visualization techniques.