**Title**: Predictive Analytics in Public Health: Evaluating Diabetes Risk Factors Using CDC's BRFSS Data

**Abstract**:

This report uses the 2015 CDC's Behavioral Risk Factor Surveillance System (BRFSS) data to identify significant predictors of diabetes, employing both baseline and advanced ensemble statistical methods. Logistic Regression and Random Forest were highlighted for their effectiveness in prediction accuracy through parameter optimization and cross-validation.

**Introduction**:

The study addresses the prevalence and economic impact of diabetes in the U.S., underscoring the importance of predictive analytics for early diagnosis and management. It elaborates on the biological process of insulin utilization and the progressive nature of diabetes.

**Methodology**:

Data from the BRFSS 2015, encompassing responses from 253,680 participants, were analyzed. Significant predictors were identified using logistic regression, with subsequent application of baseline methods (LDA, QDA, Naïve Bayes, Logistic Regression, KNN) and ensemble methods (Random Forest, XG Boosting) to evaluate diabetes risk.

**Results and Findings**:

Before and after cross-validation results were discussed, showing Random Forest's superior performance post-optimization. Significant predictors including High Blood Pressure, High Cholesterol, Heart Disease or Attack, and Difficulty Walking were consistently impactful across different models.

**Future** **Research**:

The report recommends expanding the predictor variables to include dietary and lifestyle factors and exploring further advanced machine learning techniques. It suggests that using a larger dataset could improve the robustness of the findings.

**Further** **Reading**:

References for statistical methods and additional reading on diabetes can be found in the faraway package documentation on CRAN and related articles.