**Diabetes Health Indicators Dataset**

**Context**

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. After different foods are broken down into sugars during digestion, the sugars are then released into the bloodstream. This signals the pancreas to release insulin. Insulin helps enable cells within the body to use those sugars in the bloodstream for energy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed.

Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

The scale of this problem is also important to recognize. The Centers for Disease Control and Prevention has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk. While there are different types of diabetes, type II diabetes is the most common form and its prevalence varies by age, education, income, location, race, and other social determinants of health. Much of the burden of the disease falls on those of lower socioeconomic status as well. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approaching $400 billion dollars annually.

**Content**

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

This dataset contains 3 files:

1. diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables
2. diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.
3. diabetes \_ binary \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

Explore some of the following research questions:

1. Can survey questions from the BRFSS provide accurate predictions of whether an individual has diabetes?
2. What risk factors are most predictive of diabetes risk?
3. Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?
4. Can we create a short form of questions from the BRFSS using feature selection to accurately predict if someone might have diabetes or is at high risk of diabetes?

**Acknowledgements**

It it important to reiterate that I did not create this dataset, it is just a cleaned and consolidated dataset created from the BRFSS 2015 dataset already on Kaggle. That dataset can be found [here](https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system) and the notebook I used for the data cleaning can be found [here](https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset-notebook).

**Inspiration**

Zidian Xie et al for Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques using the 2014 BRFSS was the inspiration for creating this dataset and exploring the BRFSS in general. [Link](https://www.cdc.gov/pcd/issues/2019/19_0109.htm)

**References**

[Statistics About Diabetes | ADA](https://diabetes.org/about-diabetes/statistics/about-diabetes)

[National Diabetes Statistics Report | Diabetes | CDC](https://www.cdc.gov/diabetes/data/statistics-report/index.html#print)

**Classification Use Case: Health Indicators for Diabetes**

**Objective:**

Develop a machine learning model to classify individuals into three categories based on their diabetes status: no diabetes (or only during pregnancy), prediabetes, and diabetes.

*Steps:*

**1. Data Preprocessing:**

- Handle missing values through imputation or elimination.

- Normalize or standardize the feature variables to ensure they're on the same scale.

- Address class imbalance using techniques like SMOTE or undersampling to improve model performance.

**2. Feature Selection:**

- Use techniques like Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), or correlation matrices to identify and select the most predictive features.

**3. Model Selection:**

- Experiment with different algorithms such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) for classification.

- Consider using ensemble methods or deep learning approaches if computational resources allow.

**4. Model Training and Validation:**

- Split the dataset into training and validation sets to evaluate model performance.

- Utilize cross-validation techniques like k-fold cross-validation to ensure the model's generalizability.

**5. Performance Evaluation:**

- Evaluate model performance using metrics appropriate for imbalanced datasets, such as Precision, Recall, F1-Score, and the ROC-AUC score.

- Perform error analysis to identify where the model is making mistakes and why.

**6. Model Interpretation:**

- Use techniques like SHAP or LIME to interpret the model's predictions and understand the importance of different features.

**Prediction Use Case: Type II Diabetes Prediction Based on Location and Demographics**

**Objective:**

Develop a predictive model to assess the risk of type II diabetes in the populations of Florida and New York, utilizing demographic and location-based data alongside health indicators.

**Data Acquisition for Prediction Model:**

To accomplish the prediction objective, additional data on type II diabetes prevalence by location and demographics in the US is necessary. Potential sources for this data include:

- Centers for Disease Control and Prevention (CDC): The CDC's Diabetes Atlas and other datasets offer state-specific diabetes statistics that can be valuable for this analysis.

- U.S. Census Bureau: Demographic data by state, including age, income, education level, and race/ethnicity, which are critical social determinants of health impacting diabetes risk.

- Behavioral Risk Factor Surveillance System (BRFSS): Additional years of BRFSS data may provide insights into trends and changes in diabetes prevalence and risk factors over time.

*Steps for Prediction:*

1. \*\*Data Integration\*\*:

- Combine the health indicators dataset with demographic and location-based data for Florida and New York.

- Ensure proper alignment of features across datasets and handle any discrepancies in data collection methods or definitions.

2. \*\*Predictive Modeling\*\*:

- Given the nature of the prediction task, regression models (for continuous outcomes) or classification models (for categorical outcomes, e.g., high vs. low risk) can be considered.

- Explore linear regression, logistic regression, and more complex models like Random Forest or Gradient Boosting for prediction.

3. \*\*Evaluation and Deployment\*\*:

- Evaluate the model's predictive performance using appropriate metrics, such as Mean Absolute Error (MAE) or Area Under the ROC Curve (AUC) for classification tasks.

- Deploy the model in a suitable format for stakeholders, possibly integrating it into health care systems for real-time risk assessment based on the input variables.

This ML project, with a focused timeline of 3 weeks, requires efficient project management and prioritization of tasks to meet the dual objectives of classification and prediction. Collaboration with domain experts in healthcare and public health will be crucial to ensure the models developed are not only technically sound but also practically relevant and ethically considerate in their application.