Diabetes Prediction & Risk Factor Analysis

Team Members: Taylor Erickson

Description

The goal of this project is to identify the strongest predictors of diabetes as a health outcome.

Additionally, the project will investigate whether or not there is a meaningful difference between the strongest predictors of pre-diabetes and the strongest predictors of diabetes.

Prior Work

There are many examples of prior work on this subject, which are accessible through Kaggle (dataset location) and medical journals:

Kaggle Examples from this dataset:

https://www.kaggle.com/code/tumpanjawat/diabetes-eda-cluster-catboost

https://www.kaggle.com/code/anastasiyaigonina/diabetes-eda-hypothesis-testing-predictions

https://www.kaggle.com/code/frengzkermova/diabetes-prediction

General Medical Reviews & Studies:

https://diabetesjournals.org/care/article/21/Supplement_3/C3/18526/Epidemiology-of-Type-2-Diabetes-Risk-Factors

https://europepmc.org/article/med/21163426

https://journals.sagepub.com/doi/pdf/10.1177/2040622314548679

Datasets

Dataset: Diabetes Health Indicators from the BRFSS Survey

Dataset files:

1. Raw survey results for diabetes, pre-diabetes, and no diabetes classes (*diabetes_012_health_indicators_BRFSS2015.csv*)
2. Balanced subset of total survey data with even split between diabetes and no-diabetes (*diabetes_binary_5050split_health_indicators_BRFSS2015.csv*)
3. Transformed diabetes classes into binary, diabetes vs. no diabetes (*diabetes_binary_health_indicators_BRFSS2015.csv*)

Link: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Description: The dataset is comprised of a subset of the Behavioral Risk Factor Surveillance System (BRFSS) dataset, a survey conducted annually by the CDC, with features and questions related to diabetes and diabetes risk factors. There are three files - the first is just the raw output of the survey results. The second file is a transformed subset, where there is an even distribution of diabetes and no diabetes. The third file is a transformed version of the first file, where pre-diabetes and diabetes are combined into one category, for a diabetes vs. no diabetes labelling.

Location: It is downloaded locally on my machine, and is accessible via the link above.

Proposed Work

Data Cleaning:

- For each attribute, perform inspection of any null or missing values - if an attribute has a significant number of missing values, drop it from the analysis.
- For each entry, identify whether there are any null or missing values - if an entry has a significant number of missing values, drop it from the analysis.

Attribute Selection/Pre-Processing:

- For each attribute, perform exploratory analysis to understand the variance and distribution of each attribute in the dataset. Drop any attribute that does not have high variance.
- Leverage ANOVA analysis and other statistical methods to identify the attributes that contribute to the largest amount of variance in our outcome feature, eliminating those that are unlikely to be meaningfully related.

Data Modelling:

- For each outcome of interest (diabetes vs. non-diabetes, diabetes vs. pre-diabetes vs. non-diabetes), select the strongest potential predictors and perform simple linear and/or logistic regression modelling to identify key predictors and their interactions. Leverage non-parametric methods to create a best-fit model if parametric approaches are insufficient.

List of Tool(s)

Language: Python

Python libraries:

Data ingestion and manipulation: Pandas

Visualization: Seaborn, Matplotlib

Statistical analysis and modelling: Scipy, Sklearn, Statsmodels

Evaluation

To evaluate the efficacy of the analysis and predictive models:

- Test the efficacy of generated model(s) against a test sample

- Compare to similar work and findings, and see whether the results are applicable

- Publish findings and source peer review and expert feedback