Description of preliminary data processing

* Drop missing values since most machine learning algorithms cannot work with missing features
* Handle categorical attribute

Description of data features selection

* Features: variables in demographics, diabetic-related health status, general health status, insurance, and income level.
* These features are chosen because of their potential correlation with our target: “Taking diabetic pills”.

Description of how data is split into training and test set

* Data is split into training and testing sets by using Scikit-learn’s train\_test\_split function.
* We end up with four sets: X is split into X\_train and X\_test sets, and y is split into y\_train and y\_test sets.

Explanation of model choice

* We choose logistic regression to predict factor(s) that are correlated with diabetes.
* All variables (except for weight and age) are categorical variables.
* The target variable only has two possible values, whether a person is taking diabetic pills. When this classification model encounters new data, it predict whether the individual would have diabetes.
* Limitation: the accuracy from a single variable is low, suggesting that multiple factors are in play.
* Benefit: logistic regression is relatively simple to execute and understand.

Exploratory analysis

https://public.tableau.com/app/profile/amy8818/viz/finalproject\_16711568635250/Story1?publish=yes