Insights into Diabetes Prevalence in the US

Team: Leyda Almodóvar, Neal Edgren, Chiara Mattamira, Shravan Patankar

Github: https://github.com/nedgren/Health-Insights

Dataset:
https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation

Background and Project Overview:

We developed a comprehensive analysis for the risk of diabetes at a county-level using a rich dataset of socioeconomic, demographic, and health indicators. We obtained our dataset from County Health Rankings & Roadmaps. Our target variable is the percentage of adults with diabetes. Our predictive features are 48 divided into five major categories: demographic, socioeconomic, health habits, environmental, and access to care.

Our Objectives:

- Identify key risk predictors for diabetes in the US
- Provide insights to help make informed policy decisions
- Understand which populations are at risk at a local level

Data cleaning: We removed

- rows with aggregated state data
- counties with more than 35 null values
- categorical data
- features with more than 100 missing values
- features that are redundant or obviously correlated

We imputed using k-nearest neighbors with k = 10.

Stakeholders: Federal, State, and Local Health Organizations, Insurance Companies, Social Service Agencies, Pharmaceutical Corporations.

Modeling approach:

We split the data into 80% training set and 20% testing set. 20% of the training set was reserved as a validation set. After doing some exploratory analysis, we ran some baseline models (mean model, random sampling) on all the features and then we compared linear regression, random forest, and XGBoost. Since XGBoost performed significantly better, we performed cross-validation for this model and tuned

hyperparameters.
We then looked at several models using XGBoost: Full model (includes all 48 features), Health behaviors only, Socioeconomic only, Demographic only, Physical Environment only, and Access to Care only.

Key Performance Indicators: Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error

Results:

For the full model, we found that the percentage of physically inactive people was the most important feature, followed by percentage of children in poverty, and percentage of people who completed high school in a given county. The RMSE for the full model is 0.31.

We also looked at importance by feature group, calculated the RMSE for each, and found that the top features by group are as follows:

- ➢ Health behaviors:
  - ○ Top features: % physically inactive, % insufficient sleep, % adults with obesity.
  - ○ RMSE: 0.52
- ➢ Socioeconomic factors:
  - ○ Top features: % children in poverty, % completed high school, % children in single-parent households
  - ○ RMSE: 0.72
- ➢ Demographics:
  - ○ Top features: % hispanic non-white, % black, % asian
  - ○ RMSE: 1.02
- ➢ Physical environment:
  - ○ Top features: % food insecure, % households with broadband access, food environment index
  - ○ RMSE: 1.34
- ➢ Access to care:
  - ○ Top features:% uninsured, % with annual mammograms, % uninsured children
  - ○ RMSE: 1.74

We compared the prevalence of diabetes between counties with low median income and counties with high median income. We found that demographic features are more predictive for low-income counties, while health behavior features are more predictive for high-income counties. There were also noticeable differences in socioeconomic features.

We also split the data by percentage of non-White and found that for counties below the median income, socioeconomic factors were more important and for counties above the median, health behaviors were more prevalent highlighting the need for tailored interventions that address both health behaviors and socio-economic challenges specific to each racial and income group.

We concluded that diabetes prevalence is not only predicted by health behaviors, as it can be expected but also by socioeconomic factors, with the latter ones being especially important in counties with lower household income.

Future directions:

- Focus on a specific state/geographical region
- Make a more inferential model
- Further reduce features list