Predicting Diabetes Rate Using the Food Environment Atlas Project Executive Summary

Team members: Mercy Amankwah, Danielle Brager, Nicole Bruce, Monalisa Dutta, Cyril Dennis Enyi
Github: https://github.com/db4014/food-atlas-2024

Overview:
Our objective is to develop predictive regression and classification models to anticipate the diabetes rate in different regions of the United States, and to identify key contributing factors. We use data science tools to predict the diabetes rate given various food environment indicators such as poverty rate, access to snap benefits, food stores, restaurants, and other community characteristics regarding food, health, nutrition, physical activities, and socioeconomic status. We utilize the data from the Food Environment Atlas (https://www.ers.usda.gov/data-products/food-environment-atlas/) to predict the diabetes rate using all these indicators. This project holds significant value for health agencies and local governments, enabling them to allocate resources more efficiently and advocate for improved access to healthy food.

Stakeholders:
- Residents, health agencies, and local government would want to lobby for better access to food by using the diabetes rate as a factor

- Grocery store chains, farmer's markets, etc. would use the diabetes rate as an indicator of where there is additional demand for more stores

KPI (Key Performance Indicator):
- Ability to find features that relate to diabetes rate
- Ability to predict diabetes rate based on a combination of features
- Build effective predictive models for diabetes rates in US counties using the USDA's Food Environment Atlas. The effectiveness of the model is measured by the model's accuracy, precision, recall scores for classification models, and mean squared error for regression models.
- Utilize our model to inform a county that they are at risk of increasing their diabetes rate and the associated causative factors as identified by our model, allowing communities to take precautionary measures.

Approach (Data Exploration):
- Our primary resource for the dataset is the Food Environment Atlas dataset found on the USDA website (https://www.ers.usda.gov/data-products/food-environment-atlas), which provides diabetes prevalence data for each county in 2008 and 2013. The dataset comprises 284 variables, including diabetes rates for 2008 and 2013.
- Since it is hard to work with so many variables, and not all the variables are relevant to the study, we first computed the correlation matrix between all these variables and dropped the features that are highly correlated with each other.

- We defined four regions northeast, southeast, midwest, and west, then separated the data into these regions. These classified regions were introduced into the dataset as one of our features for modeling.
- For the classification model, we defined a new feature named health which is either 0 or 1 or 2 depending on whether the diabetes rate is low, medium, or high respectively (we determined the threshold using the violinplot).

Models:

In our modeling approach, we:

- explore the dataset using scatter plots to visualize relationships between variables.
- compute correlations between features to identify significant relationships
- use the correlation matrix and the pairplot reduces the dimensionality of our model by selecting only the most relevant features for predicting diabetes rates
- compare the performance of various regression models (Linear regression, Support vector regressor, KNeighbors regressor, Random forest regressor, AdaBoost regressor, Gradient boosting regressor, XGBoost regressor, Histogram Gradient Boosting Regression Tree) using mean squared error as our evaluation metric
- compare the performance of various classification models (Linear Discriminant Analysis, Logistic Regression, Support Vector Classifier, Quadratic Discriminant Analysis, KNeighbors Classifier) using accuracy, precision, and recall scores as our evaluation metric

Results & Strategies:

- Tree-based ensemble methods (Random Forest, Gradient Boosting, XGBoost, and Histogram Gradient Boosting Regression Tree) significantly outperformed all the other models. The best regression model was the Gradient boost.
- After doing a gridsearch, we built a Gradient boost model that gave a mean squared error of 0.73 for the 2008 data and 1.28 for the 2013 data.
- In the classification models, the Random Forest classifier performed the best with an accuracy score of 0.81 and a precision of 84% .
- The most important predictive features are by far the percentage of SNAP participants, SNAP benefits per capita in a county, and the region the county belongs to.

Future Iterations:

Two straightforward improvements would be to collect and use more years of diabetes data and try integrating time-series data to capture trends and seasonal variations in food availability, lifestyle changes, and their impact on diabetes rates. Also, one could get the latitude and longitude of the counties and use that as one of the features for building the models. We anticipate that it will be one of the most important features. This way we can provide more localized predictions. We can also conduct longitudinal studies to observe how changes in the food environment impact diabetes rates over time. One could also include data on nutritional quality and diversity of available foods, beyond ordinary availability or consumption.

Acknowledgements:
We would like to thank Alexis Johnson for her enthusiastic mentorship and helpful feedback during this project. We are also very grateful to Steven Gubkin for the amazing lectures during the boot camp. We also want to thank Roman Holowinsky, Alec Clott, and The Erdős Institute for their support during this boot camp.