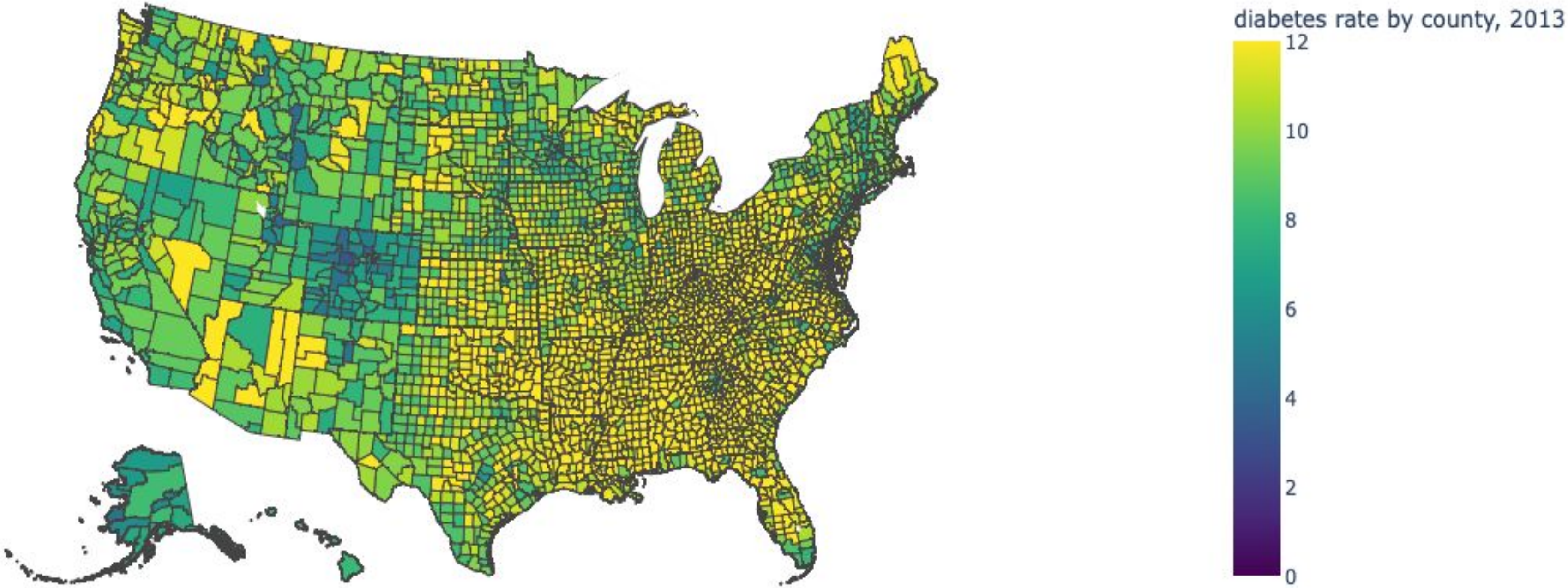
Predicting Diabetes Rate Using the Food Environment Atlas

Mercy Amankwah, Danielle Brager, Nicole Bruce, Monalisa Dutta, and Cyril Dennis Enyi

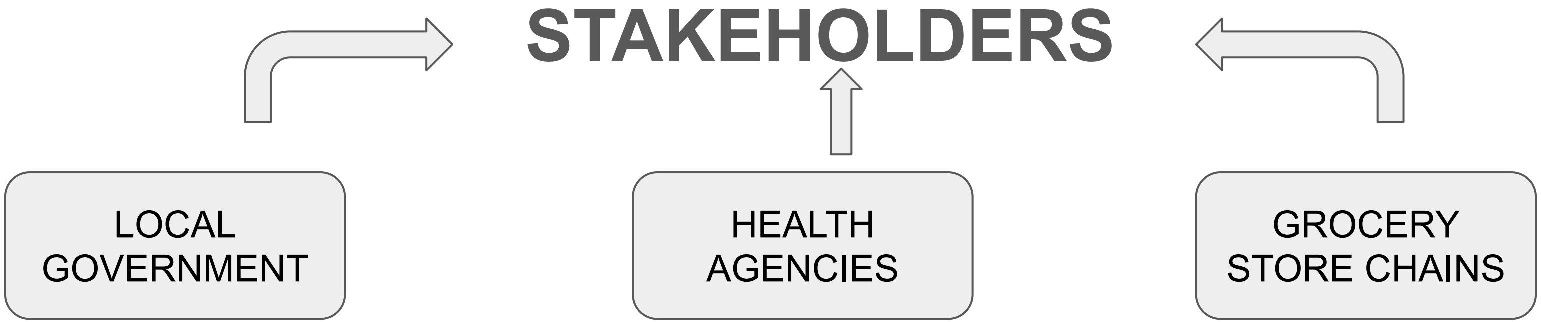
Introduction

Diabetes is a major global health issue, affecting millions and incurring significant costs



Dataset

- We utilize the Food Environment Atlas (FEA) dataset to predict diabetes rate given various food environment indicators such as poverty rate, access to SNAP benefits, access to food stores, and other community characteristics
- The FEA dataset is compiled by the USDA’s Economic Research Service



STAKEHOLDERS

LOCAL GOVERNMENT

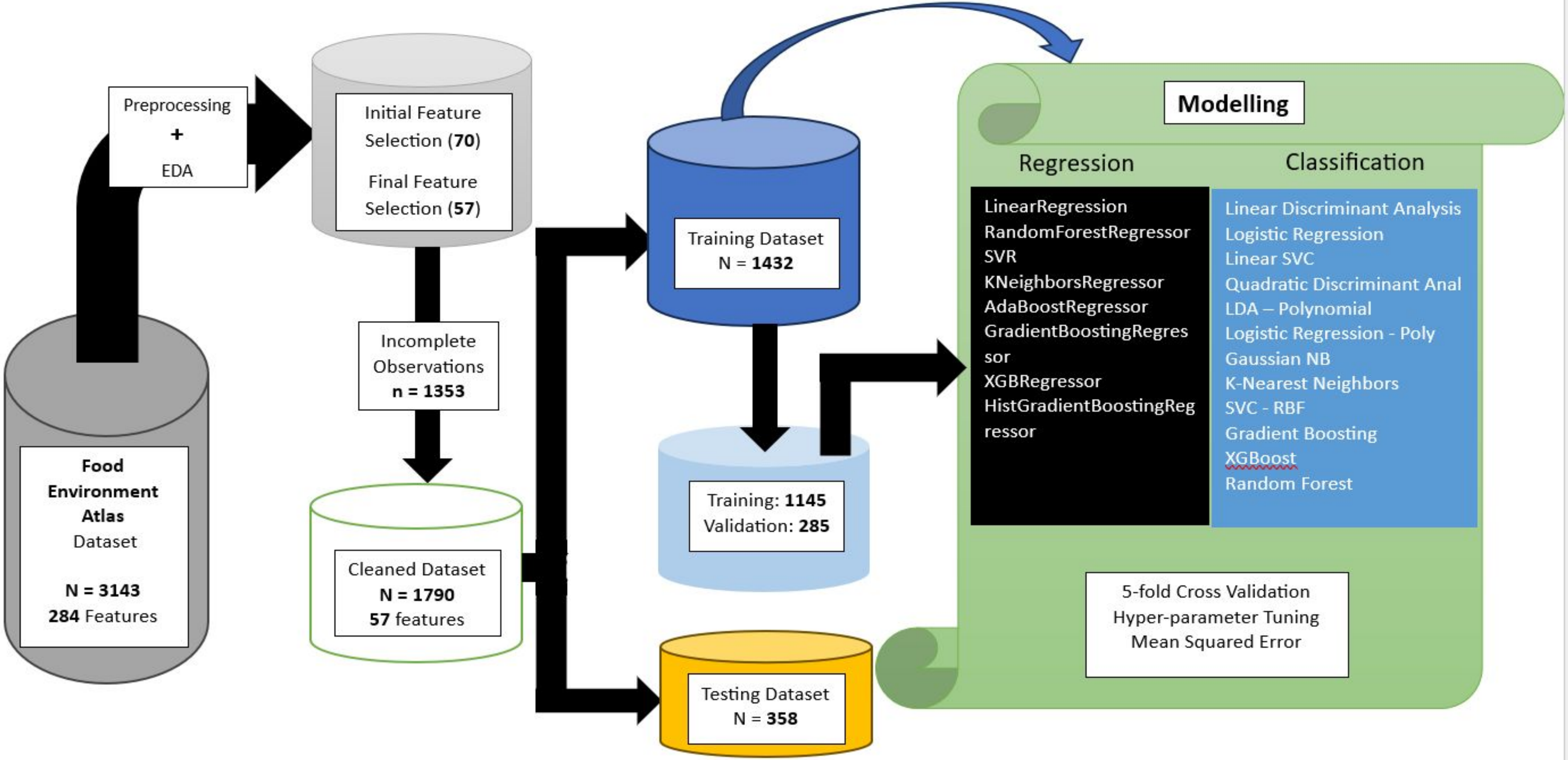
HEALTH AGENCIES

GROCERY STORE CHAINS

KPIs

- Develop a model to predict diabetes incidence in counties within the United States.
- Identify additional factors affecting diabetes rate aside poverty (the obvious).
- Use model predictions to educate counties to anticipate changes in diabetes prevalence and propose potential preventive measures.

Model Design



Food Environment Atlas Dataset

N = 3143

284 Features

Preprocessing
+
EDA

Initial Feature Selection (**70**)

Final Feature Selection (**57**)

Incomplete Observations
n = 1353

Cleaned Dataset
N = 1790
57 features

Training Dataset
N = **1432**

Training: **1145**
Validation: **285**

Testing Dataset
N = **358**

Modelling

Regression

- LinearRegression
- RandomForestRegressor
- SVR
- KNeighborsRegressor
- AdaBoostRegressor
- GradientBoostingRegressor
- XGBRegressor
- HistGradientBoostingRegressor

Classification

- Linear Discriminant Analysis
- Logistic Regression
- Linear SVC
- Quadratic Discriminant Anal
- LDA – Polynomial
- Logistic Regression - Poly
- Gaussian NB
- K-Nearest Neighbors
- SVC - RBF
- Gradient Boosting
- XGBoost
- Random Forest

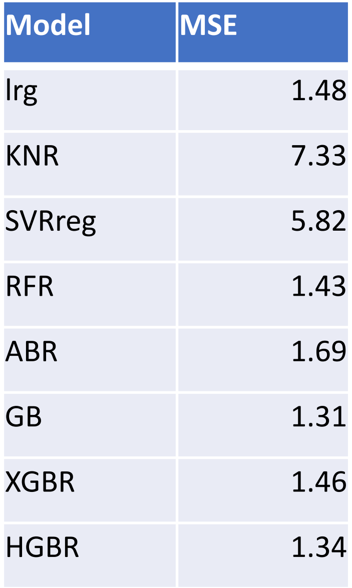
5-fold Cross Validation
Hyper-parameter Tuning
Mean Squared Error

Results: Regression Models

- Random Forest, Gradient boost, XGBoost, and Histogram Gradient Boosting Regression Tree significantly outperformed all the other models. The best regression model was the Gradient boost.

- Gradient boost model gave a mean squared error of 1.28.

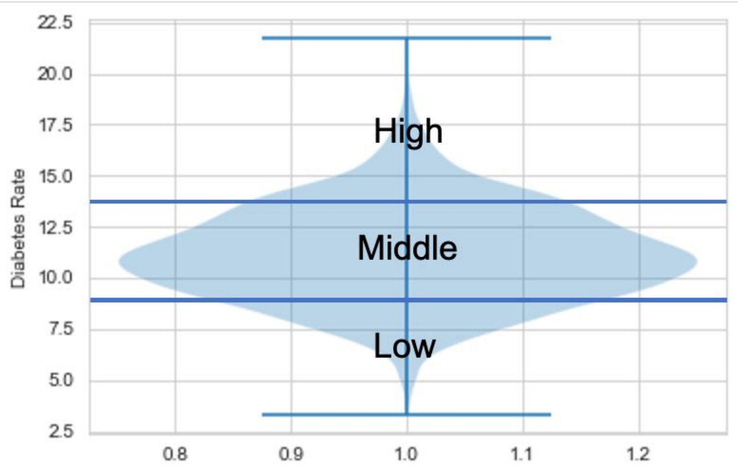
- The most important predictive features are the percentage of SNAP participants, SNAP benefits per capita in a county and the region the county belongs to.

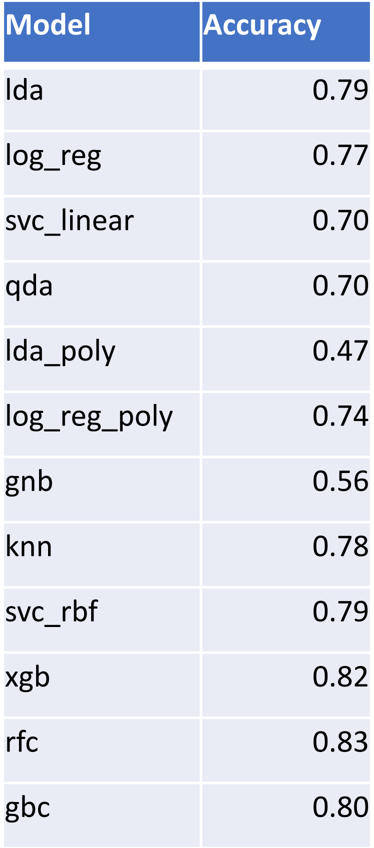


Model	MSE
lrg	1.48
KNR	7.33
SVRreg	5.82
RFR	1.43
ABR	1.69
GB	1.31
XGBR	1.46
HGBR	1.34

Results: Classification Models

- Gradient boosting, random forest, and XGB classifiers outperformed other models
- Random Forest classifier performed the best with an accuracy score of 0.81 and a precision of 84%





Model	Accuracy
lda	0.79
log_reg	0.77
svc_linear	0.70
qda	0.70
lda_poly	0.47
log_reg_poly	0.74
gnb	0.56
knn	0.78
svc_rbf	0.79
xgb	0.82
rfc	0.83
gbc	0.80

Future work

- Future work involves reducing the error and improving the accuracy associated with predicting a county’s diabetes rate from food access data
- Collect and use more years of diabetes data and try integrating time-series
- Use latitude and longitude of the counties as one of the features for building the models to provide more localized predictions

THANK YOU