12.19.24

Predicting Health Outcomes Based on Access to Food

Neha Mathew

Lorem Ipsum

1. Research Question
2. Data Acquisition
3. Data Wrangling
4. Exploratory Data Analysis
5. Machine Learning Model
6. Conclusion

1. **Research Question**
2. Data Acquisition
3. Data Wrangling
4. Exploratory Data Analysis
5. Machine Learning Model
6. Conclusion

Does limited access to healthy food options correlate with higher diabetes prevalence?

1. Research Question
2. **Data Acquisition**
3. Data Wrangling
4. Exploratory Data Analysis
5. Machine Learning Model
6. Conclusion

USDA Food Access Data

From the United States Department of Agriculture's Economic Research Service, the dataset contains information about US county's ability to access supermarkets, supercenters, grocery stores, or other sources of healthy and affordable food. This dataset shows the population of people outside various radii of grocery stores, broken down by county. It also takes into account the within those populations are low income or do not have a vehicle.

https://corgis-edu.github.io/corgis/csv/food_access/

Key	List of...	Comment	Example Value
County	String	County name	"Autauga County"
Population	Integer	Population count from 2010 census	54571
State	String	State name	"Alabama"
Housing Data.Residing in Group Quarters	Float	Count of tract population residing in group quarters	455.0
Housing Data.Total Housing Units	Integer	Occupied housing unit count from 2010 census	20221
Vehicle Access.1 Mile	Float	Housing units without vehicle count beyond 1 mile from supermarket	834.0
Vehicle Access.1/2 Mile	Float	Housing units without vehicle count beyond 1/2 mile from supermarket	1045.0
Vehicle Access.10 Miles	Float	Housing units without vehicle count beyond 10 miles from supermarket	222.0
Vehicle Access.20 Miles	Float	Housing units without vehicle count beyond 20 miles from supermarket	0.0

CDC Diagnosed Diabetes

From the CDC, a breakdown of the rates of diagnosed diabetes by county. Includes the Social Vulnerability Index, which refers to the demographic and socioeconomic factors (such as poverty, lack of access to transportation, and crowded housing) that adversely affect communities that encounter hazards and other community-level stressors.

https://gis.cdc.gov/grasp/diabetes/diabetesatlas-sdoh.html

County FIPS	County	State	Diagnosed Diabetes-2021 (Percentage)	Overall SVI-2018 (Percentile)
01001	Autauga County	Alabama	8	0.4354
01003	Baldwin County	Alabama	9.9	0.2162
01005	Barbour County	Alabama	9.2	0.9959
01007	Bibb County	Alabama	9.1	0.6003
01009	Blount County	Alabama	8.5	0.4242
01011	Bullock County	Alabama	9	0.8898
01013	Butler County	Alabama	9.8	0.8653
01015	Calhoun County	Alabama	12	0.8252
01017	Chambers County	Alabama	10.7	0.7382

1. Research Question
2. Data Acquisition
3. **Data Wrangling**
4. Exploratory Data Analysis
5. Machine Learning Model
6. Conclusion

- Loading and Cleaning Data
  - Load and clean datasets (remove nulls, handle missing data).
  - Merge datasets using “County” as the key.
  - Drop unnecessary columns.
  - Check and fix non-numeric data to prepare for machine learning.
  - Checked for missing values and duplicates.
- Standardize features: Use StandardScaler for numerical scaling.
- Ensured no non-numeric data for machine learning.

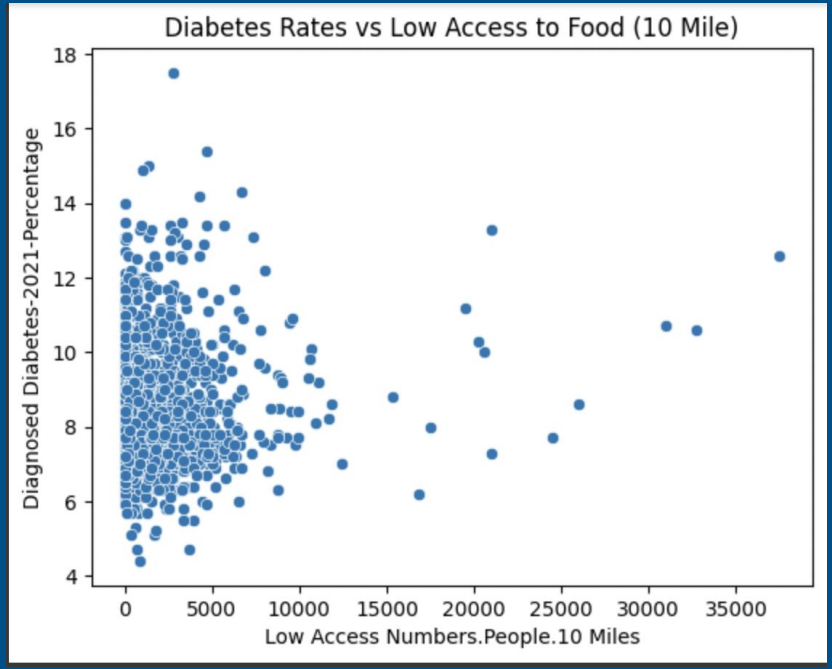
1. Research Question
2. Data Acquisition
3. Data Wrangling
4. **Exploratory Data Analysis**
5. Machine Learning Model
6. Conclusion

Scatter Plot

The majority of the data points are concentrated on the left side (low access numbers below ~10,000).

There is no clear upward or downward trend, suggesting a weak or no direct correlation between diabetes rates and low access to food within a 10-mile radius.

Some outliers (high diabetes rates) exist at higher access numbers, but they are sparse.

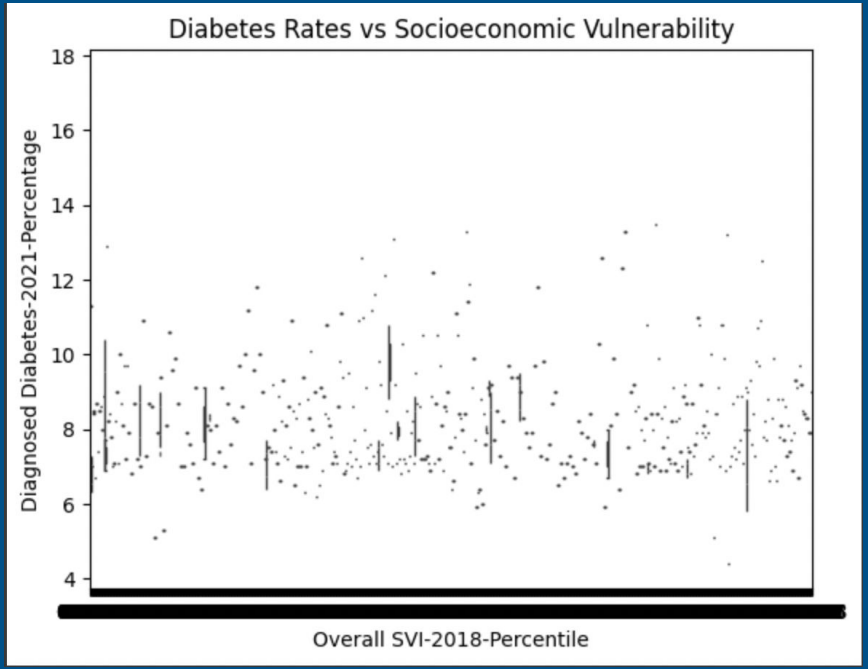


Box Plot

There is a spread of diabetes rates across all levels of socioeconomic vulnerability.

While the points appear noisy, there are slight groupings where higher diabetes rates occur at higher SVI percentiles.

The plot might indicate a weak positive relationship: areas with higher socioeconomic vulnerability tend to have slightly higher diabetes rates.



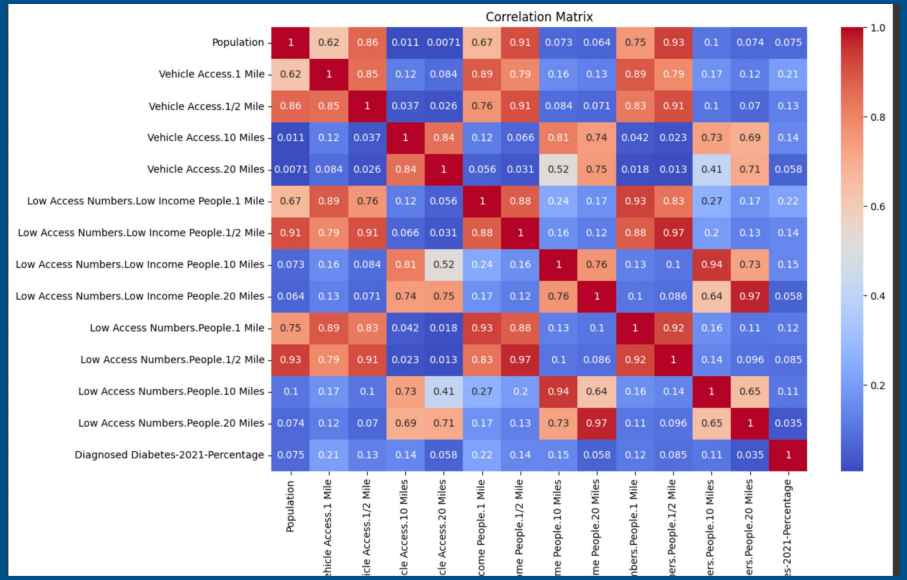
Correlation Matrix

Weak correlations exist between diabetes rates and all other variables. The highest observed correlation is ~0.22 with "Vehicle Access.1 Mile," but this is still weak.

Low Access Numbers (10 Mile, 20 Mile) show minimal correlation (~0.10 and ~0.11) with diabetes rates.

Other low access measures and socioeconomic factors also exhibit weak relationships.

Overall, it looks like there are not any linear relationships between access to food and diabetes outcomes

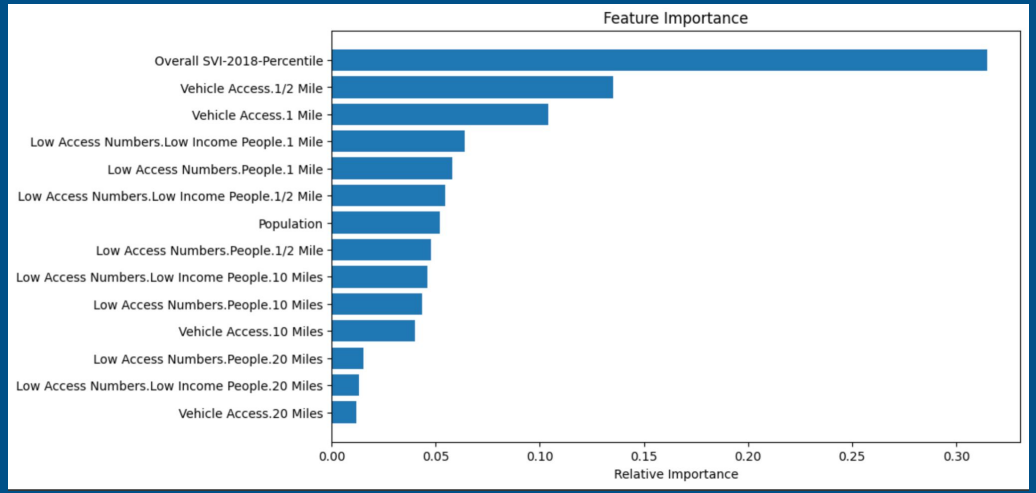


1. Research Question
2. Data Acquisition
3. Data Wrangling
4. Exploratory Data Analysis
5. **Machine Learning Model**
6. Conclusion

Model Performance Summary

- Best Model:
  - Random Forest 2 (scaled with poly fit) achieved the best performance with RMSE: 0.4734 and R^2: 0.9003, explaining 90% of the variance in diabetes rates.
- Underperforming Models:
  - Gradient Boosting models (including optimized versions) were less effective, with R^2 scores ranging between 0.46 and 0.55, indicating that these models captured less variance.
  - Cross-validated Random Forest struggled, showing signs of overfitting or inconsistencies across data splits.

Model	RMSE	R^2 Score
Random Forest 1	1.2942	0.3008
Random Forest 2	0.4734	0.9003
Gradient Boosting	1.0013	0.5540
Optimized Random Forest	0.8502	0.6785
Optimized Gradient Boosting	1.0960	0.4657
Cross-Validated Random Forest	1.2783	0.2682



1. Research Question
2. Data Acquisition
3. Data Wrangling
4. Exploratory Data Analysis
5. Machine Learning Model
6. **Conclusion**

Final Thoughts

Non-Linearity

The polynomial feature transformation for the second random forest model created a very effective model to explain this data. If I had more time I would probably try a few more degrees of polynomial to see if another model would be better. But this also runs the risk of overfitting.

Feature Importance

Based on the feature importance graph, SVI was the best predictor for diabetes rates. This indicates that income and distance from food sources play a part in predicting diabetes rates but the other factors that SVI is calculated with create a better predictor.

Fast Food Dataset

My original project was tracking concentration of fast food restaurants in the US in comparison to food access as well as diabetes outcomes. I did not have enough time to try to incorporate this third dataset. In future I would love to figure out a way to add this to the model to paint a more detailed picture.