# Food Desert Prediction

Emmy Son

# Introduction

Food deserts—areas with limited access to affordable, nutritious food—exacerbate health and economic inequalities across communities. While traditional studies have focused on mapping existing food deserts, our project aims to **predict future food deserts** using machine learning, providing actionable insights for policymakers and community organizations.

By combining **geospatial clustering** and predictive modeling, we developed tools to identify high-risk areas, visualize patterns of food access, and understand the systemic factors driving food insecurity. This approach allows for data-driven interventions to improve food accessibility in vulnerable regions.

# Methods

## Research Objectives

1. **Identify and predict food deserts:** Use K-Means clustering and Random Forest classifiers to uncover patterns in at-risk communities.
2. **Develop a visualization dashboard:** Build an interactive tool to display food desert risks across demographics, counties, and states.
3. **Hypothesis:** Machine learning models, informed by socio-economic and demographic factors, can predict emerging food deserts and guide early interventions.

### Preliminary Findings

* **Historical context:** Practices such as redlining continue to influence food access disparities today.
* **Data sources:** USDA Food Access Research Atlas (2019) including census, income, SNAP, and supermarket data at the tract level.
* **Tech stack:** Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn), Geopandas, Deepnote, Tableau.

Our EDA explored relationships between socio-economic variables and food desert status:

**Key observations:**

* Food desert tracts typically have higher poverty rates and lower median family income.
* Race, income, and transportation access strongly correlate with food insecurity.
* Visual comparisons showed systematic disparities in resources between food desert and non-food desert tracts.

We began by cleaning the dataset, filtering out irrelevant data and imputing missing values to ensure completeness. Next, we performed exploratory analysis to identify the most informative features, using correlation matrices to highlight key variables such as income, transportation access, and racial demographics. To enhance the predictive power of our models, we derived additional features, including the distance to the nearest supermarket and cumulative transportation access metrics.

For modeling, we experimented with two algorithms: K-Nearest Neighbors (KNN) and Random Forest. KNN was chosen for its ability to evaluate local similarity patterns, while Random Forest was selected for its capability to capture more complex interactions between variables. Ultimately, Random Forest outperformed KNN, effectively modeling how factors like income, race, and transportation intersect to predict the presence of food deserts.

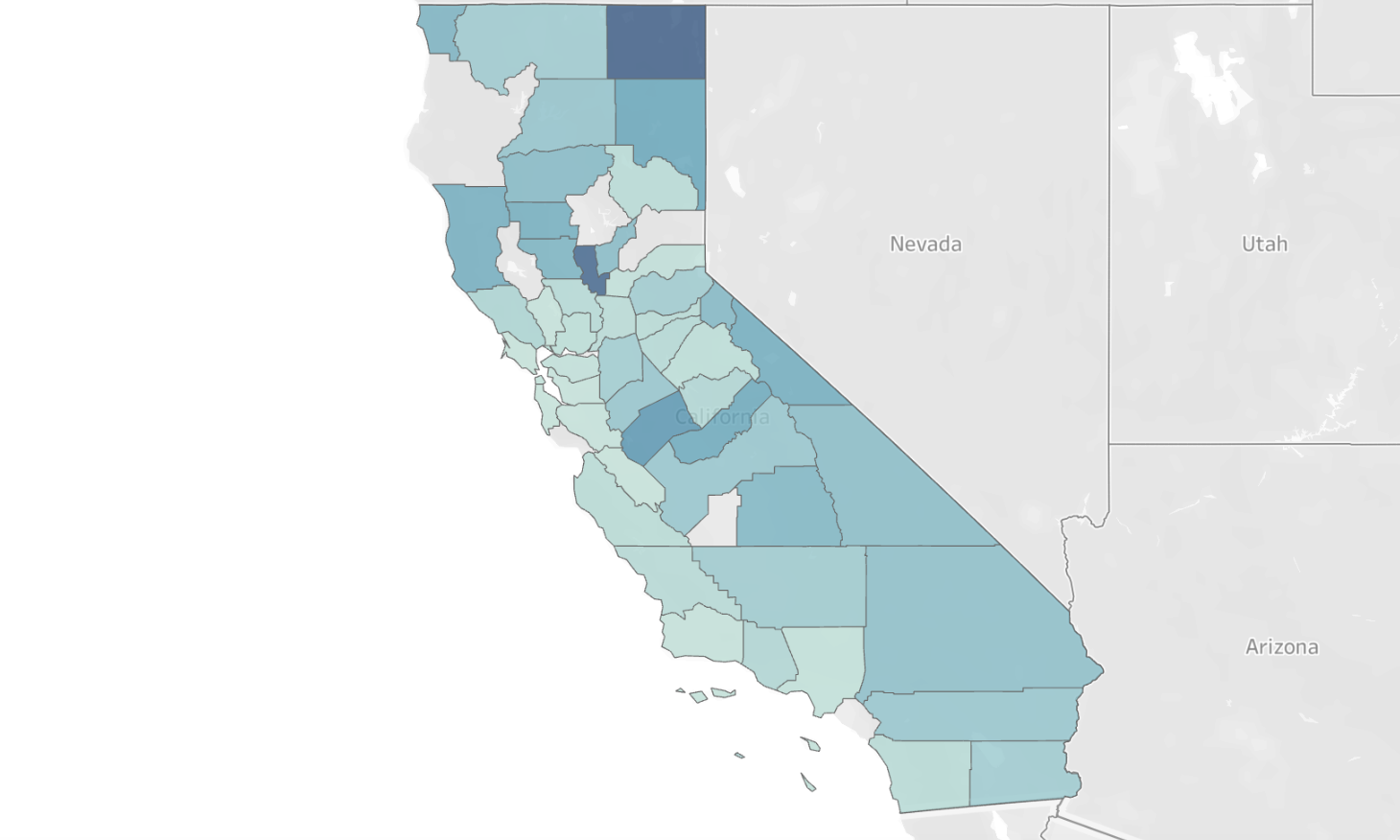
# Results

Our models were evaluated based on accuracy, precision, recall, and F1-score to determine their effectiveness in predicting food deserts. The Random Forest classifier outperformed K-Nearest Neighbors across all metrics, demonstrating its ability to capture complex interactions between socioeconomic and geographic factors. Specifically, Random Forest achieved an overall accuracy of 97%, with strong precision (0.97), recall (0.85), and F1-score (0.97) when weighted by support. In contrast, KNN showed lower performance, particularly in recall, reflecting its limitations in capturing nuanced patterns in the dataset. These results highlight the importance of modeling interactions between income, racial demographics, and transportation access to accurately identify at-risk regions.

| **Model** | **Accuracy** | **Precision (Macro Avg.)** | **Recall (Macro Avg.)** | **F1-score (Macro Avg.)** | **Support** |
| --- | --- | --- | --- | --- | --- |
| K-Nearest Neighbors | 0.83 | 0.72 | 0.77 | 0.74 | 1605 |
| Random Forest | 0.97 | 0.97 | 0.85 | 0.97 | 1605 |

The Random Forest model’s superior performance demonstrates that food deserts are driven by the intersection of multiple factors, including poverty, racial disparities, and access to transportation and supermarkets. Visualizations from our Tableau dashboard further confirm that regions with higher predicted probabilities of being food deserts consistently correspond to areas with these combined vulnerabilities, providing actionable insights for policymakers and stakeholders aiming to target interventions effectively.

**Tableau Visualization:**



# Discussion/Reflection

Our analysis highlights that food deserts are the result of multiple intersecting systemic factors rather than a single variable. Poverty, racial disparities, and limited access to transportation and healthy food collectively contribute to the emergence of these regions. The Random Forest model performed exceptionally well, achieving high accuracy while capturing complex feature interactions, making it a useful tool for policymakers to identify at-risk areas. Exploratory data analysis and visualizations reinforced these findings, showing that disadvantaged communities face compounding barriers to food access.

Looking ahead, integrating community survey data could provide local validation and improve model accuracy. Expanding the analysis to additional regions would help assess the generalizability of our approach. Finally, collaborating with local organizations could translate these insights into actionable interventions, targeting areas with the highest risk and maximizing the impact of policy measures.