Course: DATA-245 Machine Learning

PROJECT PROPOSAL

The Food-Desert Effect: Nutrition Inequality and Metabolic Health

Professor Name: Vishnu S. Pendyala, Ph.D.

Group3
Savitha Vijayarangan
Rishi Visweswar Boppana
Kapil Reddy Sanikommu
Jane Heng

PROJECT PROPOSAL

The Food-Desert Effect: Nutrition Inequality and Metabolic Health

1. Abstract

Communities across the world experience stark nutritional inequalities due to limited access to affordable, healthy food options - commonly known as "food deserts." This project investigates how food access, socioeconomic factors, and education levels collectively influence community-level obesity and diabetes prevalence. Using USDA Food Access Research Atlas, CDC Diabetes and Obesity Data, and Census socio-economic indicators, I will perform regression and clustering to reveal relationships between food accessibility and metabolic health outcomes. Principal Component Analysis (PCA) will be used for dimensionality reduction and to visualize socio-nutritional clusters. The goal is to derive a Metabolic Vulnerability Score (MVS) - a predictive indicator of community-level metabolic health risk due to nutritional inequality.

2. Motivation

While nutrition discussions often focus on individual dietary choices, the underlying geography of access plays a critical, yet overlooked, role in shaping health outcomes. Communities without nearby grocery stores or transportation access are often forced into calorie-dense, nutrient-poor diets, leading to higher rates of obesity, diabetes, and cardiovascular diseases. The motivation for this project lies in bridging the data gap between geography, inequality, and public health. Understanding this relationship could inform local policymakers and public health advocates, guiding interventions that promote equitable access to nutrition.

3. Literature Survey

Previous research has shown strong spatial correlations between food deserts and obesity prevalence:

- Walker et al. (2010) found that low-income neighborhoods have 30–40% fewer supermarkets than high-income ones.
- Morland & Filomena (2007) identified that predominantly Black neighborhoods have higher fast-food density relative to grocery access.
- CDC (2023) reports that obesity rates are consistently higher in counties classified as food deserts.

However, most studies stop at correlation. Few have built predictive models or quantified metabolic vulnerability across geography. My project aims to close this gap by integrating multiple public datasets and developing a quantitative scoring model that can identify at-risk communities.

4. Methodology

Data Sources:

- USDA Food Access Research Atlas – distance to grocery stores, vehicle access, income level.
- CDC County Health Data – obesity and diabetes prevalence rates.
- U.S. Census Data – education, poverty, and median household income.

Processing & Integration:

1. Data cleaning and merging on county or census tract IDs.
2. Handling missing data and normalization of numerical features.
3. Feature engineering (e.g., grocery distance × income ratio).

Analysis Pipeline:

- Exploratory Analysis: Correlation heatmaps, distribution visualization.
- Regression Models: Multiple linear regression to quantify impact of access variables on obesity and diabetes rates.
- Clustering (K-Means/Hierarchical): Identify regional patterns of "nutritional risk."
- PCA: Reduce dimensions to visualize social–health clusters.
- Predictive Model: Compute "Metabolic Vulnerability Score" using weighted regression coefficients.

Evaluation:

- R^2 and Adjusted R^2 for regression accuracy.
- Silhouette Score for cluster validity.
- Interpretability: Feature importance and visualization.

5. Deliverables & Milestones

Deliverables:

- Cleaned, integrated dataset (shared publicly if permitted).
- Analytical notebook with regression, clustering, and PCA analysis.
- Visualization dashboard (Streamlit or matplotlib-based).
- Technical report with results and insights.
- Research-style writeup (potential for conference submission on public health analytics).

Milestones:
Week 1: Data collection and preprocessing
Week 2: Exploratory and regression analysis
Week 3: Clustering and PCA visualization
Week 4: Predictive scoring model & report writing

6. Team Members and Roles

Savitha – Team Lead and Data Integration Specialist

Responsibilities: Data collection, cleaning, merging; GitHub management; documentation.
Co-worker: Jane Heng

Jane Heng – Statistical Analyst & Regression Lead

Responsibilities: EDA, regression modeling, coefficient analysis.
Co-worker: Rishi

Rishi Visweswar Boppana – Clustering & PCA Specialist

Responsibilities: PCA analysis, clustering, visualization.
Co-worker: Kapil

Kapil Reddy Sanikommu– Predictive Modeling & Visualization Lead

Responsibilities: MVS computation, visualization dashboard, presentation.
Co-worker: Savitha Vijayarangan

7. Relevance to the Course

This project directly aligns with the core topics of the Machine Learning course — regression, clustering, PCA, data visualization, and interpretability.
It demonstrates how classical ML methods can be applied to a real-world societal problem (nutrition inequality and health outcomes) using open public datasets.

The approach mirrors the course learning objectives (CLOs):

- Applying appropriate ML techniques to solve meaningful real-world problems.
- Interpreting model results to communicate data-driven insights.
- Understanding ethical implications and data limitations.

8. Technical Difficulty

The project involves integrating heterogeneous datasets from USDA, CDC, and the Census - each with different schemas, formats, and temporal spans.
Key technical challenges include:

- Data integration and cleaning across geographical identifiers (FIPS codes).
- Managing missing and suppressed data values in public datasets.
- Designing feature engineering strategies (e.g., grocery distance × income ratio).
- Balancing model complexity and interpretability using regularized regression (Ridge/Lasso).
    - Conducting dimensionality reduction (PCA) and unsupervised clustering to identify latent community-level patterns.
    - Each component requires methodological rigor and coding proficiency in Python (pandas, scikit-learn, matplotlib).
    - While conceptually accessible, the analytical pipeline's cross-domain integration and interpretability make it technically demanding.

9. Novelty

While previous research has examined correlations between poverty and obesity, few studies build **predictive or scoring models** that integrate geography, transport, and income to quantify **structural vulnerability**.
This project introduces a **Metabolic Vulnerability Score (MVS)** — a new, interpretable metric that combines socio-economic and spatial access features to identify "nutrition inequality hotspots."
The novelty lies in:

- Translating qualitative "food desert" concepts into **quantitative vulnerability metrics**.
- Applying **machine learning techniques** (regression + clustering + PCA) for a **social good** context.
- Producing a **visual, interpretable model** rather than black-box predictions. Thus, the project goes beyond replicating known results — it generates new data-driven insights and reusable tools

10. Impact

The project has meaningful **academic and social impact** potential:

- **Policy relevance:** Outputs can guide local agencies in targeting interventions (e.g., funding grocery stores, improving transport).
- **Health equity contribution:** Identifies regions disproportionately burdened by structural food inaccessibility.
- **Research impact:** Can evolve into a **research paper or conference presentation** on public health analytics or geospatial data science.
- **Scalability:** The methodology can be adapted globally using other national datasets (e.g., India’s NFHS, WHO nutrition data).

By bridging data science and public health, this project demonstrates how interpretable ML can address **real-world inequality —** a lasting educational and social contribution.

11. Heilmeier Catechism

1. What are you trying to do? Quantify how food access inequality drives community-level metabolic health issues.
2. How is it done today? Mostly through descriptive correlations; few predictive, multi-variable models exist.
3. What’s new in your approach? Integrating food access, transport, and income into a single predictive 'Metabolic Vulnerability Score.'
4. Who cares? Public health researchers, local governments, and social policy analysts.
5. What difference will it make? Identify at-risk areas proactively, helping design targeted health interventions.
6. What are the risks and payoffs? Risk: incomplete datasets; Payoff: interpretable model for community health analytics.
7. How will success be measured? Model accuracy (R^2), clustering validity, clarity of interpretable insights.
8. How long will it take? 4–5 weeks of structured progress.