IENG 593: Course Project

In this project, you will work with a county-level food environment dataset compiled from multiple public sources. Each row corresponds to a U.S. county (keyed by 5-digit *Federal Information Processing Standards (FIPS)* code) and includes indicators across the following categories: **ACCESS** (e.g., supermarket proximity; percent of population with low access), **STORES** (grocery, convenience, supercenter counts and per-capita rates), **RESTAURANTS** (fast-food and full-service densities), **ASSISTANCE** (participation and availability for the *Supplemental Nutrition Assistance Program (SNAP)*, the *Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)*, and school meals), **INSECURITY** (food insecurity indicators), **TAXES** (food-related taxes where applicable), **LOCAL** (farmers' markets and local food activity), **HEALTH** (adult obesity, adult diabetes, recreation facilities), and **SOCIOECONOMIC** (income, education, unemployment, demographics). We provide two files: a curated `train` set *with* targets and a `test` set with targets withheld. This test set will be utilized to evaluate the performance of your approach. Treat the data as cross-sectional (no time-series modeling is required, but if you want to try, you can go ahead). Though you are not required to use any time series approach but make sure there's no leakage by using the latest available feature year that *does not exceed* the target's year. Your notebook should document any pre-processing techniques, including but not limited to filtering, imputation, transformations (e.g., logs/splines), and the final feature set used for modeling and evaluation.

Data Files Provided

- `train.csv` (county-level features + targets).
- `test.csv` (same features, targets removed).

Along with the data, a separate `variablesList.csv` is provided that lists every variable name, its definition/meaning, its category (e.g., ACCESS, STORES, RESTAURANTS, ASSISTANCE, INSECURITY, TAXES, LOCAL, HEALTH, and SOCIOECONOMIC categories) and its unit of measurement. Use this dictionary to guide feature selection, interpretability, and to ensure consistent reporting of units in your figures and tables. Note that "No data" fields are referenced with N/A or with -9999. Counties that didn't exist in a particular year are referenced with -8888.

Targets

- **Regression A:** `target_reg_foodinsec2123` (household food insecurity %, three-year average, 2021–23).
- **Regression B:** `target_reg_diabetes19` (adult diabetes prevalence, 2019).

- **Classification:** `label_cls_obesity_hotspot` (1 if the county's adult obesity rate in 2022 is at/above the *national* 75th percentile; 0 otherwise).

The goal is to understand how food environments relate to food security and health outcomes.

Key indicator definitions

- **Household food insecurity (%)** is the share of households reporting limited or uncertain access to adequate food; the project uses the three-year average for 2021–23. A related indicator, *very low food security*, reflects disrupted eating patterns and reduced food intake.
- **Adult obesity prevalence** is the share of adults with BMI $\geq 30\,\mathrm{kg/m^2}$, typically estimated from CDC BRFSS.[1]
- **Adult diabetes prevalence** reflects diagnosed diabetes in adults and is available at the county level from CDC's U.S. Diabetes Surveillance System.[2]

What To Do

1. **Exploratory Data Analysis (EDA) & Preprocessing.** Provide a concise feature summary table (types, missingness, basic stats), a correlation/association overview (heatmap or table), and a few targeted visual checks (histograms/boxplots for key variables; scatterplots for relationships with the targets). Document your preprocessing choices (imputation, transformations such as logs/splines, encoding, scaling) and justify any feature pruning (e.g., high missingness or strong redundancy or through LASSO/ridge).
2. **Regression A.** Model `target_reg_foodinsec2123`. You are free to choose the modeling approach, but you must need to comparatively *explain and justify* it. Use appropriate validation for model and hyperparameter selection (cross-validation; nested CV is encouraged) and report suitable regression metrics (e.g., RMSE and MAE). Include brief diagnostics and discuss interpretability.
3. **Regression B.** Model `target_reg_diabetes19` under the same principles. Compare what differs from Regression A (difficulty, influential features, generalization, etc.). Follow the appropriate validation discipline and report necessary metrics.
4. **Classification.** Model `label_cls_obesity_hotspot`. Choose and justify your classifier(s). Use appropriate validation and report classification metrics (e.g., ROC–AUC and PR–AUC; include a confusion matrix at a clearly stated threshold and a short calibration check).
5. **Unsupervised Learning.** Perform dimensionality reduction and clustering on a standardized feature subset. Summarize principal patterns (e.g., component loadings or brief factor interpretations) and characterize clusters by salient features. You are encouraged to investigate whether clusters line up with broader structures such as *state or region*, *urban–rural status*, or *feature families* (ACCESS, STORES, RESTAURANTS, ASSISTANCE, INSECURITY, TAXES, LOCAL, HEALTH, SOCIOECONOMIC). You may cluster counties using

[1] CDC: Adult obesity overview and definition: https://www.cdc.gov/obesity/adult-obesity-facts/.

[2] CDC USDSS home: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html. Methods overview and county-level estimation details: https://www.cdc.gov/diabetes/php/data-research/data-statistics/index.html.

PCA scores or standardized features and then characterize each cluster by key feature means. For interpretation, relate clusters to the *training* targets only (e.g., cluster-level averages for obesity, diabetes, or food insecurity). Feel free to be independent and try reasonable variations as long as you clearly document your choices.

6. **Uncertainty, Stability, & Selection.** For the final chosen model *in each task*, quantify *uncertainty* in the evaluation metric (e.g., bootstrap or repeated CV), assess *stability* of the selected features/importance across resamples, and describe your *model selection strategy* (criteria, complexity control, and how you avoided peeking at the test set).

7. **Test Predictions.** Produce *one* CSV with your prediction for the provided `test` data with columns `FIPS`, `y_pred_foodinsec2123`, `y_pred_diabetes19`, and `p_hat_obesityhot`. Do not include any targets.

8. **Reproducibility.** Use pipelines, fixed random seeds, and clear code organization. State all external packages used.

9. **Slides.** Up to 10 slides total: problem framing, data snapshot (with the feature summary/correlation), modeling approach and validation plan, key results (few plots), final models selected, uncertainty/stability takeaways, and concise conclusions/limitations.

Grading Rubric (100 pts)

Test-set performance (held-out evaluation) *Scored on the single submissions file*	**30**
Slides & presentation (10 slides) *Problem framing, data snapshot, methods overview and selection, key results, key visuals, uncertainty, conclusions.*	**20**
Technical rigor & Methodology (total)	**50**
Data preparation & preprocessing	12
Validation protocol & model selection	12
Hyperparameter strategy & complexity control	8
Feature selection/engineering	8
Uncertainty & Reproducibility	10
Total	**100**

Team Formation

Form a team of exactly three students and notify me by email no later than October 31 (include all members' names and WVU emails in a single message with the subject line: `''IENG 593/481 Project Team | <Team Name>"`.

Deliverables & Submission

Submit a single **zip** containing exactly three items: (1) your Jupyter notebook `.ipynb` (with all analysis and Markdown commentary), (2) a slide deck (PPT or PDF, max 10 slides), and (3) `predictions.csv` with columns `FIPS`, `y_pred_foodinsec2123`, `y_pred_diabetes19`, `p_hat_obesityhot`.

One team member must email the zip by **Dec 11, 11:59 PM**, copying the other two members. Use the subject line: `IENG 593/481 Project -- <Team Name>`.