## Capstone Project: Environmental and Socioeconomic Factors Impacting Cardiovascular Disease (CVD) in California

### Project Overview

This project explores the relationship between environmental exposures, social vulnerabilities, and cardiovascular disease (CVD) prevalence across California communities using the CalEnviroScreen 4.0 dataset. The goal is to understand which factors most strongly influence CVD rates and to build predictive models to assist public health planning.

### Data Source

The CalEnviroScreen 4.0 dataset includes over 80 environmental, health, and demographic indicators across 8,000+ California census tracts, capturing:

* Air quality (PM2.5, ozone)
* Pollution exposure (diesel particulate matter, toxic releases)
* Demographics (poverty, education, housing burden, linguistic isolation)
* Health outcomes (CVD, asthma, low birth weight)

### Modeling Approaches

Two primary models were developed:

* **Polynomial Ridge Regression (degree 2):** to capture nonlinear effects and feature interactions.
* **Random Forest Regression:** to capture complex, nonlinear relationships and interactions automatically.

### Key Results

#### Polynomial Ridge Regression

* **Test RMSE:** 3.760
* **Test R²:** 0.43
* Identified **Education, Poverty, Ozone, Linguistic Isolation,** and **Unemployment** as key factors.
* Highlighted significant interactions (e.g., PM2.5 × Groundwater Threats).

#### Random Forest Regression

* Using all features:
  + **Test RMSE:** 3.226
  + **Test R²:** 0.58
* Using **Top 7 Features** (Education, Poverty, Ozone, Unemployment, PM2.5, Toxic Releases, Drinking Water):
  + **Test RMSE:** 3.112
  + **Test R²:** 0.61

### Cross-Validation Results

Using 5-fold cross-validation on the Random Forest with the top 7 features:

* **Mean CV RMSE:** 3.130 ± 0.065
* **Mean CV R²:** 0.579 ± 0.010

This indicates stable performance across subsets, confirming the robustness of the feature set and model.

### Interpretation

By focusing on the top 7 predictive features, the Random Forest model further improved predictive accuracy while reducing complexity:

* RMSE reduced from 3.760 (Ridge) and 3.226 (RF with all features) to 3.112.
* R² improved from 0.43 (Ridge) and 0.58 (RF with all features) to 0.61.

This demonstrates that targeted, high-importance environmental and socioeconomic indicators can provide accurate, interpretable CVD risk predictions to aid public health decisions.

### Next Steps and Recommendations

* Incorporate additional or higher-resolution data for refined predictions.
* Explore advanced ensemble methods (e.g., XGBoost) for further performance gains.
* Use these findings to guide interventions focusing on pollution reduction and socio-economic support in high-risk communities.
* Perform temporal trend analyses to understand causal pathways and changes over time.

This project demonstrates that machine learning methods can effectively leverage environmental and socioeconomic data to understand and predict cardiovascular disease patterns, supporting targeted interventions and environmental justice efforts in California.