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

### Project Overview

This capstone investigates how environmental exposures and social vulnerabilities relate to cardiovascular disease (CVD) prevalence across California communities using the CalEnviroScreen 4.0 dataset. The dataset includes detailed environmental, demographic, and health indicators for 8,000+ census tracts, enabling the analysis of patterns and drivers of cardiovascular health disparities.

### Goals

* Understand spatial and statistical patterns of CVD across California.
* Identify environmental and socioeconomic factors most strongly associated with CVD rates.
* Build a predictive baseline regression model for CVD risk.
* Inform public health and environmental justice discussions.

### Key Findings

* **Top Predictors:** Education, Poverty, and Ozone levels are the strongest predictors of CVD rates.
* Environmental factors such as PM2.5, Lead, and Diesel Particulate Matter contribute moderately to CVD variations.
* **Interaction Effects:** Interactions between social and environmental factors (e.g., Poverty × PM2.5, Education × Ozone) significantly influence CVD rates.
* **Model Performance:**
  + Test R²: 0.450
  + Test RMSE: 3.695
  + 5-Fold CV RMSE: 3.650
  + 5-Fold CV R²: 0.440
* The model captured nonlinear and interaction effects while retaining interpretability, emphasizing the compounding impacts of pollution and social vulnerability on cardiovascular health.

### Methodology

* **Data Cleaning:** Handled outliers, removed negative or unrealistic CVD values, and scaled features.
* **Exploratory Data Analysis (EDA):**
  + Correlation analysis showed Poverty (0.50) and Education (0.48) had the highest correlation with CVD.
  + Environmental factors like Ozone (0.35) and PM2.5 (0.24) also correlated with CVD.
  + Scatter plots illustrated clear positive relationships between CVD rates and socioeconomic vulnerabilities.
* **Feature Engineering:** Generated polynomial and interaction features to capture nonlinear effects.
* **Modeling:**
  + Applied Ridge regression with polynomial features (degree 3) for balanced bias-variance tradeoff.
  + Used GridSearchCV to optimize regularization hyperparameters.
  + Evaluated using RMSE and R² to ensure interpretability and predictive performance.

### Recommendations

* **Policy:** Interventions addressing poverty and educational disparities will significantly impact cardiovascular health outcomes.
* **Environment:** Monitoring and mitigating ozone and PM2.5 pollution remain critical for reducing CVD rates.
* **Future Modeling:** Consider advanced models (e.g., XGBoost, Random Forest) and geospatial analysis to visualize high-risk areas for targeted interventions.

### Conclusion

This analysis demonstrates that social vulnerabilities and environmental exposures collectively influence cardiovascular disease rates in California communities. By leveraging environmental justice data, the project provides insights for public health interventions and community planning to reduce disparities in cardiovascular health outcomes.