Climate Risk In Marginalized Communities

Team: Bailey Forster, Zoe Kearney, Reeya Kumbhojkar, Viraj Meruliya, Braeden Reinoso

GitHub: https://github.com/zkearney7730/EJ-Erdos-Project

Project Overview: The Environmental Protection Agency (EPA) monitors concentrations of air toxins in the US, including particulate matter smaller than 2.5 micrometers (PM2.5). These fine particles are able to enter the lungs and bloodstream, posing a significant health risk. There is also evidence that people of color in the US are at increased risk for adverse health effects due climate change and pollution (see review article: Berberian et al. 2022). **Our goal** is to create a model that uses 2021 ACS 5-Year Estimates Data Profiles and EPA data to identify tracts likely to be at risk of high PM2.5 levels.

Stakeholders: climate and health researchers, policy makers, government.

KPI:
- Precision: % of predicted positives that are true positives
- Recall: % of true positives that are predicted as positive
- f-beta: calculated from precision and recall. Beta>1 indicates greater weight on recall, while beta<1 indicates greater weight on precision.

Our model is optimized by maximizing f2 (f-beta with beta=2). Since the goal of this model is to identify high-risk tracts for further study, we have decided to place a greater emphasis on recall over precision.

Approach: We have developed a binary classification model that predicts whether a census tract is high-risk for PM2.5 levels. We consider PM2.5> 9 $\mu g/m^3$ high risk as supported by a recent EPA proposal. Many rural tracts had insufficient environmental or demographic data, so we applied an urban cutoff of population density > 500/km^2.

- **Training:** 68% of the data was used to train 5 different model types: logistic regression, decision tree, random forest, AdaBoost, and XGBoost. GridSearchCV was used for hyperparameter tuning and cross validation on each model separately.
- **Validation:** 12% of the data was used to compare 5 model candidates (one of each type).
- **Testing:** We reserved the remaining 20% of the data to perform a final test on the best model.

19 features are considered: (See ReadMe for detailed definitions and units)

- **Population:** % black, % asian, % hispanic/latino, % white population, and population density
- **Economic:** median income, % below poverty, primary industry share, secondary industry share.
- **Environmental:** ozone, diesel particulate matter, superfund proximity, hazardous waste proximity, wastewater discharge, underground storage tanks, RMP facility proximity, housing units built before 1960, and traffic proximity.

Results: XGBoost model with all 19 features achieved **93% accuracy** and **89% f2-score.** Feature importance gives new insights into causes and distribution of PM2.5 risk. Ozone and Diesel PM are among the top contributors, thus these other air toxins are good indicators of high PM2.5. **Without Ozone and Diesel PM** the model still provides a reasonable prediction of risk with **84% accuracy** and **78% f2-score.**

Future Directions:

- Separate classification models for target groups and use feature importance to identify the most significant **PM2.5 sources** or **health risks** affecting minority groups.
  - E.g. control for areas with high hispanic/latino populations
  - Allow policy makers and government officials to create targeted solutions
  - NOTE: need CDC health data with current tract boundaries to be made available
- Multinomial model: low, medium, high risk based on WHO, EPA, US standards
- Rural model: how does feature importance change?