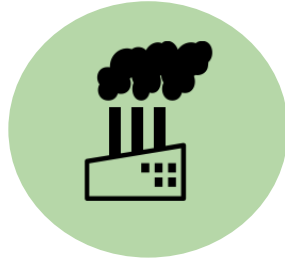# Environmental Justice: *Asthma Rates and at Risk Communities*

By: Sidhika Tripathee, Ali Baratta, and Ghazi Randhawa

1.
Introduction

Background and Motivation

Background

Air Pollution

- The silent killer
- 'Lost years of life'

Environmental Injustice

- Are communities of color afflicted by higher level of asthma rates?

Can we use machine learning to determine which factors have the most significant impact on asthma rates in a community?
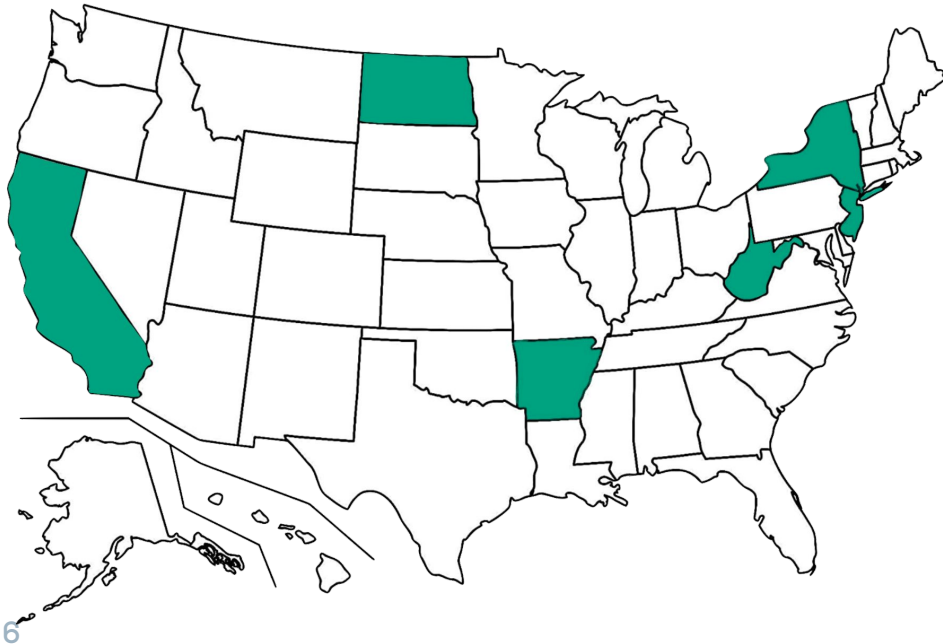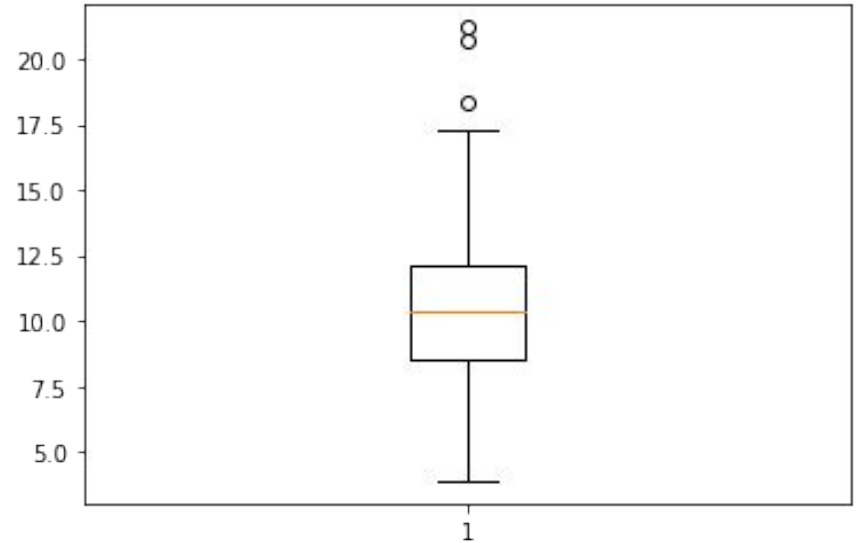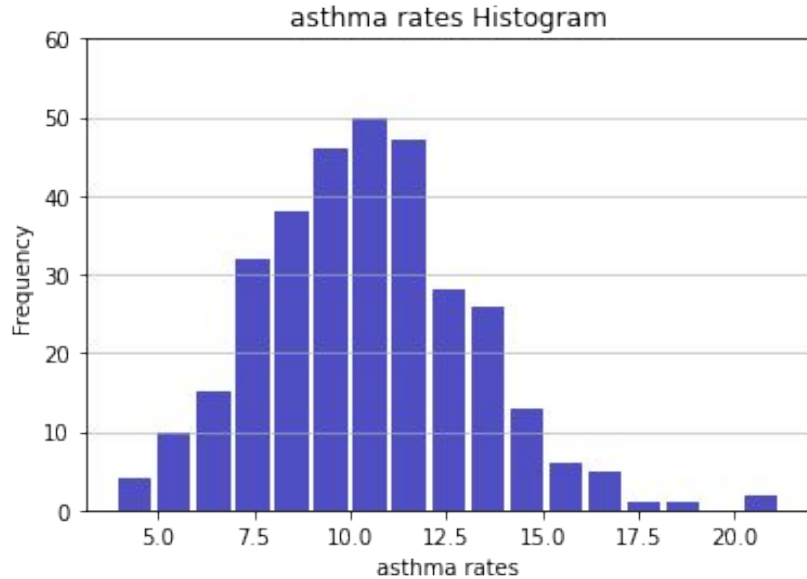
2.
Our Dataset

Collection, Preprocessing, Labeling
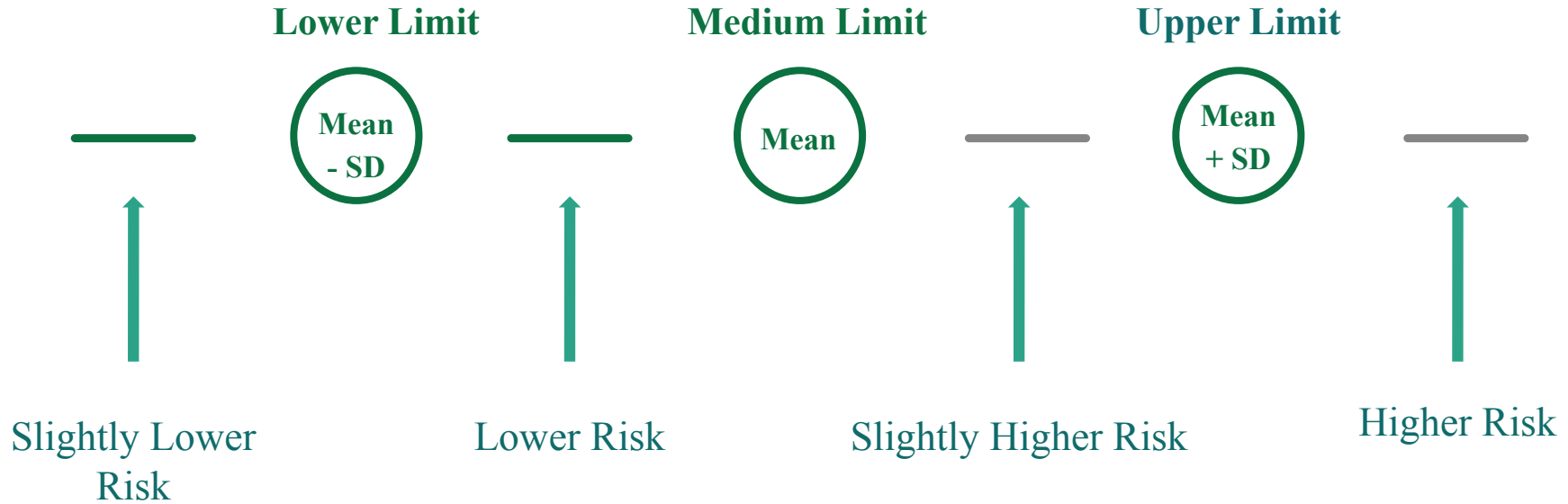
Data Collection

- Obtained county data from NY, NJ, ND, CA, WV, AR (324 total counties)
- For each county collected information on:
  - # of waste facility locations
  - PM2.5 concentration
  - Racial demographics
  - Population
  - Asthma rates (label)

Data Preprocessing

asthma rates Histogram

- created four labels: slighter lower, lower, slightly higher, higher risk

Labeling

Lower Limit

Medium Limit

Upper Limit

Mean - SD

Mean

Mean + SD

Slightly Lower Risk
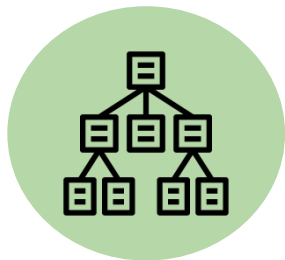
Lower Risk

Slightly Higher Risk

Higher Risk

3.
Experimental Design

An overview of our process

Our Process

Regression Models

- Logistic Regression
- ID3
- Random Forests
- SVMs

Hypertuning Parameters

- Pipeline
- Grid Search
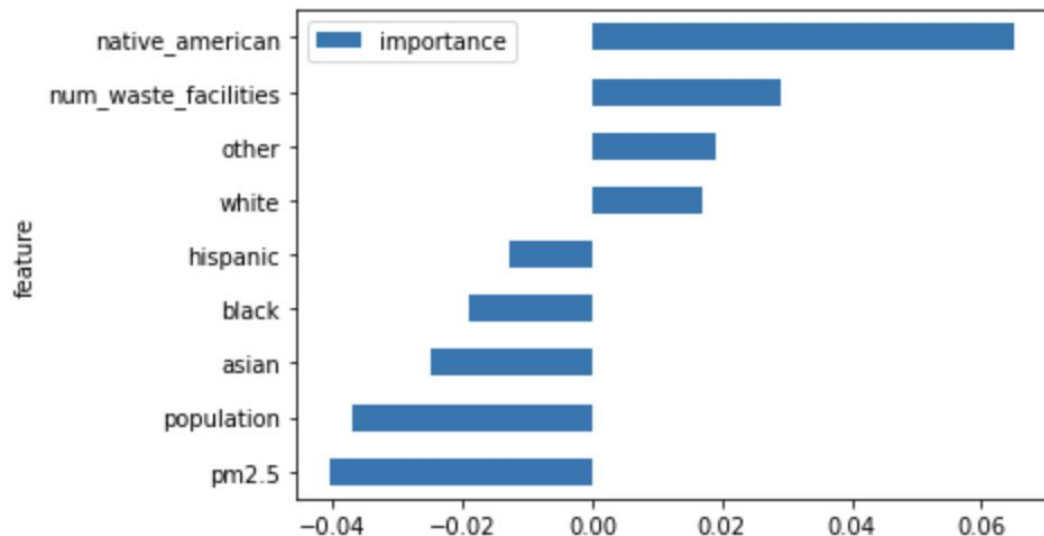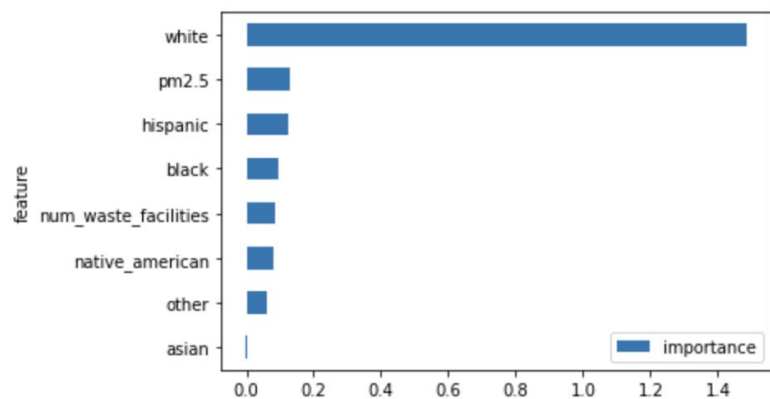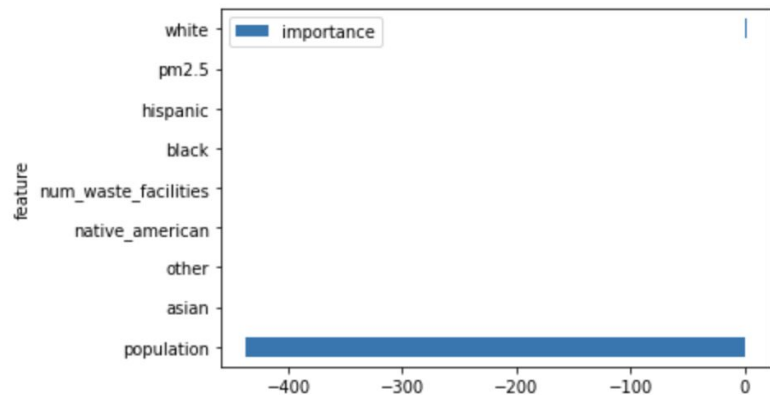- Feature Importance graphs

Analysis

- Accuracy among models
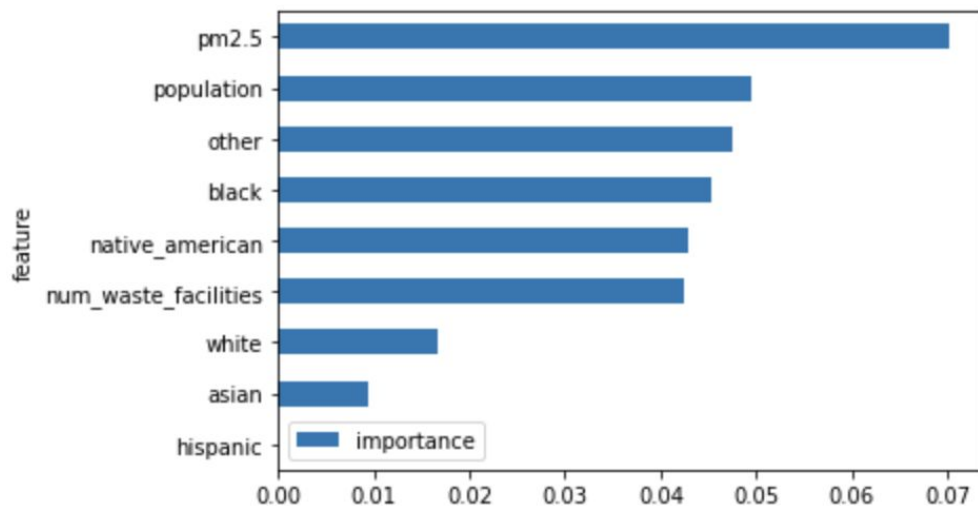- Feature weights
- Impact of changing labeling techniques

4.
Results

Feature Importance, Model Performance
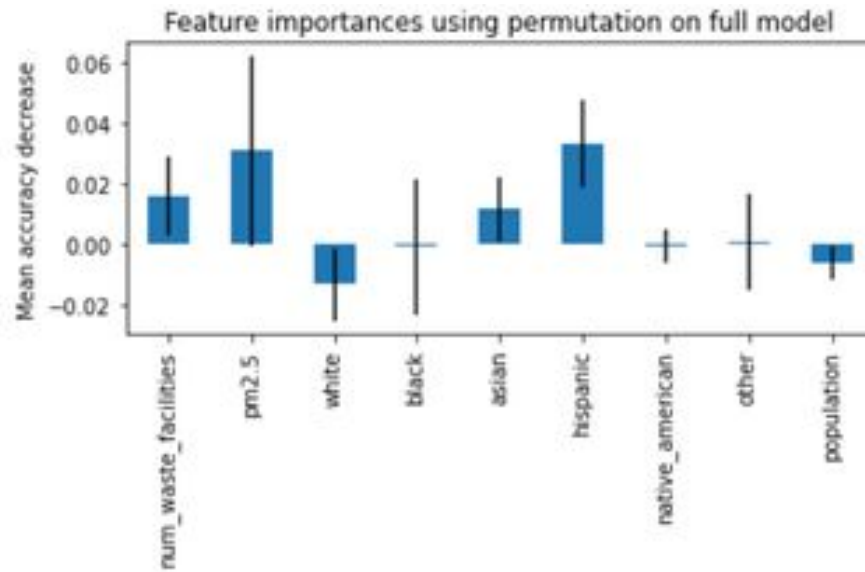
Logistic Regression

Decision Trees

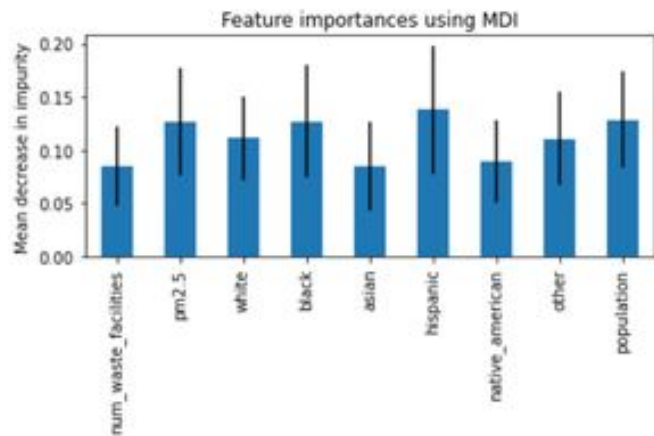

Feature	Feature Importance
# waste facilities	0.042
PM2.5 concentration	0.07
White	0.017
Black	0.045
Asian	0.01
Hispanic	0.0
Native American	0.043
Other	0.047
Population	0.049

Random Forests

Feature importances using MDI

Feature importances using permutation on full model

SVMs
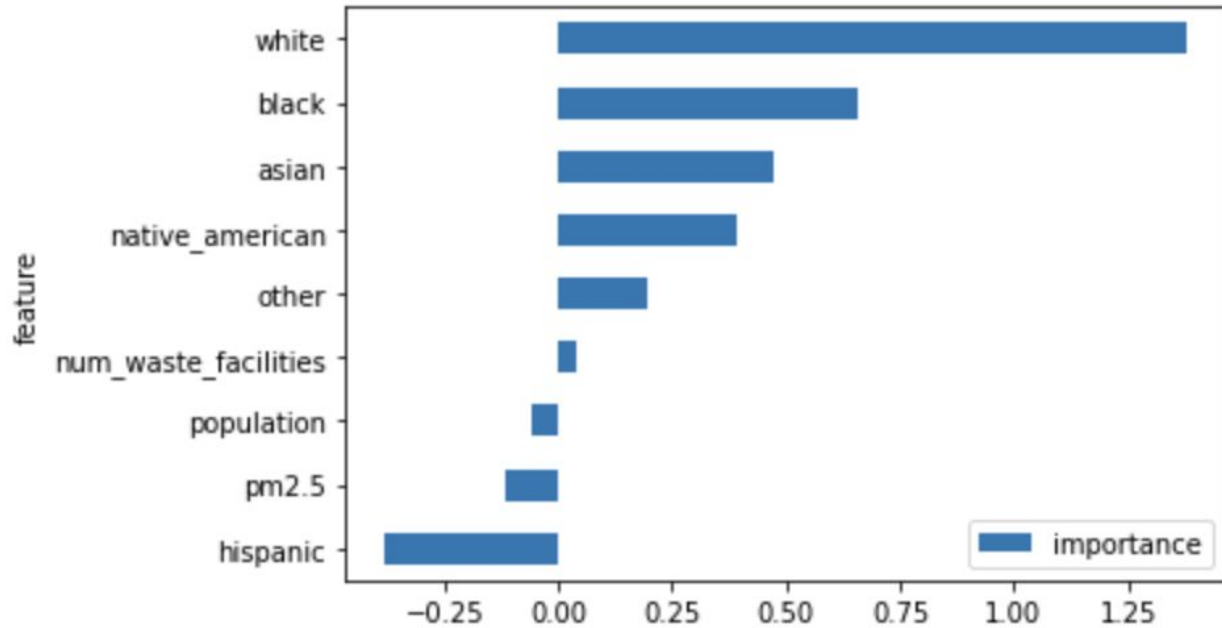
Model Performance

	Hypertuned Parameters	Highest Accuracy
Logistic Regression	**Penalty:** 12 **C:** 0.43	34.6%
Decision Trees	**max depth:** 5-6	38.5%
Random Forests	**# of estimators:** 45,	38.5%
SVM	**C:** 1000 **Gamma:** 0.1 **Kernel:** linear	42.0%

CV = 4

Shifting Benchmarks

Cutoff

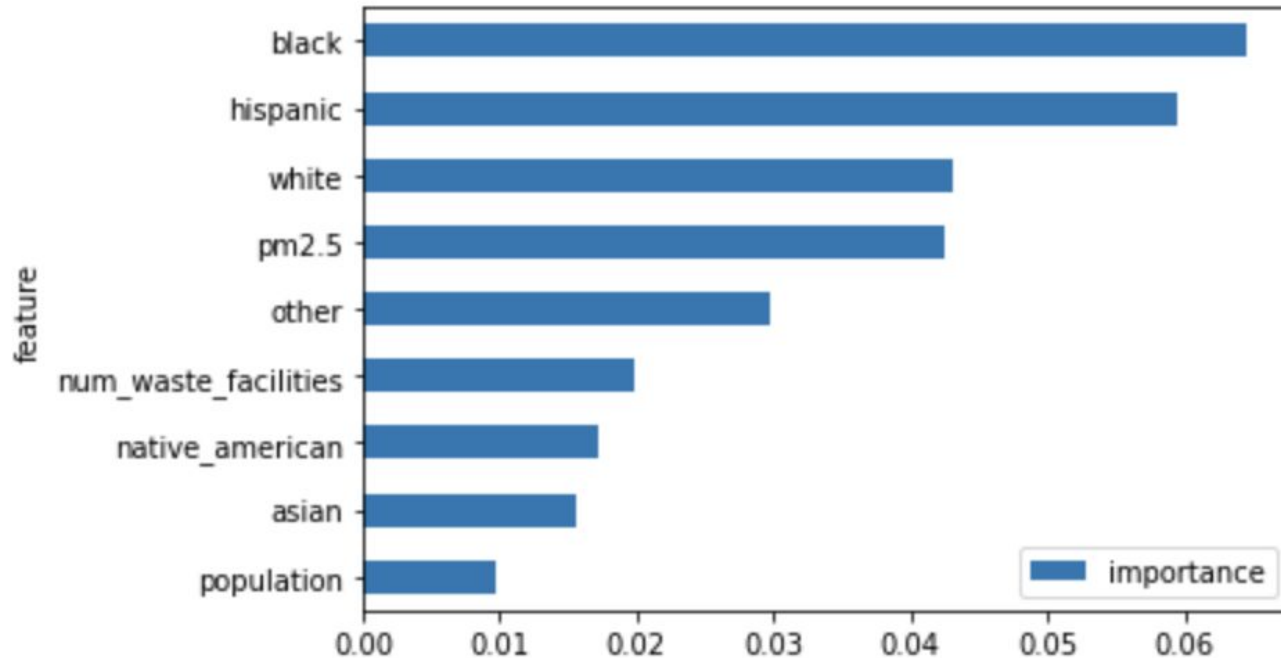Lower Risk

Mean + 0.5*SD

Higher Risk
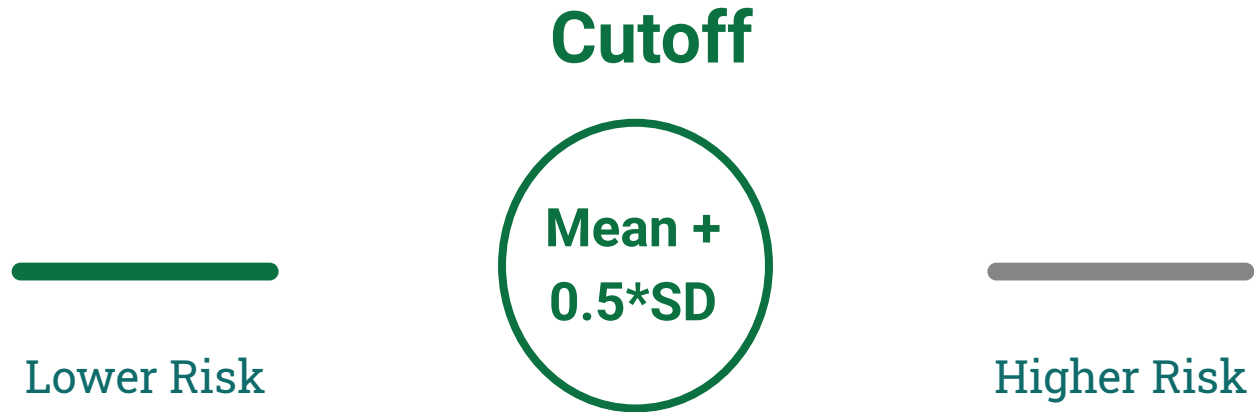
Shifting Benchmarks - Logistic Regression

Accuracy: 75.38%
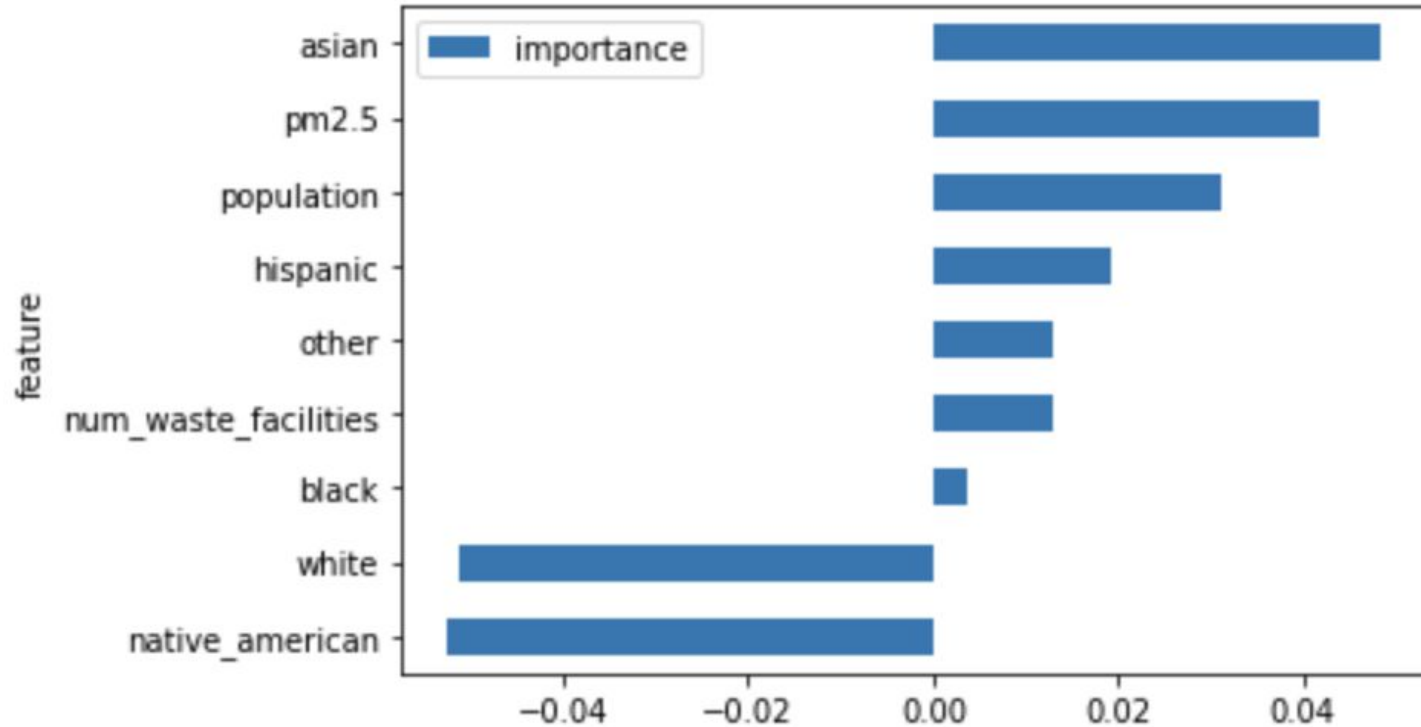
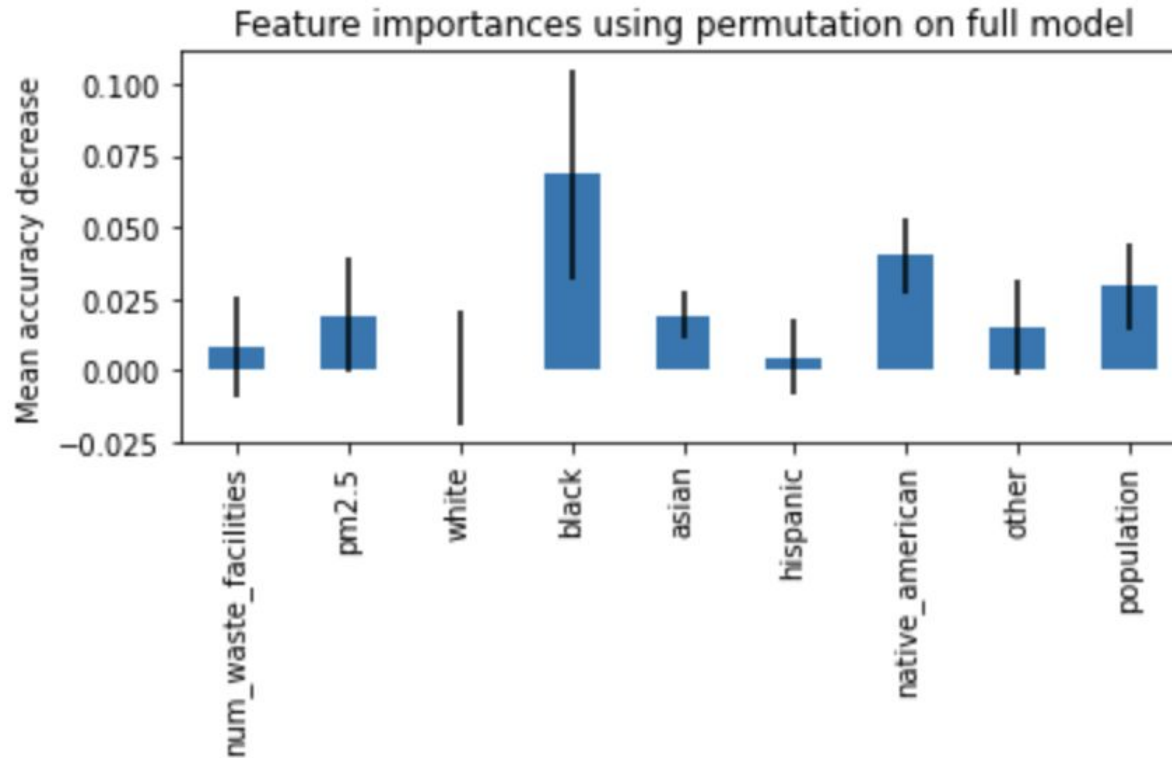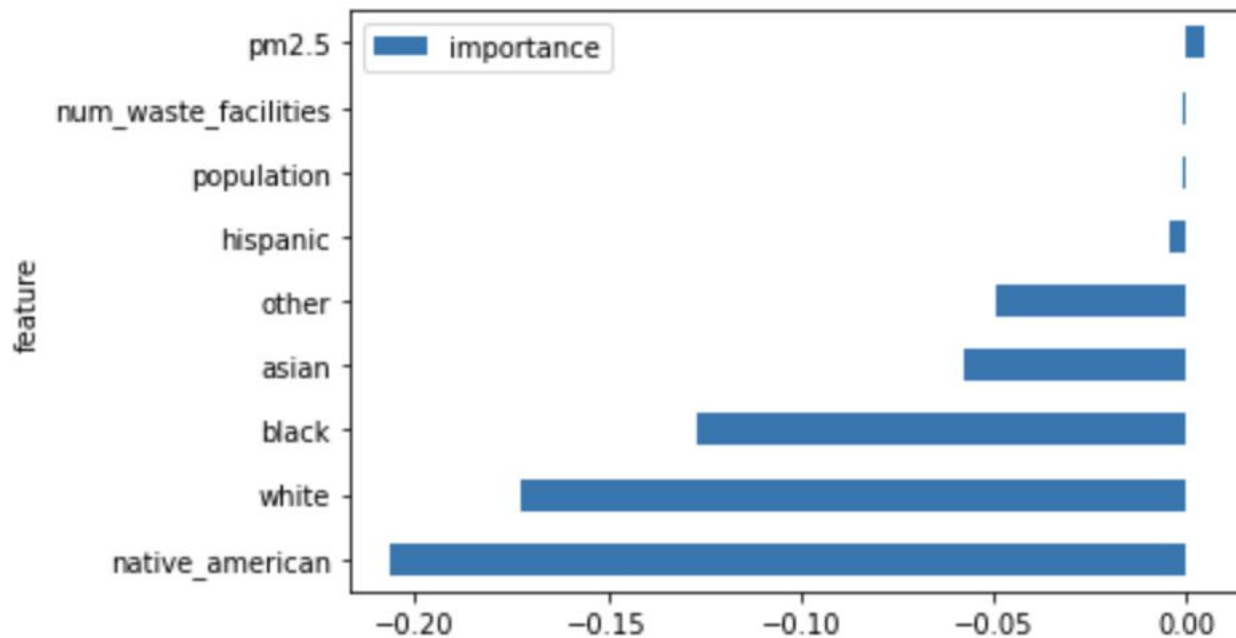# Shifting Benchmarks - Decision Trees

Accuracy: 72.3%

Shifting Benchmarks - Random Forests

Accuracy Rate: 71.0%

Feature importances using permutation on full model

Shifting Benchmarks - SVMs
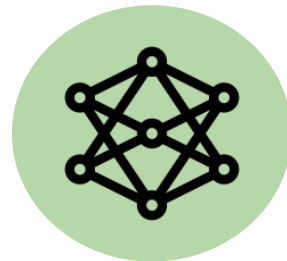
Accuracy: 71.3%

Improvements

More Data!
- Add county data from all states
- Additional features (income, pop density, car ownership rates)

Account for Feature Correlations
- # waste facilities + PM2.5 levels
- Racial demographics + waste facilities

Other Models
- Neural networks
- Naive Bayes

5.
Impact

Importance of our results

Impact

- Model predicts that marginalized communities tend to have higher risks of asthma rates

- Model helps generate evidence of environmental injustice

References

- Data:
    - U.S. Chronic Disease Indicators: Asthma (2016)
    - EPA Historical Air Quality
    - Daily Census Tract-Level PM2.5 Concentrations, 2016
    - Solid Waste Landfill Facilities
- Air Pollution and Asthma | AAFA.org
- Environmental Justice in Chester | Chester Digital
- Summary of the Clean Air Act
- Summary of Executive Order 12898 - Federal Actions to Address Environmental Justice in Minority Populations and Low-Income Populations
- The Distributive Effects of Risk Prediction in Environmental Compliance: Algorithmic Design, Environmental Justice, and Public Policy

Thank you!

Questions?