# WiDS Datathon 2024 Challenge #1: Equity in Healthcare

By Betty Hagos

## Overview

The WiDS Datathon 2024 focuses on a prediction task using a roughly 39k record dataset (split into training and test sets) representing patients and their characteristics (age, race, BMI, zip code), their diagnosis and treatment information (breast cancer diagnosis code, metastatic cancer diagnosis code, metastatic cancer treatments, … etc.), their geo (zip-code level) demographic data (income, education, rent, race, poverty, …etc), as well as toxic air quality data (Ozone, PM25 and NO2) that tie health outcomes to environmental conditions. Each row in the data corresponds a single patient and her Diagnosis Period.

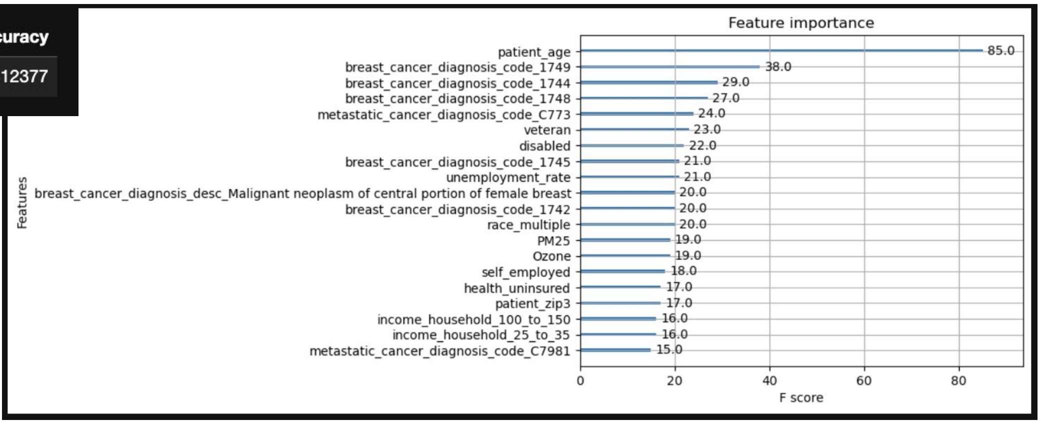
## Objective

The task is to assess whether the likelihood of the patient's Diagnosis Period being less than 90 days is predictable using these characteristics and information about the patient. Ultimately, we want to uncover any issues of equity in metastatic breast cancer diagnosis.

## Results

After an initial exploratory data analysis of the 83 column dataset, given the missing values and the underlying binary logistic regression task, I built a model using XGBoost & hyperparameter tuning to select the most important features for the next steps of the challenge.

| | Model | F1 | Recall | Precision | Accuracy |
|---|---|---|---|---|---|
| 0 | XGBoost CV | 0.865323 | 0.96493 | 0.784388 | 0.812377 |


Feature importance

The model performed very well and the most predictive variables in the data set include age, veteran status, disability status, income, race_multiple, Ozone, PM25, uninsured, self employed, and zip code in addition to specific medical indicators. This suggests that there are issues of equity to continue to explore.

## Next Steps

Visualize the relationships between the demographic data, environmental data, patient characteristics and outcomes to unearth the story of inequity in healthcare.

I am considering using Logistic Regression and/or Naive Bayes to calculate the posterior probabilities for the next step, which is to find the likelihood of a patient's diagnosis being less than 90 days given these predictive features.